

Hanna Hedeland, Thomas Schmidt, Kai Wörner (eds.)

Multilingual Resources and Multilingual Applications

Proceedings of the Conference of the
German Society for Computational Linguistics and
Language Technology (GSCL) 2011



Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Sonderforschungsbereich
Mehrsprachigkeit



Hanna Hedeland, Thomas Schmidt, Kai Wörner (eds.)

Multilingual Resources and Multilingual Applications

Proceedings of the Conference of the German Society for
Computational Linguistics and Language Technology (GSCL) 2011

© *Hanna Hedeland, Thomas Schmidt, Kai Wörner*
Hamburger Zentrum für Sprachkorpora
Max Brauer-Allee 60
D-22765 Hamburg

Die „Arbeiten zur Mehrsprachigkeit – Folge B“ publizieren Forschungsarbeiten aus dem Sonderforschungsbereich 538 *Mehrsprachigkeit*, der von der Deutschen Forschungsgemeinschaft im Juli 1999 an der Universität Hamburg eingerichtet wurde. Wir danken der DFG für ihre Unterstützung.

Die „Arbeiten zur Mehrsprachigkeit – Folge B“ sind bei der Deutschen Bibliothek in Frankfurt/Main mit der Seriennummer ISSN 0176-559X eingetragen.

Redaktion:
Martin Elsig, Svenja Kranich, Thomas Schmidt, Manuela Schönenberger
Technische Umsetzung:
Thomas Schmidt

Collaborative Research Center: Multilingualism
Sonderforschungsbereich 538: Mehrsprachigkeit
University of Hamburg

Founded in July 1999, the Collaborative Research Centre on Multilingualism conducts research on patterns of language use in multilingual environments, bilingual language acquisition, and the role of multilingualism and language contact for language change.

In the current, fourth funding period (2008–2011), the Centre comprises two main research branches, each of which focuses on a central set of common issues, and a third branch of projects dealing with practical application solutions. Branch E, *Multilingual Language Acquisition*, consists of four projects, with a common focus on the nature of “critical phases” in language acquisition. Branch H, *Historical Aspects of Multilingualism and Variation*, consists of eight projects, dealing with questions of language change and language contact. This branch also comprises projects of former separate branch K, *Multilingual Communication*.

Since 2007, a new Branch T, *Transfer Projects*, has been active. It consists of five projects whose goal is to develop concrete solutions for practical problems relating to multilingual situations, based on research results derived from the Centre’s research activities.

Languages currently studied at the Research Centre include the following: Danish, Catalan, English, Faroese, French, German, German Sign Language, Icelandic, Irish, Polish, Portuguese, Spanish, Swedish, and Turkish, as well as several historical or regional sub-varieties of some of these languages.

Chair:

Prof. Dr. Christoph Gabriel
christoph.gabriel@uni-hamburg.de

Co-chairs:

Prof. Dr. Kurt Braunmüller
braunmueller@rrz.uni-hamburg.de

Prof. Dr. Barbara Hänel-Faulhaber
Barbara.Haenel@uni-hamburg.de

Local Organizing Comittee

- Thomas Schmidt
- Kai Wörner
- Timm Lehmberg
- Hanna Hedeland

Program Committee

- Maja Bärenfänger (Universität Gießen)
- Stefanie Dipper (Universität Bochum)
- Kurt Eberle (Lingnio Heidelberg)
- Alexander Geyken (Berlin-Brandenburgische Akademie der Wissenschaften)
- Ullrich Heid (Universität Hildesheim)
- Claudia Kunze (Qualisys GmbH)
- Lothar Lemnitzer (Berlin-Brandenburgische Akademie der Wissenschaften)
- Henning Lobin (Universität Gießen)
- Ernesto de Luca (Technische Universität Berlin)
- Cerstin Mahlow (Universität Zürich)
- Alexander Mehler (Universität Bielefeld)
- Wolfgang Menzel (Universität Hamburg)
- Georg Rehm (Deutsches Forschungszentrum für Künstliche Intelligenz)
- Josef Ruppenhofer (Universität Saarbrücken)
- Thomas Schmidt (Universität Hamburg)
- Roman Schneider (Institut für Deutsche Sprache Mannheim)
- Bernhard Schröder (Universität Duisburg)
- Manfred Stede (Universität Potsdam)
- Angelika Storrer (Universität Dortmund)
- Maik Stührenberg (Universität Bielefeld)
- Thorsten Trippel (Universität Tübingen)
- Cristina Vertan (Universität Hamburg)
- Andreas Witt (Institut für Deutsche Sprache Mannheim)
- Christian Wolff (Universität Regensburg)
- Kai Wörner (Universität Hamburg)

Call for Papers

The Conference of the German Society for Computational Linguistics and Language Technology (GSCL) in 2011 will take place from 28th to 30th September 2011 at the University of Hamburg. The main conference theme is Multilingual Resources and Multilingual Applications.

Contributions to any topic related to Computational Linguistics and Language Technology are invited, but we especially encourage submissions that are related to the main theme. The topic “Multilingual Resources and Multilingual Applications” comprises all aspects of computational linguistics and speech and language technology in which issues of multilingualism, of language contrasts or of language independent representations play a major role. This includes, for instance:

- representation and analysis of parallel corpora and comparable corpora
- multilingual lexical resources
- machine translation
- annotation and analysis of learner corpora
- linguistic variation in linguistic data and applications
- localisation and internationalisation

Conference languages are English and German; contributions are welcome in both languages. Three types of submission are invited:

- Regular talk – Submission of an extended abstract
- Poster – Submission of an abstract
- System presentation – Submission of an abstract

Only contributions in electronic form will be accepted. We do not provide style sheets for submissions at this stage; constraints on length are given below. Submissions must be submitted via the electronic conference system. Accepted abstracts will be published as a book of abstracts on the occasion of the conference. Extended versions of the papers will be published as a special issue of the Journal for Language Technology and Computational Linguistics after the conference.

GSCL

*German Society for
Computational Linguistics
& Language Technology*



Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

hzsk hamburger zentrum
für sprachkorpora

Contents

I Invited Talks

- Constructing Parallel Lexicon Fragments Based on English FrameNet Entries:
Semantic and Syntactic Issues* 9
Hans C. Boas
- The Multilingual Web: Opportunities, Borders and Visions*..... 19
Felix Sasaki
- Combining Various Text Analysis Tools for Multilingual Media Monitoring*..... 25
Ralf Steinberger

II Regular Papers

- Generating Inflection Variants of Multi-Word Terms for French and German*..... 33
Simon Clematide, Luzia Roth
- Tackling the Variation in International Location Information Data: An Approach
Using Open Semantic Databases* 39
Janine Wolf, Manfred Stede, Michaela Atterer
- Towards Multilingual Biographical Event Extraction - Initial Thoughts on the
Design of a New Annotation Scheme*..... 45
Michaela Geierhos, Jean-Leon Bouraoui, Patrick Watrin
- The Corpus of Academic Learner English (CALE): A New Resource for the Study
of Lexico-Grammatical Variation in Advanced Learner Varieties*..... 51
Marcus Callies, Ekaterina Zaytseva
- From Multilingual Web-Archives to Parallel Treebanks in Five Minutes*..... 57
Markus Killer, Rico Sennrich, Martin Volk
- Querying Multilevel Annotation and Alignment for Detecting Grammatical
Valence Divergencies* 63
Oliver Čulo
- SPIGA – A Multilingual News Aggregator*..... 69
Leonhard Hennig, Danuta Ploch, Daniel Prawdzik, Benjamin Armbruster,
Christoph Büscher, Ernesto William De Luca, Holger Düwiger, Şahin Albayrak
- From Historic Books to Annotated XML: Building a Large Multilingual
Diachronic Corpus* 75
Magdalena Jitca, Rico Sennrich, Martin Volk
- Visualizing Dependency Structures*..... 81
Chris Culy, Verena Lyding, Henrik Dittmann
- A Functional Database Framework for Querying Very Large Multi-Layer Corpora*..... 87
Roman Schneider

<i>Hybrid Machine Translation for German in taraXÜ: Can Translation Costs Be Decreased Without Degrading Quality?</i>	93
Aljoscha Burchardt, Christian Federmann, Hans Uszkoreit	
<i>Annotation of Explicit and Implicit Discourse Relations in the TüBa-D/Z Treebank</i>	99
Anna Gastel, Sabrina Schulze, Yannick Versley, Erhard Hinrichs	
<i>Devil's Advocate on Metadata in Science</i>	105
Christina Hoppermann, Thorsten Trippel, Claus Zinn	
<i>Improving an Existing RBMT System by Stochastic Analysis</i>	111
Christian Federmann, Sabine Hunsicker	
<i>Terminology Extraction and Term Variation Patterns: A Study of French and German Data</i>	117
Marion Weller, Helena Blancafort, Anita Gojun, Ulrich Heid	
<i>Ansätze zur Verbesserung der Retrieval-Leistung kommerzieller Translation-Memory-Systeme</i>	123
Dino Azzano, Uwe Reinke, Melanie Sauer	
<i>WikiWarsDE: A German Corpus of Narratives Annotated with Temporal Expressions</i>	129
Jannik Strötgen, Michael Gertz	
<i>Translation and Language Change with Reference to Popular Science Articles: The Interplay of Diachronic and Synchronic Corpus-Based Studies</i>	135
Sofia Malamatidou	
<i>A Comparable Wikipedia Corpus: From Wiki Syntax to POS Tagged XML</i>	141
Noah Bubenhofer, Stefanie Haupt, Horst Schwinn	
<i>A German Grammar for Generation in OpenCCG</i>	145
Jean Vancoppenolle, Eric Tabbert, Gerlof Bouma, Manfred Stede	
<i>Multilingualism in Ancient Texts: Language Detection by Example of Old High German and Old Saxon</i>	151
Zahurul Islam, Roland Mittmann, Alexander Mehler	
<i>Multilinguale Phrasenextraktion mit Hilfe einer lexikonunabhängigen Analysekomponente am Beispiel von Patentschriften und nutzergenerierten Inhalten</i>	157
Daniela Becks, Julia Maria Schulz, Christa Womser-Hacker, Thomas Mandl	
<i>Die Digitale Rätoromanische Chrestomathie – Werkzeuge und Verfahren für die Korpuserstellung durch kollaborative Volltexterschließung</i>	163
Claes Neuefeind, Jürgen Rolshoven, Fabian Steeg	
<i>Ein umgekehrtes Lehnwörterbuch als Internetportal und elektronische Ressource: Lexikographische und technische Grundlagen</i>	169
Peter Meyer, Stefan Engelberg	
<i>Localizing a Core HPSG-Based Grammar for Bulgarian</i>	175
Petya Osenova	

III Poster Presentations

<i>Autorenunterstützung für die Maschinelle Übersetzung</i>	183
Melanie Siegel	
<i>Experimenting with Corpus-Based MT Approaches</i>	187
Monica Gavrilă	
<i>Method of POS-Disambiguation Using Information about Words Co-Occurrence (For Russian)</i>	191
Edward Klyshinsky, Natalia Kochetkova, Maxim Litvinov, Vadim Maximov	
<i>Von TMF in Richtung UML: In drei Schritten zu einem Modell des übersetzungsorientierten Fachwörterbuchs</i>	197
Georg Löckinger	
<i>Annotating for Precision and Recall in Speech Act Variation: The Case of Directives in the Spoken Turkish Corpus</i>	203
Şükriye Ruhi, Thomas Schmidt, Kai Wörner, Kerem Eryılmaz	
<i>The SoSaBiEC Corpus: Social Structure and Bilinguality in Everyday Conversation</i>	207
Veronika Ries, Andy Lücking	
<i>DIL, ein zweisprachiges Online-Fachwörterbuch der Linguistik (Deutsch-Italienisch)</i>	211
Carolina Flinz	
<i>Knowledge Extraction and Representation: The EcoLexicon Methodology</i>	215
Pilar León Araúz, Arianne Reimerink	
<i>Processing Multilingual Customer Contacts via Social Media</i>	219
Michaela Geierhos, Yeong Su Lee, Matthias Bargel	
<i>ATLAS – A Robust Multilingual Platform for the Web</i>	223
Diman Karagiozov, Svetla Koeva, Maciej Ogrodniczuk, Cristina Vertan	
<i>Multilingual Corpora at the Hamburg Centre for Language Corpora</i>	227
Hanna Hedeland, Timm Lehmberg, Thomas Schmidt, Kai Wörner	
<i>The English Passive and the German Learner – Compiling an Annotated Learner Corpus to Investigate the Importance of Educational Settings</i>	233
Verena Möller, Ulrich Heid	
<i>Register, Genre, Rhetorical Functions: Variation in English Native-Speaker and Learner Writing</i>	239
Ekaterina Zaytseva	
<i>Tools to Analyse German-English Contrasts in Cohesion</i>	243
Kerstin Kunz, Ekaterina Lapshinova-Koltunski	
<i>Comparison and Evaluation of Ontology Extraction Systems</i>	247
Stefanie Reimers	

IV System Presentations

<i>New and Future Developments in EXMARaLDA</i>	253
Thomas Schmidt, Kai Wörner, Hanna Hedeland, Timm Lehmborg	
<i>Der VLC Language Index</i>	257
Dirk Schäfer, Jürgen Handke	
<i>Topological Fields, Constituents and Coreference: A New Multi-Layer Architecture for TüBa-D/Z</i>	259
Thomas Krause, Julia Ritz, Amir Zeldes, Florian Zipser	
<i>MT Server Land Translation Services</i>	263
Christian Federmann	

Invited Talks

Constructing parallel lexicon fragments based on English FrameNet entries: Semantic and syntactic issues

Hans C. Boas

The University of Texas at Austin

Department of Germanic Studies and Department of Linguistics

1 University Station, C3300, Austin, TX 78712-0304, U.S.A.

E-mail: hcb@mail.utexas.edu

Abstract

This paper investigates how semantic frames from FrameNet can be re-used for constructing FrameNets for other languages. Section one provides a brief overview of Frame Semantics (Fillmore, 1982). Section 2 introduces the main structuring principles of the Berkeley FrameNet project. Section three presents a typology of FrameNets for different languages, highlighting a number of important issues surrounding the universal applicability of semantic frames. Section four shows that while it is often possible to re-use semantic frames across languages in a principled way it is not always straightforward because of systematic syntactic differences in how lexical units express the semantics of frames. Section five summarizes the issues discussed in this paper.

Keywords: Computational Lexicography, FrameNet, Frame Semantics, Syntax

1. Frame Semantics

Research in Frame Semantics (Fillmore, 1982; 1985) is empirical, cognitive, and ethnographic in nature. It seeks to describe and analyze what users of a language understand about what is communicated by their language (Fillmore & Baker, 2010). Central to this line of research is the notion of *semantic frame*, which provides the basis for the organization of the lexicon, thereby linking individual word senses, relationships between the senses of polysemous words, and relationships among semantically related words. In this conception of the lexicon, there is a network of hierarchically organized and intersecting frames through which semantic relationships between collections of concepts are identified (Petrucek et al., 2004). A frame is any system of concepts related in such a way that to understand any one concept it is necessary to understand the entire system; introducing any one concept results in all of them becoming available. In Frame Semantics, word meanings are thus characterized in terms of experience-based schematizations of the speaker's world, i.e. frames. It is held that understanding any element in a frame requires access to an understanding

of the whole structure (Petrucek & Boas, 2003).¹ The following section shows how the concept of semantic frame has been used to structure the lexicon of English for the purpose of creating a lexical database.

2. The Berkeley FrameNet Project

The Berkeley FrameNet Project (Lowe et al., 1997; Baker et al., 1998; Fillmore et al., 2003a; Ruppenhofer et al., 2010) is building a lexical database that aims to provide, for a significant portion of the vocabulary of contemporary English, a body of semantically and syntactically annotated sentences from which reliable information can be reported on the valences or combinatorial possibilities of each item targeted for analysis (Fillmore & Baker, 2001). The method of inquiry is to find groups of words whose frame structures can be described together, by virtue of their sharing common schematic backgrounds and patterns of expressions that can combine with them to form larger phrases or sentences. In the typical case, words that share a frame can be used in paraphrases of each other. The general purposes of the project are both to provide

¹See Petrucek (1996), Ziem (2008), and Fillmore & Baker (2010) on how different theories employ the notion of "frame."

reliable descriptions of the syntactic and semantic combinatorial properties of each word in the lexicon, and to assemble information about alternative ways of expressing concepts in the same conceptual domain (Fillmore & Baker, 2010).

To illustrate, consider the sentence *Joe stole the watch from Michael*. The verb *steal* is said to evoke the **Theft** frame, which is also evoked by a number of semantically related verbs such as *snatch*, *shoplift*, *pinch*, *filch*, and *thieve*, among others, as well as nouns such as *thief* and *stealer*.² The **Theft** frame represents a scenario with different Frame Elements (FEs) that can be regarded as instances of more general semantic roles such as **AGENT**, **PATIENT**, **INSTRUMENT**, etc. More precisely, the **Theft** frame describes situations in which a **PERPETRATOR** (the person or other agent that takes the **GOODS** away) takes **GOODS** (anything that can be taken away) that belong to a **VICTIM** (the person (or other sentient being or group) that owns the **GOODS** before they are taken away by the **PERPETRATOR**). Sometimes more specific information is given about the **SOURCE** (the initial location of the **GOODS** before they change location).³ The necessary background information to interpret *steal* and other semantically related verbs as evoking the **Theft** frame also requires an understanding of illegal activities, property ownership, taking things, and a great deal more (see Boas, 2005b; Bertoldi et al., 2010; Dux, 2011).

Based on the frame concept, FrameNet researchers follow a lexical analysis process that typically consists of the following steps according to Fillmore & Baker (2010:321-322): (1) Characterizing the frames, i.e. the situation types for which the language has provided special expressive means; (2) Describing and naming the Frame Elements (FEs), i.e. the aspects and components of individual frames that are likely to be mentioned in the phrases and sentences that are instances of those frames; (3) Selecting lexical units (LUs) that belong to the frame, i.e. words from all parts of speech that evoke and depend on the conceptual

²Names of frames are in courier font. Names of Frame Elements (FEs) are in small caps font.

³Besides so-called core Frame Elements, there are also peripheral Frame Elements that describe more general aspects of a situation, such as **MEANS** (e.g. *by trickery*), **TIME** (e.g. *two days ago*), **MANNER** (e.g. *quietly*), or **PLACE** (e.g. *in the city*).

background associated with the individual frames; (4) Creating annotations of sentences sampled from a very large corpus showing the ways in which individual LUs in the frame allow frame-relevant information to be linguistically presented; (5) Automatically generating lexical entries, and the valence descriptions contained in them, that summarize observations derivable from them (see also Atkins et al., 2003; Fillmore & Petruck, 2003; Fillmore et al., 2003b; Ruppenhofer et al., 2010).

The results of this work-flow are stored in FrameNet (<http://framenet.icsi.berkeley.edu>), an online lexical database (Baker et al., 2003) currently containing information about more than 1,000 frames and more than 10,000 LUs.⁴ Users can access FrameNet data in a variety of ways. The most prominent methods include searching for individual frames or specific LUs.

Valence Patterns:

These frame elements occur in the following syntactic patterns:

Number Annotated	Patterns				
1 TOTAL	Frequency	Goods	Perpetrator	Victim	
(1)	AVP Dep	NP Obj	NP Ext	PP[from] Dep	
1 TOTAL	Goods	Instrument	Time		
(1)	NP Obj	NP Ext	PP[after] Dep		
1 TOTAL	Goods	Manner	Perpetrator		
(1)	NP Obj	AVP Dep	NP Ext		
1 TOTAL	Goods	Means	Perpetrator	Place	Victim
(1)	NP Ext	PP[in] Dep	CNI --	PP[at] Dep	INI --
1 TOTAL	Goods	Means	Perpetrator	Source	
(1)	NP Obj	PPing[by] Dep	NP Ext	INI --	
1 TOTAL	Goods	Means	Perpetrator	Victim	
(1)	NP Obj	PPing[by] Dep	NP Ext	INI --	

Figure 1: Partial valence table for *steal.v* in the **Theft** frame

Each entry for a LU in FrameNet consists of the following parts: (1) A description of the frame together with definitions of the relevant FEs, annotated examples sentences illustrating the relevant FEs in context, and a list of other LUs evoking the same frame; (2) An annotation report displaying all the annotated corpus

⁴For differences between FrameNet and other lexical databases such as WordNet see Boas (2005a/2005b/2009).

sentences for a given LU; (3) A lexical entry report which summarizes the syntactic realization of the FEs and the valence patterns of the LU in two separate tables (see Fillmore et al., 2003B; Fillmore, 2007).

Figure 1 above illustrates an excerpt from the valence patterns in the lexical report of *steal* in the Theft frame. The column on the far left lists the number of annotated example sentences (in the annotation report) illustrating the individual valence patterns. The rows represent so-called frame element configurations together with their syntactic realizations in terms of phrase type and grammatical function. For example, the third frame element configuration from the top lists the FEs GOODS, MANNER, and PERPETRATOR. The GOODS are realized syntactically as a NP Object, the MANNER as a dependent ADVP, and the PERPETRATOR as an external NP. Such systematic valence tables allow researchers to gain a better understanding of how the semantics of frames are realized syntactically.⁵

3. FrameNets for other languages

3.1. Similarities and differences

Following the success of the Berkeley FrameNet for English, a number of FrameNets for other languages were developed over the past ten years. Based on ideas outlined in Heid (1996), Fontenelle (1997), and Boas (2001/2002/2005a), researchers aimed to create parallel FrameNets by re-using frames constructed by the Berkeley FrameNet project for English. While FrameNets for other languages aim to re-use English FrameNet frames to the greatest extent possible, they differ in a number of important points from the original FrameNet (see Boas, 2009).

For example, projects such as SALSA (Burchardt et al., 2009) aim to create full-text annotation of an entire German corpus instead of finding isolated corpus sentences to identify lexicographically relevant information as is the case with the Berkeley FrameNet and Spanish FrameNet (Subirats, 2009). FrameNets for other languages also differ in what types of resources

they use as data pools. That is, besides exploiting a monolingual corpus as is the case with Japanese FrameNet (Ohara, 2009) or Hebrew FrameNet (Petrucci, 2009), projects such as French FrameNet (Pitel, 2009) or BiFrameNet (Fung & Chen, 2004) also employ multilingual corpora and other existing lexical resources. Another difference concerns the tools used for data extraction and annotation. While the Japanese and Spanish FrameNets adopted the Berkeley FrameNet software (Baker et al., 2003) with slight modifications, other projects such as SALSA developed their own tools to conduct semi-automatic annotation on top of existing syntactic annotations found in the TIGER corpus, or they integrate off-the shelf software as is the case with French FrameNet or Hebrew FrameNet. FrameNets for other languages also differ in the methodology used to produce parallel lexicon fragments. While German FrameNet (Boas, 2002) and Japanese FrameNet (Ohara, 2009) rely on manual annotations, French FrameNet and BiFrameNet use semi-automatic and automatic approaches to create parallel lexicon fragments for French and Chinese. Finally, FrameNets for other languages also differ in their semantic domains and the goals they pursue. While most non-English FrameNets aim to create databases with broad coverage, other projects focus on specific lexical domains such as football (a.k.a. soccer) language (Schmidt, 2009) or the language of criminal justice (Bertoldi et al., 2010). Finally, while the data from almost all non-English FrameNets are intended to be used by a variety of audiences, Multi FrameNet⁶ is intended to support vocabulary acquisition in the foreign language classroom (see Atzler, 2011).

3.2. Re-using (English) semantic frames

To exemplify how English FrameNet frames can be re-used for the creation of parallel lexicon fragments consider Boas' (2005a) discussion of the English verb *answer* evoking the *Communication_Response* frame and its counterpart *responder* in Spanish FrameNet. The basic idea is that since the two verbs are translation equivalents they should evoke the same semantic frame, which should in turn be used as a common structuring device for combining the respective

⁵For details about the different phrase types and grammatical functions, including the different types of null instantiation (CNI, DNI, and INI) (Fillmore 1986), see Fillmore et al. 2003b, Boas 2009, Fillmore & Baker 2010, and Ruppenhofer et al. 2010.

⁶<http://www.coerll.utexas.edu/coerll/taxonomy/term/627>

English and Spanish lexicon fragments. Since the MySQL databases representing each of the non-English FrameNets are similar in structure to the English MySQL database in that they share the same type of conceptual backbone (i.e., the semantic frames and frame relations), this step involves determining which English LUs are equivalent to corresponding non-English LUs.

However, before creating parallel lexicon fragments for Spanish and linking them to their English counterparts via their semantic frame it is necessary to first conduct a detailed comparison of the individual LUs and how they realize the semantics of the frame. To begin, consider the different ways in which the FEs of the *Communication_response* frame are realized with *answer*.

FE Name	Syntactic Realization
SPEAKER	NP.Ext, PP_ <i>by</i> .Comp, CNI
MESSAGE	INI, NP.Obj, PP_ <i>with</i> .Comp, QUO.Comp, Sfin.Comp
ADDRESSEE	DNI
DEPICTIVE	PP_ <i>with</i> .Comp
MANNER	AVP.Comp, PPing_ <i>without</i> .Comp
MEANS	PPing_ <i>by</i> .Comp
MEDIUM	PP_ <i>by</i> .Comp, PP_ <i>in</i> .Comp, PP_ <i>over</i> .Comp
TRIGGER	NP.Ext, DNI, NP.Obj, Swh.Comp

Table 1: Partial realization table for the verb *answer* (Boas 2005a)

Table 1 shows that there is a significant amount of variation in how FEs of the *Communication_Response* frame are realized with *answer*. For example, the FE *DEPICTIVE* has only one option for its syntactic realization, i.e. a PP complement headed by *with*. Other FEs such as *SPEAKER* and *MANNER* exhibit more flexibility in how the FEs of the frame are realized

syntactically while yet another set of FEs such as *MESSAGE* and *TRIGGER* exhibit the highest degree of syntactic variation. Now that we know the full range of how the FEs of the *Communication_Response* frame are realized syntactically with *answer* we can take the next step towards creating a parallel lexical entry for its Spanish counterpart *responder*.

This step involves the use of bilingual dictionaries and parallel corpora in order to identify possible Spanish translation equivalents of *answer*. While this procedure may seem trivial, it is a rather lengthy and complicated process because it is necessary to consider the full range of valence patterns (the combination of FEs and their syntactic realizations) of the English LU *answer* listed in FrameNet. It lists a total of 22 different frame element configurations, totaling 32 different combinations in which these sequences may be realized syntactically. As the full valence table for *answer* is rather long we focus on only one out of the 22 frame element configurations, namely that of *SPEAKER* (Sp), *MESSAGE* (M), *TRIGGER* (Tr), and *ADDRESSEE* (A) in Table 2.

	Sp	M	Tr	A
a.	NP.Ext	NP.Obj	DNI	DNI
b.	NP.Ext	PP_ <i>with</i> .Comp	DNI	DNI
c.	NP.Ext	QUO.Comp	DNI	DNI
d.	NP.Ext	Sfin.Comp	DNI	DNI

Table 2: Excerpt from the Valence Table for *answer* (Boas 2005a)

As Table 2 shows, the frame element configuration exhibits a certain amount of variation in how the FEs are realized syntactically: All four valence patterns have the FE *SPEAKER* realized as an external noun phrase, and the FEs *TRIGGER* and *ADDRESSEE* not realized overtly at the syntactic level, but null instantiated as Definite Null Instantiation (DNI). In other words, in sentences such as *He answered with another question* the FEs *TRIGGER* and *ADDRESSEE* are understood in context although they are not realized syntactically.

With the English-specific information about *answer* and the more general frame information in place we are now

in a position to search for the corresponding frame element configuration of its Spanish translation equivalent *responder*. Taking a look at the lexical entry of *responder* in Spanish FrameNet we see that the variation of syntactic realizations of FEs is similar to that of *answer* in Table 1.

FE Name	Syntactic Realizations
SPEAKER	NP.Ext, NP.Dobj, CNI, PP_por.COMP
MESSAGE	AVP.AObj, DNI, QUO.DObj, queSind.DObj, queSind.Ext
ADDRESSEE	NP.Ext, NP.IObj, PP_a.IObj, DNI, INI
DEPICTIVE	AJP.Comp
MANNER	AVP.AObj, PP_de.AObj
MEANS	VPndo.AObj
MEDIUM	PP_en.AObj
TRIGGER	PP_a.PObj, PP_de.PObj, DNI

Table 3: Partial Realization Table for the verb *responder* (Boas 2005a)

Spanish FrameNet also offers a valence table that includes for *responder* a total of 23 different frame element configurations. Among these, we find a combination of FEs and their syntactic realization that is comparable in structure to that of its English counterpart in Table 2 above.

	Sp	M	Tr	A
a.	NP.Ext	QUO.DObj	DNI	DNI
b.	NP.Ext	QueSind.DObj	DNI	DNI

Table 4: Excerpt from the Valence Table for *responder* (Boas 2005a)

Comparing Tables 2 and 4 we see that *answer* and *responder* exhibit comparable valence combinations with the FEs SPEAKER and MESSAGE realized syntactically while the FEs TRIGGER and ADDRESSEE are not realized syntactically, but are instead implicitly understood (they

are definite null instantiations). With a Spanish counterpart in place it now becomes possible to link the Spanish set of frame element configurations in Table 4 with its English counterpart in Table 2 via the Communication_Response frame as the following Figure illustrates.

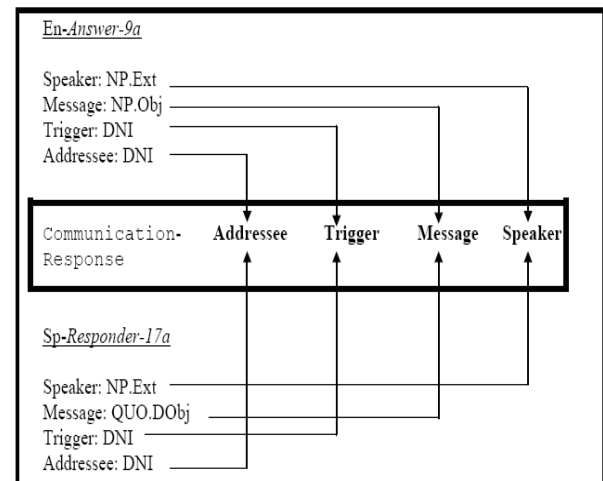


Figure 2: Linking partial English and Spanish lexicon fragments via semantic frames (Boas 2005a)

Figure 5 shows how the lexicon fragments of *answer* and *responder* are linked via the Communication_Response frame. The 'a' index points to the respective first lines in the valence tables of the two LUs (cf. Tables 2 and 4) and identifies the two syntactic frames as being translation equivalents of each other. At the top of Figure 2 we see the verb *answer* with one of its 22 frame element configurations, i.e. SPEAKER, TRIGGER, MESSAGE, and ADDRESSEE. Figure 2 shows for this configuration one possible set of syntactic realizations of these FEs, that given in row (a) in Table 2 above. The 9a designation following *answer* indicates that this lexicon fragment is the ninth configuration of FEs out of a total of 22 frame element configurations listed in the complete realization table. Of the ninth frame element configuration 'a' indicates that it is the first of a list of various possible syntactic realizations of these FEs (there are a total of four, cf. Table 2 above). As already pointed out, the FE SPEAKER is realized syntactically as an external NP, MESSAGE as an object NP, and both TRIGGER and ADDRESSEE are null instantiated.

The bottom of Figure 2 shows *responder* with the first of the 17 frame element configurations (recall that there are a total of 23). For one of these configurations, we see one subset of syntactic realizations of these FEs, namely the first row catalogued by Spanish FrameNet for this configuration (see row (a) in Table 3).

The two parallel lexicon fragments at the top and the bottom of Figure 2 are linked by indexing their specific semantic and syntactic configurations as equivalents within the `Communication_Response` frame. This linking is indicated by the arrows pointing from the top and the bottom of the partial lexical entries to the mid-section in Figure 2, which symbolizes the `Communication_Response` frame at the conceptual level, i.e. without any language-specific specifications. Note that this procedure does not automatically link the entire lexical entries of *answer* and *responder* to each other. Establishing such a correspondence link connects only the relevant frame element configurations and their syntactic realizations in Tables 2 and 4 via the common semantic frame, because they can be regarded as translation equivalents.

Although linking the two lexicon fragments this way results in a systematic way of creating parallel lexicon fragments based on semantic frames (which serve as interlingual representations), it is not yet possible to automatically create or connect such parallel lexicon fragments. This means that one must carefully compare each individual part of the valence table of a LU in the source language with each individual part of the valence table of a LU in the target language. This step is extremely time intensive because it involves a detailed comparison of bilingual dictionaries as well as electronic corpora to ensure matching translation equivalents. Recall that Figure 2 represents only a very small set of the full lexical entries of *answer* and *responder*. The procedure outlined above will have to be repeated for each of the 32 different valence patterns of *answer* – and its (possible) Spanish equivalents. The following section addresses a number of other issues that need to be considered carefully when creating parallel lexicon fragments based on semantic frames.

4. Cross-linguistic problems

Creating parallel lexicon entries for existing English

FrameNet entries and linking them to their English counterparts raises a number of important issues, most of which require careful (manual) linguistic analysis. While some of these issues apply to the creation of parallel entries across the board, others differ depending on the individual languages or the semantic frame. The following subsections, based on Boas (to appear), briefly address some of the most important issues, which all have direct bearing on how the semantics of a frame are realized syntactically across different languages.

4.1. Polysemy and profiling differences

While translation equivalents evoking the same frame are typically taken to describe the same types of scenes, they sometimes differ in how they profile FEs. For example, Boas (2002) discusses differences in how *announce* and various German translation equivalents evoke the `Communication_Statement` frame. When *announce* occurs with the syntactic frame [NP.Ext _ NP.Obj] to realize the `SPEAKER` and `MESSAGE` FEs as in *They announced the birth of their child*, German offers a range of different translation equivalents, including *bekanntgeben*, *bekanntmachen*, *ankündigen*, or *anzeigen*. Each of these German LUs comes with its own specific syntactic frames that express the semantics of the `Communication_Statement` frame. When *announce* is used to describe situations in which a message is communicated via a medium such as a loudspeaker (e.g. *Joe announced the arrival of the pizza over the intercom*), German offers *ansagen* and *durchsagen* as more specific translation equivalents of *announce* besides the more general *ankündigen*. Thus, by providing different LUs German offers the option of profiling particular FEs of the `Communication_Statement` frame, thereby allowing for the representation of subtle meaning differences of the frame and the perspective given of a situation (see Ohara, 2009 on similar profiling differences between English and Japanese LUs evoking the `Risk` frame).

4.2. Differences in lexicalization patterns

Languages differ in how they lexicalize particular types of concepts (see Talmy, 1985), which may directly influence how the semantics of a particular frame are

realized syntactically. For example, in a comparative study of English, Spanish, Japanese, and German motion verbs in *The Hound of the Baskervilles* (and its translations), Ellsworth et al. (2006) find that there are a number of differences in how the various concepts of motion are associated with different types of semantic frames. More specifically, they show that English *return* (cf. *The wagonette was paid off and ordered to return to Coombe Tracey forthwith, while we started to walk to Merripit House*) and Spanish *regresar* both evoke the Return frame, whereas the corresponding German *zurückschicken* evokes the Sending frame. These differences demonstrate that although the concept of motion is incorporated into indirect causation, the frames expressing indirect causation may vary from language to language (see Burchardt et al., 2009 for a discussion of more fine-grained distinctions between verbs evoking the same frame in English and German).

4.3 Polysemy and translation equivalents

Finding proper translation equivalents is typically a difficult task because one has to consider issues surrounding polysemy (Fillmore & Atkins, 2000; Boas, 2002), zero translations (Salkie, 2002; Boas 2005a; Schmidt, 2009), and contextual and stylistic factors (Altenberg & Granger, 2002; Hasegawa et al., 2010), among others. To illustrate, consider Bertoldi's (2010) discussion of contrastive legal terminology in English and Brazilian Portuguese. Based on the English Criminal Process frame (see FrameNet), Bertoldi finds that while there are some straightforward translation equivalents of English LUs in Portuguese, others involve a detailed analysis of the relevant polysemy patterns.

Consider Figure 3, which compares English and Portuguese LUs in the Notification_of_charges frame. The first problem discussed by Bertoldi (2010) addresses the fact that although there are corresponding Portuguese LUs such as *denunciar*, they do not evoke the same semantic frame as the English LUs, but rather a frame that could best be characterized as evoking the Accusation frame. The second problem is that six Portuguese translation equivalents of the English LUs evoking only the Notification_of_charges frame, i.e. *acusar*, *acusação*, *denunciar*,

denuncia, *pronunciar*, and *pronuncia*, potentially evoke three different frames.

English	Portuguese	English
Accuse.v	Acusar	Incriminate; blame; arraign; renounce; accuse; prosecute; charge; indict.
	Denunciar	Denounce; accuse; inform against; report; proclaim.
Charge.n	Acusação	Accusation; charge; incrimination; denunciation; prosecution; indictment.
Charge.v	Acusar	
	Pronunciar	Indict; arraign.
Indict.v	Acusar	
	Denunciar	
Indictment.n	Pronúncia	Indictment; arraignment.

Figure 3: English LUs from the Notification_of_Charges frame and their Portuguese translation equivalents (Bertoldi, 2010: 6)

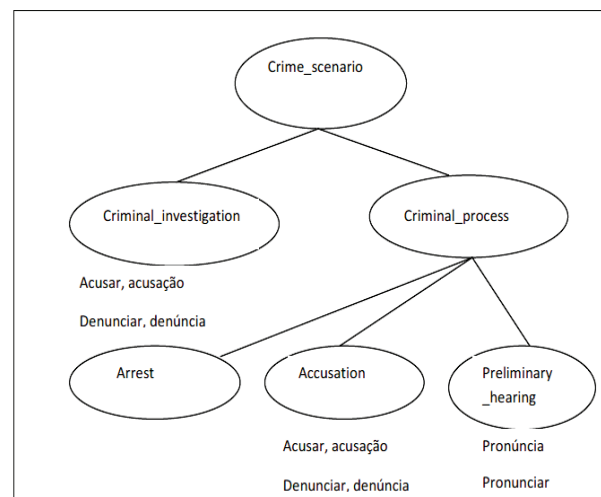


Figure 4: LUs evoking multiple frames in the Portuguese Crime_scenario frame (Bertoldi, 2010:7)

This leads Bertoldi to claim that the LUs *acusar*, *acusação*, *denunciar*, and *denuncia* may evoke two different Criminal_Process sub-frames, besides

other general language, non-legal specific frames, as is illustrated by Figure 4. Bertolid's analysis shows that finding translation equivalents is not always an easy task and that one needs to pay close attention to different polysemy networks across languages, which may sometimes be influenced by systematic differences such as differences between legal systems.

4.4 Universal frames?

Claims about the universality of certain linguistic features are abundant in the literature. When it comes to semantic frames the question is whether frames derived on the basis of English are applicable to the description and analysis of other languages (and vice versa). While a number of studies on motion verbs (Fillmore & Atkins, 2000; Boas, 2002; Burchardt et al., 2009; Ohara, 2009) and communication verbs (Boas, 2005a; Subirats, 2009), among other semantic domains, suggest that there are frames that can be re-used for the description and analysis of other languages, there also seem to be culture-specific frames that may not be re-usable without significant modification.

One set of examples comes from the English *Personal_Relationship* frame, whose semantics appears to be quite culture-specific. Atzler (2011) shows that concepts such as dating (*to date*) seem to be quite specific to Anglo culture and may not be directly applicable to the description of similar activities in German. Another, perhaps more extreme example, is the term *sugar daddy*, which has no exact counterpart in German, but instead requires a lengthy paraphrase in German to render the concept of this particular type of relationship in German.

A second example comes from the intransitive Finnish verb *saunoa* (literally 'to sauna'), which has no direct English counterpart because it is very culture-specific, and in effect evokes a particular type of frame. To this end, Leino (2010:131) claims that this verb (and correspondingly the Finnish *Sauna* frame) "expresses a situation in which the referent of the subject goes to the sauna, is in the sauna, participates in the sauna event, or something of the like." Dealing with such culture-specific frames thus requires quite lengthy paraphrases to arrive at an approximation of the semantics of the frame in English.

5. Conclusions and outlook

This paper has outlined some of the basic steps underlying the creation of parallel lexicon fragments. Employing semantic frames for this purpose is still a work in progress, but the successful compilation of several FrameNets for languages other than English is a good indication that this methodology should be pursued further.

Clearly, the problems outlined in the previous section need to be solved. The first problem, polysemy and profiling differences, is perhaps the most daunting one. Decades of linguistic research into these issues (see, e.g. Leacock & Ravin, 2000; Altenberg & Granger, 2002) seem to suggest that there is no easy solution that could be implemented to arrive at an automatic way of analyzing, comparing, and classifying different polysemy and lexicalization patterns across languages. This means that for now these issues need to be addressed manually, in the form of careful linguistic analysis, in the near future.

The same can be said about the problems surrounding lexicalization patterns, zero translations, and the universality of frames. Without a detailed catalogue of linguistic analyses of these phenomena in different languages, and a comparison across language pairs, any efforts regarding the effective linking of parallel lexicon fragments, whether on the basis of semantic frames or not, will undoubtedly hit many roadblocks.

6. Acknowledgements

Work on this paper was supported by a fellowship for experienced researchers from the Alexander von Humboldt Foundation, as well as by Title VI grant #P229A100014 (Center for Open Educational Resources and Language Learning) to the University of Texas at Austin.

7. References

- Altenberg, B., Granger, S. (2002): Recent trends in cross-linguistic studies. In B. Altenberg & S. Granger (Eds.), *Lexis in Contrast*. Amsterdam/Philadelphia: John Benjamins, pp. 3-50.
- Atkins, B.T.S., Fillmore, C.J., Johnson, C.R. (2003): *Lexicographic relevance: Selecting information from*

- corpus evidence. *International Journal of Lexicography*, 16(3), pp. 251-280.
- Atzler, J. (2011): *Twist in the line: Frame Semantics as a vocabulary teaching and learning tool*. Doctoral Dissertation, The University of Texas at Austin.
- Baker, C.F., Fillmore, C.J., Lowe, J.B. (1998): The Berkeley FrameNet Project. In *COLING-ACL '98: Proceedings of the Conference*, pp. 86-90.
- Baker, C.F., Fillmore, C.J., Cronin, B. (2003): The Structure of the FrameNet Database. In *International Journal of Lexicography*, 16(3), pp. 281-296.
- Bertoldi, A. (2010): *When translation equivalents do not find meaning equivalence: a contrastive study of the frame Criminal_Process*. Manuscript. UT Austin.
- Bertoldi, A., Chishman, R., Boas, H.C. (2010): Verbs of judgment in English and Portuguese: What contrastive analysis can say about Frame Semantics. *Calidoscopio*, 8 (3), pp. 210-225.
- Boas, H.C. (2001): Frame Semantics as a framework for describing polysemy and syntactic structures of English and German motion verbs in contrastive computational lexicography. In *Proceedings of Corpus Linguistics 2001*, pp. 64-73.
- Boas, H.C. (2002): Bilingual FrameNet dictionaries for machine translation. In: *Proceedings of the Third International Conference on Language Resources and Evaluation*, Vol. IV, pp. 1364-1371.
- Boas, H.C. (2005a): Semantic frames as interlingual representations for multilingual lexical databases. *International Journal of Lexicography*, 18(4), pp. 445-478.
- Boas, H.C. (2005b): From theory to practice: Frame Semantics and the design of FrameNet. In S. Langer & D. Schnorbusch (Eds.), *Semantik im Lexikon*. Tübingen: Narr, pp. 129-160.
- Boas, H.C. (2009): Recent trends in multilingual lexicography. In H.C. Boas (Ed.), *Multilingual FrameNets in Computational Lexicography: Methods and Applications*. Berlin/New York: Mouton de Gruyter, pp. 1-36.
- Boas, H.C. (to appear): Frame Semantics and Translation. In I. Antunano and A. Lopez (Eds.), *Translation in Cognitive Linguistics*. Berlin/New York: Mouton de Gruyter.
- Burchardt, A., Erk, K., Frank, A., Kowalski, A., Pado, S., & Pinkal, M. (2009): Using FrameNet for the semantic analysis of German: annotation, representation, and automation. In H.C. Boas (Ed.), *Multilingual FrameNets in Computational Lexicography: Methods and Applications*. Berlin/New York: Mouton de Gruyter, pp. 209-244.
- Dux, R. (2011): *A frame-semantic analysis of five English verbs evoking the Theft frame*. M.A. Report. The University of Texas at Austin.
- Ellsworth, M., Ohara, K., Subirats, C., & Schmidt, T. (2006): *Frame-semantic analysis of motion scenarios in English, German, Spanish, and Japanese*. Presentation given at the 4th International Conference on Construction Grammar (ICCG-4), Tokyo.
- Fillmore, C.J. (1982): Frame Semantics. In *Linguistic Society of Korea (Ed.), Linguistics in the Morning Calm*. Seoul: Hanshin, pp. 111-138.
- Fillmore, C.J. (1985): Frames and the semantics of understanding. *Quadernie di Semantica*, 6, pp. 222-254.
- Fillmore, C.J. (2006): Pragmatically controlled zero anaphora. *BLS*, 12, pp. 95-107.
- Fillmore, C.J. (2007): Valency issues in FrameNet. In: T. Herbst & K. Götz-Vetteler (Eds.), *Valency: theoretical, descriptive, and cognitive issues*. Berlin/New York: Mouton de Gruyter, pp. 129-160.
- Fillmore, C.J., Atkins, B.T.S. (2000): Describing polysemy: The case of 'crawl'. In Y. Ravin and C. Lacock (Eds.), *Polysemy*. Oxford: Oxford University Press, pp. 91-110.
- Fillmore, C.J., Baker, C.F. (2010): A frames approach to semantic analysis. In: B. Heine and H. Narrog (Eds.), *The Oxford Handbook of Linguistic Analysis*. Oxford: Oxford University Press, pp. 313-340.
- Fillmore, C.J., Petruck, M.R.L. (2003): FrameNet Glossary. *International Journal of Lexicography*, 16(3), pp. 359-361.
- Fillmore, C.J., Johnson, C.R., Petruck, M.R.L. (2003a): Background to FrameNet. *International Journal of Lexicography*, 16(3), pp. 235-251.
- Fillmore, C.J., Petruck, M.R.L., Ruppenhofer, J., Wright, A. (2003b): FrameNet in action: The case of Attaching. *International Journal of Lexicography*, 16(3), pp. 297-333.

- Fontenelle, T. (1997): Using a bilingual dictionary to create semantic networks. *International Journal of Lexicography*, 10(4), pp. 275-303.
- Fung, P., Chen, B. (2004): BiFrameNet: Bilingual frame semantics resource construction by cross-lingual induction. In *Proceedings of COLING 2004*.
- Hasegawa, Y., Lee-Goldman, R., Ohara, K., Fujii, S., Fillmore, C.J. (2010): On expressing measurement and comparison in English and Japanese. In H.C. Boas (Ed.), *Contrastive Studies in Construction Grammar*. Amsterdam/Philadelphia: John Benjamins, pp. 169-200.
- Heid, U. (1996): Creating a multilingual data collection for bilingual lexicography from parallel monolingual lexicons. In *Proceedings of the VIIth EURALEX International Congress*, pp. 559-573.
- Leino, J. (2010): Results, cases, and constructions: Argument structure constructions in English and Finnish. In H.C. Boas (Ed.), *Contrastive Studies in Construction Grammar*. Amsterdam/Philadelphia: John Benjamins, pp. 103-136.
- Lowe, J.B., Baker, C.F., Fillmore, C.J. (1997): A frame-semantic approach to semantic annotation. In *Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?* Held April 4-5, in Washington, D.C., in conjunction with ANLP-97.
- Ohara, K. (2009): Frame-based contrastive lexical semantics in Japanese FrameNet: The case of risk and kakeru. In H.C. Boas (Ed.), *Multilingual FrameNets in Computational Lexicography: Methods and Applications*. Berlin/New York: Mouton de Gruyter, pp. 163-182.
- Petruck, M.R.L. (1996): Frame Semantics. In J. Verschueren, J.-O. Östman, J. Blommaert, C. Bulcaen (Eds.), *Handbook of Pragmatics*. Amsterdam/Philadelphia: John Benjamins, pp. 1-13.
- Petruck, M.R.L. (2009): Typological considerations in constructing a Hebrew FrameNet. In H.C. Boas (Ed.), *Multilingual FrameNets in Computational Lexicography: Methods and Applications*. Berlin/New York: Mouton de Gruyter, pp. 183-208.
- Petruck, M.R.L., Boas, H.C. (2003): All in a day's week. In . Hajicova, A. Kotesovcova, J. Mirovsky (Eds.), *Proceedings of CIL 17*. CD-ROM. Prague: Matfyzpress.
- Petruck, M.R.L., Fillmore, C.J., Baker, C.F., Ellsworth, M., Ruppenhofer, J. (2004): Reframing FrameNet data. In *Proceedings of the 11th EURALEX International Conference*, pp. 405-416.
- Pitel, G. (2009): Cross-lingual labeling of semantic predicates and roles: A low-resource method based on bilingual L(atent) S(emantic) A(nalysis). In H.C. Boas (Ed.), *Multilingual FrameNets in Computational Lexicography: Methods and Applications*. Berlin/New York: Mouton de Gruyter, pp. 245-286.
- Ruppenhofer, J., Ellsworth, M., Petruck, M.R.L., Johnson, C., Scheffczyk, J. (2010): *FrameNet II: Extended theory and practice*. Available at <http://framenet.icsi.berkeley.edu>
- Salkie, R. (2002): Two types of translation equivalence. In B. Altenberg and S. Granger (Eds.), *Lexis in Contrast*. Amsterdam/Philadelphia: John Benjamins, pp. 51-72.
- Schmidt, T. (2009): The Kicktionary – A multilingual lexical resource of football language. In H.C. Boas (Ed.), *Multilingual FrameNets in Computational Lexicography: Methods and Applications*. Berlin/New York: Mouton de Gruyter, pp. 101-134.
- Subirats, C. (2009): Spanish FrameNet: A frame-semantic analysis of the Spanish lexicon. In H.C. Boas (Ed.), *Multilingual FrameNets in Computational Lexicography: Methods and Applications*. Berlin/New York: Mouton de Gruyter, pp. 135-162.
- Talmy, L. (1985): Lexicalization patterns: semantic structures in lexical forms. In T. Shopen (Ed.), *Language Typology and Syntactic Description*. Cambridge: Cambridge University Press, pp. 57-149.
- Ziem, A. (2008): *Frames und sprachliches Wissen*. Berlin/New York: Mouton de Gruyter.

The Multilingual Web: Opportunities, Borders and Visions

Felix Sasaki

DFKI, LT-Lab / Univ. of Applied Sciences Potsdam

Alt-Moabit 91c, 10559 Berlin

E-mail: felix.sasaki@dfki.de

Abstract

The Web is growing more and more in languages other than English, leading to the opportunity of a truly multilingual, global information space. However, the full potential of multilingual information creation and access across language borders has not yet been developed. We report about a workshop series and project called “MultilingualWeb” that aims at analyzing borders or “gaps” within Web technology standardization hindering multilinguality on the Web. MultilingualWeb targets at scientific, industrial and user communities who need to collaborate closely. We conclude with a first concrete outcome of MultilingualWeb: the upcoming formation of a cross-community, W3C standardization activity that will close some of the gaps that already have been recognized.

Keywords: Multilinguality, Web, standardization, language technology, metadata

1. Introduction: Missing Links between Languages on the Web

A recent blog post discussed “Languages of the World (Wide Web)”¹. Via impressive visualizations, it showed the amount of content per language and number of links between languages. By no surprise English is a dominant language on the Web, and every other language has a certain number of links to English web pages. Nevertheless, the amount of content in many other languages is continuously and rapidly growing. Unfortunately, the links between these languages and links to English are rather few.

What does this mean? First, it demonstrates that English is a lingua franca on the Web. Users who are not capable or willing to use this lingua franca cannot communicate with others and are not part of the global information society; they are residents of local silos on the Web. Second, the desire to communicate in one’s own language is high and is growing.

Several issues need to be resolved to tear down the walls between language communities on the web. One key issue is the availability of standardized technologies to create content in your own language, and to access

content across languages. The need to resolve this issue led to the creation of the “MultilingualWeb” project.

2. MultilingualWeb: Overview

MultilingualWeb <http://www.multilingualweb.eu/> is an EU-funded thematic network project exploring standards and best practices that support the creation, localization and use of multilingual web-based information. It is lead by the World Wide Web Consortium (W3C), the major stakeholder for creating the technological building blocks of the web. MultilingualWeb encompasses 22 partners <http://www.multilingualweb.eu/partners>, both from research and various industries, related to content creation, localization, various software providers etc. The project main part is a series of four public workshops, to discuss what standards and best practices currently exist, and what gaps need to be filled. The project started in April 2010; as of writing, two workshops have been held. They have been of enormous success, in terms of the number of participants, awareness esp. in social media, and the outcome of discussions. In the reminder of this abstract, we will discuss current findings² of the project and will take a look at what the two upcoming workshops and future projects might bring.

¹ See

<http://googleresearch.blogspot.com/2011/07/languages-of-world-wide-web.html>

² More details on the findings can be found in workshop reports on the project website <http://www.multilingualweb.eu>.

3. About Terminology and Communities

One gap is related to the communities, industry and technology stacks that need to be aware of standards related to the multilingual Web. *Internationalization* deals with the prerequisites to create content in many languages. This involves technologies and standards related to character encoding, language identification, font selection etc. The proper internationalization of (web) technologies is required for *localization*: the adaptation to local markets and cultures. Localization often involves translation. With more and more content that needs to be translated and a growing number of target languages, the use of *language technologies* (e.g. machine translation) comes into play.

A huge success of the MultilingualWeb project is that major stakeholders from the areas of internationalization, localization and language technologies have been brought together. This is important since both in terms of research and industry projects, so far the communities do not overlap. The same is true for conference series; see e.g. the (non) overlap of attendees at Localization World, LREC or the Unicode conferences.

4. Workshop Sessions and Topics

MultilingualWeb provides a common umbrella for these communities via a set of labels used for the workshop sessions:

- Developers provide the basic technological building blocks (e.g. browsers) for multilingual web content creation.
- Creators use the technologies to create multilingual content.
- Localizers adapt the content to regions and cultures.
- Machines are used to support multilingual content creation and access, e.g. via machine translation.
- Users more and more do not only consume content, but also at the same time become contributors - see e.g. growing number of users in social networks.
- Policy makers decide about strategies for fostering multilinguality on the Web. They play an important role in big multinational companies, regional or international governmental or non-governmental bodies, standardization bodies etc.

Of course the above labels serve only as a rough orientation. But esp. for the detection of gaps (see below) they have proven to be quite useful. The following

subsections provide a brief summary of some outcomes ordered via these labels and based on the workshop reports. For further details the reader may be referred to these reports.

4.1. Developers

Developers are providing the technological building blocks that are needed for multilingual content creation and content access on the web. Many of these building blocks are still under development, and web browsers play a crucial role. During the workshops, many presentations dealt with enhancement of characters and fonts support, locale data formats, internationalized domain names and typographic support.

Gaps in this area are also related to handling of translations: although more and more web content is being translated, the key web technology HTML so far has no means to support this process. Here it is important that the need for such means is being brought to the standard development organizations, namely W3C, and esp. to the browser implementers.

Another gap is what technology stacks are being developed, and how content providers are actually adopting them. HTML5 plays a crucial role in the future of web technology development, but for many content providers its relation to other parts of the technology eco system is not clear yet.

4.2. Creators

Creators more and more need to bring content to many, esp. mobile devices. Since these devices lack computing power, many aspects of multilinguality (e.g. usage of large fonts) need to be taken care of in a specific manner. "Content" does not only mean text. It also encompasses information for multimodal and voice applications, or SMS, especially in developing countries. Navigation of content esp. across languages is another area without standardized approaches or best practices.

Like in the developer area, translation is important for content creation too. There is no standardized way to identify non-translatable content, to identify tools used for translation, translation quality etc.

4.3. Localizers

Localizers deal with the process of localization, which involves many aspects: content creation, the distribution

of content to language service providers, further distribution to translators, etc. To organize this process there is a need to improve standards and better integrate them. Metadata plays a crucial role in this respect, as we will discuss later.

Content itself is becoming more complex and fast changing - and localization approaches need to be adapted accordingly. In the area of localization, many standards have been developed: for the representation of content in the translation process, for terminology management, translation memories etc. The gap here is to understand how the standards interplay. This is not an easy task, since sometimes there are competing technologies available. Hence, currently there are quite a few initiatives dedicated to interoperability in the localization area, including the integration with web content creation and machine translation.

4.4. Machines

For machines, that is applications based on language technology, the need for standardization esp. related to metadata and the localization process is of outmost importance. Language resources are crucial in this area, including their standardized representation and means to share resources. The META-SHARE infrastructure currently being developed is expected to play an important role in this area.

While discussing developers, creators and localizers, machine translation has been mentioned already. It has become clear that a close integration of machine translation technologies to these areas is a major requirement for the better translation quality.

Machines play a crucial role in building bridges between smaller and larger languages, and to change the picture about “languages on the web” that we mentioned at the beginning of this paper.

4.5. Users

Users normally have no strong voice in the development of multilingual or other technologies. At the MultilingualWeb workshops, it became clear that the worldwide interest in multilingual content is high, but significant organizational and technical challenges need to be approached for reaching people in continents such as Africa and Asia.

Multilingual social media are becoming more important

and can be supported by language technology applications like on-the-fly machine translation. However it is important to have a clear border between controlled and uncontrolled environments of content creation. Only in this way the right tools can be chosen to achieve high quality translation of small amounts of text, versus gist translation for larger text bodies.

4.6. Policy Makers

The topic of policy makers was not discussed as a separate session in the first workshop, but only in the 2nd one. Nevertheless it is of high importance: many gaps related to the multilingual web are not technical ones, but are related to e.g. political decisions about the adoption of standards. Esp. in the localization and language technology area, proprietary solutions prevailed for a long time. Here we are ahead of a radical change, and MultilingualWeb will play a crucial role in bringing the right people together.

Some technological pieces have a lot of political aspects. The META-SHARE infrastructure mentioned before is a good example. A key aspect of this infrastructure is the licensing model it will provide, since not everybody will be willing to share language resources for free.

5. Metadata for Language Related Technology in the Web

5.1. Introduction

After the broad overview of various gaps that have been detected, we will now dive deeper into gaps related to metadata. All communities we mentioned before already for a while have used such metadata:

- in internationalization, metadata is used to identify character encoding or language;
- in localization, metadata helps to organize the localization workflow, e.g. to identify parts of content that need to be translated;
- in language technology, metadata helps as a heuristic to complement language technology applications.

Such heuristics can be useful for the language technology application of automatic detection of the language of content. The heuristic here can be e.g. the language identifier given in a web page. However, to be able to judge its reliability, it is important that many stakeholders work together and that there are stable bridges between internationalization, localization and language

technology. As one concrete outcome of the MultilingualWeb project, a project has been prepared that will work on creating these bridges. The basic project idea is summarized below.

5.2. Three gaps related to Metadata

Language technology applications (machine translation, automatic summarization, cross-language information retrieval, automatic quality assurance etc.) and resources (grammars, translation memories, corpora, lexica etc.) are increasingly becoming available on the web and integrated into HTML and Web based content and accessible via web applications and web service APIs.

This approach has partially been successful in fostering interoperability between language technology resources and applications. However, it lacks the integration with the “Open Web Platform”, i.e.: with the growing set of technologies used for creating and consuming the Web in many applications, on many devices, for many (and more and more) users.

From the view of this current platform, language technology is a black box: Services like online machine translation receive textual input, and produce some output. The end users have no means to adjust language technology to their needs, and they are not able to influence language technology based processes in detail. On the other hand, providers of language technology face difficulties in adapting to specific demands by users in a timely and cost-effective manner, which is a problem also experienced by Language Service Providers as they increasingly adopt language technologies.

To address the “black box” problem, three gaps that have been detected during the MultilingualWeb workshops need to be filled. They play a role in the chain of multilingual content processing and consumption on the Web:

- An online machine translation service might make mistakes like translation of fixed terminology or named entities. This demonstrates gap no. 1: language technology does not know about metadata in the source content, e.g. “What parts of the input should be translated?”
- In the database from which the translated text has been generated, the information about translatability might have been available. However, the machine translation service does not know about that kind of

“hidden Web” information. This reveals gap no. 2: there is no description of the processes available, which were the basis for generating “surface Web” pages.

- The last gap no. 3 is about a standardized approach for identification. This means first that identification of information to fill the gaps 1 and 2 is so far not described in a standardized manner. For example, there is no commonly identified translate flag available in core web technologies like HTML. Second, it means that so far resources used by language technology applications (e.g. “what lexicon is used for machine translation?”) and the applications themselves (e.g. “general purpose machine translation versus application tailored towards a specific domain”) cannot be identified uniquely. This hinders the ad hoc creation of language technology applications on the Web, i.e. the re-combination of resources and application modules.

5.3. Addressing the Gaps: MultilingualWeb-LT

To close the gaps mentioned above, a project called MultilingualWeb-LT has been formed that is planned to start in early 2012. The consortium of MultilingualWeb-LT consists of 14 partners from the areas of CMS systems, localization service providers, language technology industry and research etc. As the forum of work gaps, the project will start a working group within W3C.

The goal of MultilingualWeb-LT is to define a standard that fills the gaps, including three mostly open source reference implementations around three topic areas, in which metadata is being used:

- Integration of CMS and Localization Chain. Modules for the Drupal CMS system will be built that support the creation of the metadata. The metadata will then be taken up in web-based tools that support the localization chain: from the process of gathering of localizable content, the distribution to translators, to the re-aggregation of the results into localized output.
- Online MT Systems. MT systems will be made aware of the metadata, which will lead to more satisfactory translation results. An online MT system

will be made sensitive to the outputs of the modified CMS described above.

- **MT Training.** Metadata aware tools for training MT systems will be built. Again these are closely related to CMS that produce the necessary metadata. They will lead to better quality for MT training corpora harvested on the Web.

The above description shows that CMS systems play a crucial role in MultilingualWeb-LT. The usage of language identifiers for deciding about the language of content (see sec. 4) can be enhanced e.g. by the MT training module mentioned above. However, since MultilingualWeb-LT will be a regular W3C working group, other W3C member organizations might join that group. This is highly desired, hoping not only that further implementations will be built, but also that consensus about and usage of the metadata stretches out to the web community.

5.4. MultilingualWeb-LT: Already a Success Story

Although MultilingualWeb-LT has not started yet, it is already a success story: It is a direct outcome of the MultilingualWeb project and of two other projects that play an important role - among others - for community building in the area of language technology research and industry.

- *FLaReNet* (Fostering Language Resources Network) has developed a common vision for the area of language resources. The FLaReNet “Blueprint of Actions and Infrastructures” is a set of recommendations to support this vision in terms of (technical) infrastructure, R&D, and politics. As part of these recommendations, the task of “putting standards in action” has been described as highly important; MultilingualWeb-LT is a direct implementation of this task.
- *META-NET* is dedicated to fostering the technological foundations of a multilingual European information society, by building a shared vision and strategic research agenda, an open distributed facility for the sharing and exchange of resources (META-SHARE), and by building bridges to relevant neighbouring technology fields. MultilingualWeb-LT is a bridge to support the exchange between the language technology community and the web community at large.

These projects and the formation of MultilingualWeb-LT itself demonstrate that a holistic view prevails, in which the differences between internationalization, localization and language technology mentioned before become of less importance, for the common aim of a truly multilingual web.

6. Upcoming Workshops and the Future

At the time of writing, two workshops are planned for the MultilingualWeb project. A workshop in September 2011 will take place in Ireland. Naturally it will have a focus in localization, since many software related companies in Ireland work on this topic.

The last workshop will take place in Luxembourg in March 2012 and will wrap up the MultilingualWeb project. However, the holistic view of a multilingual web, including the communities of internationalization, localization, language technology and the web community itself, will be put forward using the MultilingualWeb brand. The MultilingualWeb-LT project is one means to carry on that brand. It is the hope of the author that other activities will follow and that cross-community collaboration will become a common place. Only in this way we will be able to tear down language barriers on the web and to achieve a truly global information society.

7. Acknowledgements

This extended abstract has been supported by the European Commission as part of the Competitiveness and Innovation Framework Programme and through ICT PSP Grants: Agreement No. 250500 (MultilingualWeb contract) and 249119 (META-NET T4ME contract).

Combining various text analysis tools for multilingual media monitoring

Ralf Steinberger

European Commission – Joint Research Centre (JRC)

21027 Ispra (VA), Italy

E-mail: Ralf.Steinberger@jrc.ec.europa.eu, URL: <http://langtech.jrc.ec.europa.eu/>

Abstract

There is ample evidence that information contained in media reports is complementary across countries and languages. This holds both for facts and for opinions. Monitoring multilingual and multinational media therefore gives a more complete picture of the world than monitoring the media of only one language, even if it is a world language like English. Wide coverage and highly multilingual text processing is thus important. The JRC-developed *Europe Media Monitor* (EMM) family of applications gathers about 100,000 media reports per day in 50 languages from the internet, groups related articles, classifies them, detects and follows trends, produces statistics and issues automatic alerts. For a subset of 20 languages, it also extracts and disambiguates entities (persons, organisations and locations) and reported speech, links related news over time and across languages, gathers historical information about entities and produces various types of social networks. More recent R&D efforts focus on event scenario template filling, opinion mining, multi-document summarisation, and machine translation. This extended abstract gives an overview of EMM from a functionality point of view rather than providing technical detail.

Keywords: news analysis; multilingual; automatic alerting; text mining; information extraction.

1. EMM: Background and Objectives

The JRC with its 2700 employees working in five different European locations in a wide variety of scientific-technical fields is a Directorate General of the European Commission (EC). It is thus a governmental body free of national interests and without commercial objectives. Its main mandate is to provide scientific advice and technical know-how to European Union (EU) institutions and its international partners, as well as to EU member state organisations, with the purpose of supporting a wide range of EU policies. Lowering the language barrier in order to increase European integration and competitiveness is a declared EU objective.

The JRC-developed *Europe Media Monitor* (EMM) is a publicly accessible family of four news gathering and analysis applications consisting of *NewsBrief*, the *Medical Information System MedISys*, *NewsExplorer* and *EMM-Labs*. They are accessible via the single URL <http://emm.newsbrief.eu/overview.html>. The first EMM website went online in 2002 and it has since been extended and improved continuously. The initial objective was to complement the manual news clipping

services of the EC, by searching for news reports online, categorising them according to user needs, and providing an interface for human moderation (selection and re-organisation of articles; creation of layout to print in-house newspapers). EMM users thus typically have a specific information need and want to be informed about any media reports concerning their subject of interest. Monitoring the media for events that are dangerous to the public health (PH) is a typical example. EMM thus continuously gathers news from the web, automatically selects PH-related news items (e.g. on chemical, biological, radiological and nuclear (CBRN) threats including disease outbreaks, natural disasters and more), presents the information on targeted web pages, detects unexpected information spikes and alerts users about them. In addition to PH, EMM categories cover a very wide range of further subject areas, including the environment, politics, finance, security, various scientific and policy areas, general information on all countries of the globe, etc. For an overview of EMM, see Steinberger et al. (2009).

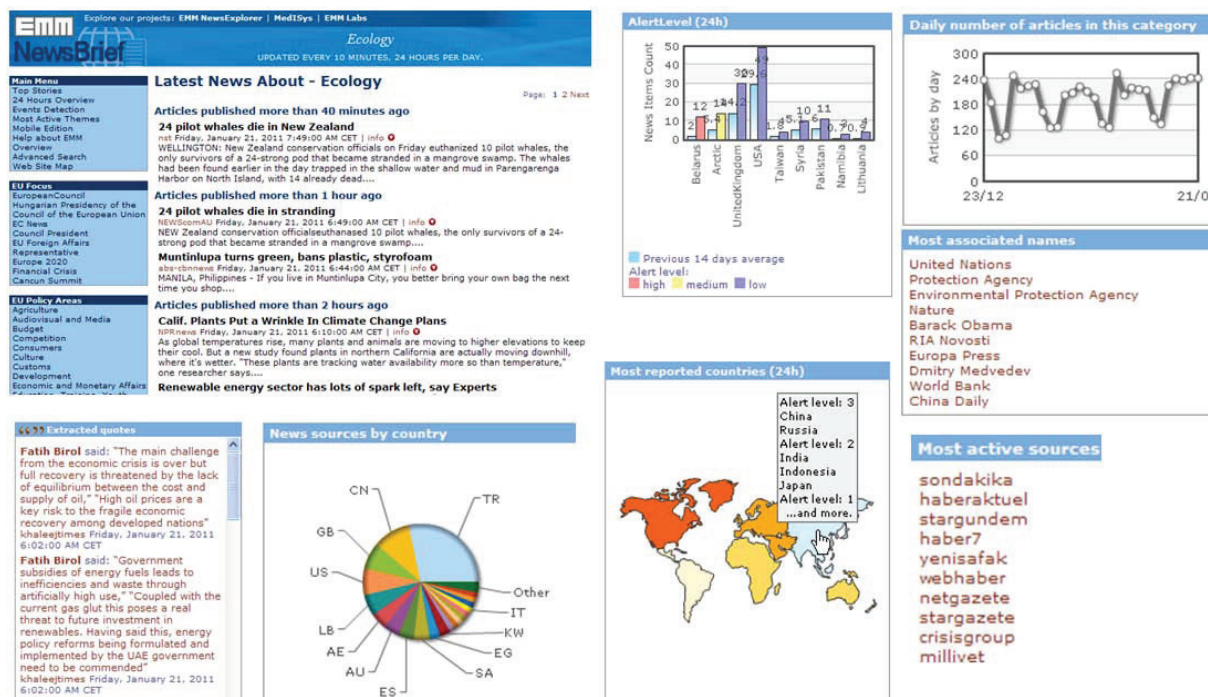


Figure 1. Various aggregated statistics and graphs showing category-based information for one category (ECOLOGY) derived from reports in multiple languages.

2. Information complementarity across languages and countries; news bias

While national EMM clients are mostly interested in the news of their own country and that of surrounding countries (e.g. for disease outbreak monitoring), they also need to follow mass gatherings (e.g. for religious, sport-related or political reasons) because participants may bring back diseases. In addition to the news in the 23 official EU languages, EMM thus also monitors news in Arabic, Chinese, Croatian, Farsi, Swahili, etc., to mention just a few of the altogether 50 languages. While major events such as wars or natural disasters are usually well-covered in world languages such as English, French and Spanish, many small events are typically only mentioned in the national or even in regional press. For instance, disease outbreaks, small-scale violent events and accidents, fraud cases, etc. are usually not reported outside the national borders. The study by Piskorski et al. (2011) comparing media reports in six languages showed that only 51 out of 523 events (of the event types violence, natural disasters and man-made disasters) were reported in more than one language. 350 out of the 523 events were found in non-English news.

Due to this information complementarity across languages and countries, it is crucial that monitoring systems like EMM process texts in many different languages. Using Machine Translation (MT) into one language (usually English) and filtering the news in that language is only a partial solution because specialist terms and names are often badly translated. The benefits of processing texts in the original language was also formulated by Larkey et al. (2004) in their *native language hypothesis*.

We observed the following benefits of applying multilingual text mining tools:

- 1) Different languages cover different geographical areas of the world, for specific subject areas as well as generally. EMM-NewsBrief's news clouds (see <http://emm.newsbrief.eu/geo?type=cluster&format=html&language=all>) show this clearly.
- 2) More information on entities (persons and organisations; see NewsExplorer entity pages) can be extracted from multilingual text. This is due to different contents found, but also to varying linguistic coverage of the text mining software.
- 3) Many more named entity variant spellings (including across scripts) are found when analysing different languages (see NewsExplorer entity pages). These

variant spellings can then be used for improved retrieval, for generating multilingual social networks, and more.

- 4) News bias – regarding the choice of facts as well as the expression of opinions – will be reduced by looking at the media coming from different countries. News bias becomes visually evident when looking at automatically generated social networks (see, e.g. Pouliquen et al., 2007, and Tanev, 2007). For instance, mentions of national politicians are usually preferred in national news, resulting in an inflated view of the importance of one's own political leaders.

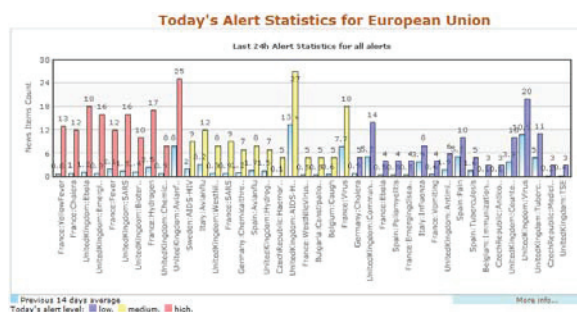
From the point of view of an organisation with a close relationship to many international users, there is thus no doubt that highly multilingual text mining applications are necessary and useful.

3. Ways to harness the benefits of multilinguality

Extracting information from multilingual media reports and merging the information into a single view is possible, but developing text mining tools for each of the languages costs effort and is time-consuming. However, there are various ways to limit the effort per language (for an overview of documented methods, see Steinberger, forthcoming). Some monitoring and automatic alerting functionality can even be achieved with relatively simple means. This section summarises the main multilingual media monitoring functionality provided by the EMM family of applications.

3.1. Multilingual category alerting

EMM categorises all incoming news items into over



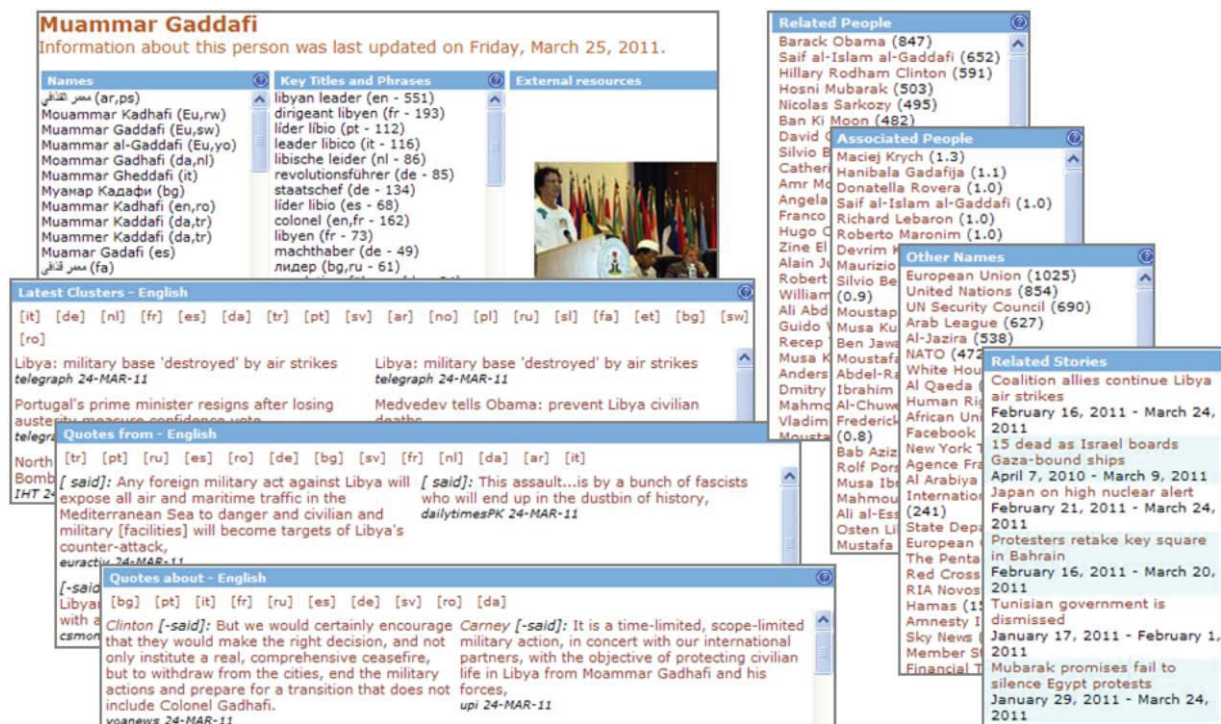


Figure 4. Information automatically gathered over time by EMM-NewsExplorer from media reports in twenty or more languages on one named entity.

The software additionally tracks related news over time, produces time lines and displays extracted meta-information about the news event. For details about the linking of related news items across languages and over time, see Pouliquen et al. (2008).

3.3. Multilingual information gathering on named entities

EMM-NewsExplorer identifies references to person and organisation names in twenty languages. It automatically identifies whether newly found names (within the same script or across different scripts) are simply spelling variants of another name or whether they are new names (for details, see Pouliquen & Steinberger, 2009). The EMM database currently contains up to 400 different automatically collected spellings for the same entity. Any EMM application making use of named entity information uses unique entity identifiers instead of concrete name spellings, allowing to merge information across documents, languages and scripts. The EMM software furthermore keeps track of titles and other expressions found next to the name, keeps statistics on where and when the names were found, and which entities get frequently mentioned together. The latter

information is used to generate social networks that are derived from the international media, thus being independent of national viewpoints. EMM software also detects quotations by and about each entity. The accumulated multilingual results are displayed on the NewsExplorer entity pages (see Figure 4), through which users can explore entities, their relations and related news. Click on any entity name in any of the EMM applications to explore this application.

3.4. Multilingual event scenario template filling

For a smaller subset of currently seven languages, the EMM-NEXUS software extracts structured descriptions of events relevant for global crisis monitoring, such as natural disasters; accidents; violent, medical and humanitarian events, etc. (Tanev et al., 2009; Piskorski et al., 2011). For each news cluster about any such event, the software detects the event type; the event location; the count of dead, wounded, displaced, arrested etc. persons; the perpetrator in the event, as well as the weapons used, if applicable. Contradictory information found in different news articles (such as differing victim counts) are resolved to produce a best guess. The aggregated

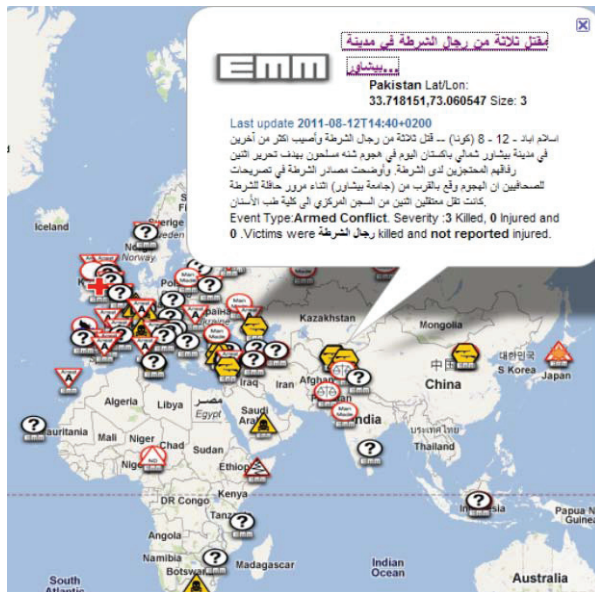


Figure 5. EMM-Labs geographical visualisation of events extracted from media reports in seven languages.

event information is then displayed on *NewsBrief* (in text form) and on *EMM-Labs* (in the form of a geographic map¹; see **Figure 5**).

4. JRC's multilingual text mining resources

The previous section gave a rather brief overview of EMM functionality without giving technical detail. Scientific-technical details and evaluation results for all applications have been described in various publications available at <http://langtech.jrc.ec.europa.eu/>.

The four main EMM applications are freely accessible for everybody. Additionally, the JRC has made available a number of resources (via the same website) that will hopefully be useful for developers of multilingual text mining systems. The *JRC-Acquis* parallel corpus in 22 languages (Steinberger et al., 2006), comprising altogether over 1 billion words was publicly released in 2006, followed by the *DGT-Translation Memory* in 2007. A new resource that can be used both as a translation memory and as a parallel corpus for text mining use is currently under preparation. *JRC-Names*, a collection of over 400,000 entity names and their multilingual spelling variants gathered in the course of seven years of daily news analysis (see Section 3.3), has been released in

¹ <http://emm.newsbrief.eu/geo?type=event&format=html&language=all> displays continuously updated live maps.

September 2011 (Steinberger et al., 2011). *JRC-Names* also comprises software to look up these known entities in multilingual text. Finally, the *JRC Eurovoc Indexing software JEX*, which categorises text in 23 different languages according to the thousands of subject domain categories of the Eurovoc thesaurus², will also be released soon.

5. Ongoing and forthcoming work

EMM customers have been making daily use of the media monitoring software for years. While being generally satisfied with the service, they would like to have more functionality and even higher language coverage. JRC's ongoing research and development work focuses on three text mining main areas: (1) Multilingual multi-document summarisation: The purpose is to automatically summarise the thousands of news clusters generated every day; (2) Machine Translation (MT): While commercial MT software currently translates Arabic and Chinese EMM texts into English and hyperlinks to *Google Translate* are offered for all other languages, the JRC is working on developing its own MT software, based on *Moses* (Koehn et al., 2007); (3) Opinion mining / sentiment analysis: EMM users are not only interested in receiving contents, but they would also like to see opinions on certain subjects. They would like to see differences of opinions across different countries and media sources, as well as trends showing changes over time. See the JRC's Language Technology website for publications showing the current progress in these fields.

6. Acknowledgements

Developing the EMM family of applications was a major multi-annual team effort. We would like to thank our present and former colleagues in the OPTIMA group for all their hard work.

7. References

Koehn P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E. (2007): *Moses: Open Source Toolkit for Statistical Machine Translation*. Proceedings of the Annual Meeting of the Association for Computational

² See <http://eurovoc.europa.eu/>.

- Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.
- Larkey, L., Feng, F., Connell, M., Lavrenko, V. (2004): Language-specific Models in Multilingual Topic Tracking. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 402-409.
- Piskorski, J., Belyaeva, J., Atkinson, M. (2011): Exploring the usefulness of cross-lingual information fusion for refining real-time news event extraction: A preliminary study. Proceedings of the 8th International Conference 'Recent Advances in Natural Language Processing'. Hissar, Bulgaria, 14-16 September 2011.
- Pouliquen B., Steinberger, R. (2009): Automatic Construction of Multilingual Name Dictionaries. In: C. Goutte, N. Cancedda, M. Dymetman & G. Foster (eds.), Learning Machine Translation. MIT Press - Advances in Neural Information Processing Systems Series (NIPS), pp. 59-78.
- Pouliquen B., Steinberger, R., Deguernel, O. (2008): Story tracking: linking similar news over time and across languages. In Proceedings of the 2nd workshop 'Multi-source Multilingual Information Extraction and Summarization' (MMIES'2008) held at CoLing'2008. Manchester, UK, 23 August 2008.
- Pouliquen, B., Steinberger, R., Belyaeva, J. (2007): Multilingual multi-document continuously updated social networks. Proceedings of the Workshop 'Multi-source Multilingual Information Extraction and Summarization' (MMIES'2007) held at RANLP'2007, pp. 25-32. Borovets, Bulgaria, 26 September 2007.
- Steinberger R. (forthcoming): A survey of methods to ease the development of highly multilingual Text Mining applications. Language Resources and Evaluation Journal LRE.
- Steinberger R., Pouliquen, B., van der Goot, E. (2009): An Introduction to the Europe Media Monitor Family of Applications. In: F. Gey, N. Kando & J. Karlgren (eds.): Information Access in a Multilingual World - Proceedings of the SIGIR 2009 Workshop (SIGIR-CLIR'2009), pp. 1-8. Boston, USA. 23 July 2009.
- Steinberger R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D. (2006): The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006), pp. 2142-2147. Genoa, Italy, 24-26 May 2006.
- Steinberger R., Pouliquen, B., Kabadjov, M., van der Goot, E. (2011): JRC-Names: A freely available, highly multilingual named entity resource. Proceedings of the 8th International Conference 'Recent Advances in Natural Language Processing'. Hissar, Bulgaria, 14-16 September 2011.
- Tanev, H. (2007): Unsupervised Learning of Social Networks from a Multiple-Source News Corpus. Proceedings of the Workshop 'Multi-source Multilingual Information Extraction and Summarization' (MMIES'2007), held at RANLP'2007, pp. 33-40. Borovets, Bulgaria, 26 September 2007.
- Tanev, H., Zavarella, V., Linge, J., Kabadjov, M., Piskorski, J., Atkinson, M., Steinberger, R. (2009): Exploiting Machine Learning Techniques to Build an Event Extraction System for Portuguese and Spanish. *LinguaMÁTICA*, 2, pp. 55-66.

Regular Papers

Generating Inflection Variants of Multi-Word Terms for French and German

Simon Clematide, Luzia Roth

Institute of Computational Linguistics, University of Zurich

Binzmühlestr. 14, 8050 Zürich

E-mail: simon.clematide@uzh.ch, luzia.roth@access.uzh.ch

Abstract

We describe a free Web-based service for the inflection of single words and multi-word terms for French and German. Its primary purpose is to provide glossary authors (instructors or students) of an open electronic learning management system with a practical way to add inflected variants for their glossary entries. The necessary morpho-syntactic processing for analysis and generation is implemented by finite-state transducers and a unification-based grammar framework in a declarative and principled way. The techniques required for German and French terms cover two typological different types of term creation and both can be easily transferred to other languages.

Keywords: morphological generation, morphological analysis, multi-word terms, syntactic analysis, syntactic generation

1. Introduction

In the age of electronic media and rapid proliferation of technical terms and concepts, the use of glossaries and their dynamic linkage into running text seems to be important and self-evident in the area of e-learning. However, depending on the morphological properties of a language, e.g. the use of compounds or multi-word terms or the degree of surface modification that inflection imposes on words, the task of constructing inflected term variants from typically uninflected glossary entries is not a trivial task.

In this article, we describe two Web services for inflected term variant generation that illustrate the different requirements regarding morphological and syntactic processing. Whereas French shows modest inflectional variation in comparison to German, French requires more effort regarding syntactic analysis of complex nominal phrases. For German, guessing the correct inflection class of unknown compounds is more important.

A linguistically informed method for inflected term variant generation involves morphological and syntactical analysis and generation. In order to ensure this bidirectional processing, declarative linguistic frameworks such as finite-state transducers and rule-based unification grammars are beneficial. For a practical system, however, one wants to be able to analyze a wider range of expressions than what should actually be generated and presented to the user, e.g.

entries in the form of back-of-the-book indexes should be understood by the system, but these forms will not appear in running text.

Figure 1: Screenshot of the glossary author interface

The main application domain for our services is the e-Learning Management Framework OLAT¹ where we provide glossary authors with an easy but fully controllable way to add inflected variants for their glossary entries. Our free Web-based generation service² is only called once for a given term, viz. when the

¹ See <http://www.olat.org> for further information about the open source project OLAT (Online Learning and Training).

² The service is realized as a Common Gateway Interface (CGI), and it delivers a simple XML document customized for further processing in the glossary back-end of the e-learning

glossary author edits an entry. As shown in Fig. 1, the glossary author is free to select or deselect any of the generated word forms.

2. Methods and Resources

In this section, we first describe the lexical and morphological resources used for French and German. In section 2.2 we discuss the implementation of the syntactic processing module.

2.1. Lexical Resources

2.1.1. Lexical resources for French

Morphalou³, a lexicon for inflected word forms in French (95,810 lemmata, 524,725 inflected forms), was used as a lexical resource to automatically build the finite-state transducer⁴ which provides all lexical information, including word forms and morphological tags.

After the first evaluation of our development set, some modifications were made to extend the vocabulary: As derivations with neo-classical elements are quite common in terminological expressions, all adjectives were additionally combined with the prefixes of a list⁵ to create derivational forms such as *audiovisuel*, *interethnique* or *biomédical*.

Furthermore, from all lexicon entries containing a hyphen the beginning from the entry including the hyphen was extracted. This string was taken as a prefix and combined with nouns to cover cases like *demi-charge*.

2.1.2. Lexical resources for German

We use the lexicon *moliède* (Clematide, 2008), which was mainly built by us by exploiting a full form lexicon generated by Morphy (Lezius, 2000), the German lexicon of the translation system OpenLogos⁶, and the morphological resource Morphisto (Zielinski & Simon, 2008). The manually curated resource contains roughly 40,000 lemmas (nouns, adjectives, verbs), and by

applying automatic rules for derivation and conversion an additional set of 100,000 lemmas is created.

As noun compounds are the most common and productive form of terms in German, a suffix-based inflection class guesser for nouns is necessary. In an evaluation experiment with 200 randomly selected nouns from a sociology lexicon⁷, about 40% of the entries were unknown. We implemented a finite-state based ending guesser by exploiting frequency counts of lemma endings (3 up to 5 characters) from our curated lexicon. Roughly 80% of the 73 unknown singular nouns got their correct inflection class. The finite-state based ending guesser is tightly coupled with the finite-state transducer derived from our lexicon. See Clematide (2009) for technical implementation details.

2.2. Morpho-syntactic Analysis and Generation

While the generation of inflected variants for single words can be easily done with the help of finite-state techniques only, this is not the case for a proper treatment of complex multi-word terms. Therefore, we decided to use a unification-based grammar framework for syntactic processing.

The Xerox Linguistic Environment (XLE) has several benefits for our purposes:

Firstly, finite-state transducers for morphological processing integrate in a seamless and efficient way. Additionally, different tokenizer transducers can be specified for analysis and generation. This proved to be useful for the treatment of French, e.g. regarding the treatment of hyphenated compounds.

Secondly, there are predefined commands in XLE for parsing a term to its functional structure, neutralizing certain morpho-syntactic features, and generating all possible strings out of an underspecified functional structure.

Thirdly, the implementation of optimality theory in XLE allows a principled way of specifying preference heuristics, for instance for the part of speech of an ambiguous word. Additionally, using optimality marks allows to analyze more constructions than what should be generated, e.g. terms in the format of back-of-the-book indexes as *Automat*, *endlich*. With the same technique different lexical specification conventions of French

management software OLAT. See <http://kitt.cl.uzh.ch/kitt/olat>.

³ See <http://www.cnrtl.fr/lexiques/morphalou> for this resource, which is freely available for educational and academic purposes.

⁴ We use the Xerox Finite State Tools (XFST) (Beesley & Karttunen, 2003), which seamlessly integrate with the Xerox Linguistic Environment (XLE), see <http://www2.parc.com/isl/groups/nlft/xle>.

⁵ http://fr.wiktionary.org/wiki/Catégorie:Préfixes_en_français

⁶ Containing approx. 120,000 entries with inflection class categorizations of varying quality, see <http://logos-os.dfki.de>.

⁷ <http://www.socioweb.org>

	Terms	Correct Generation	Incorrect Generation	Accuracy
Development Set	400	376	24	94%
			parse failure: 19	
			wrong parse: 5	
Test Set	50	48	2	98%
			parse failure: 1	
			wrong parse: 1	

Table 1: Evaluation results for French from the development set and test set

adjectives can be handled by the XLE grammar. Lexicon entries like *grand*, *e* or *grand/e* or *grand(e)* are parsed and will result in the same output *grand*, *grande*, *grands*, *grandes*.

Lastly, dealing with unknown words is supported in XLE in a way that parts of a multi-word term that do not undergo inflection may be analyzed and regenerated verbatim. This is useful for the treatment of postnominal prepositional phrases.

The use of a full-blown grammar engineering framework for the generation of inflected term variants might be seen as too much machinery at first sight. However, the experience we gained with this approach is definitely positive. Despite the expressivity of the framework, the processing time needed for the processing of one multi-word term is about 200ms on an AMD Opteron 2200 MHz. Given the fact that our service is only called when an entry is created by a glossary author, this performance is adequate.

2.2.1. French multi-word terms

As French is more analytic than German, compounding is less prominent. The words in a multi-word term are syntactically depending on each other and require syntactic processing. The most common construction for multi-word terms is a noun combined with a preposition and a noun phrase (e.g. *droit de l'individu*). Such constructions typically correspond to German compounds. Each noun may be modified by one or more adjectives. For a correct generation of all inflected variants, the core noun and its core adjectives have to be identified, as these are the only parts to be altered for inflected variants. The core part of a French multi-word term is typically the one preceding the preposition (e.g. *droit de l'individu* → *droits de l'individu*). Due to this fact, even terms with unknown words can be handled as

long as they follow the preposition. In our XLE grammar, a default parsing strategy for unknown words occurring after a preposition is built-in and for the generation side such input is copied unchanged.

Further constructions for multi-word terms are: a noun with one or more adjectives, expressions with a hyphen (e.g. *éthylène-glycol*), noun-noun combinations (e.g. *assurance maladie*) or combinations of several nouns with *et* or *ou* (e.g. *cause et effet*). For our development set of 400 terms (see section 3.1.1 for further details), we get the following distribution: terms with prepositions (190), terms with adjectives (183), noun-noun combinations (16), terms with hyphens (9), combination of type noun *et* noun (2).

2.2.2. Preference heuristics for French

If the parsing of a one-word input term results in ambiguous structures, nouns are preferred to adjectives and verbs, as glossary entries often are nouns. For ambiguous structures of multi-word input terms the sequence noun-adjective is preferred to noun-noun, e.g. *église moderne* = *noun + adjective* instead of *noun + noun*. If a term is a combination of two nouns, only the first one is inflected, e.g. *assurance maladie* → *assurances maladie*.

In expressions with a hyphen, inflection is carried out by treating the hyphenated part of the term as normal word: Core adjectives or nouns with a hyphen are inflected, all others are not, e.g. *éthylène-glycol* → *éthylène-glycols*, or *document quasi-négociable* → *documents quasi-négociables*. In these two examples, the second part of the hyphenated expression is a core noun and has to be inflected. But there are cases where both parts of the hyphenated expression are non-core nouns. They are not inflected as in the example *égalité homme-femme* → *égalités homme-femme*. This example follows the

construction of a noun-noun multi-word term and is treated as such.

2.2.3. German multi-word terms

A detailed technical report on the XLE-based generation and analysis part for German can be found in Clematide (2009). Currently, German multi-word terms are restricted to the combination of an attributive adjective and a noun that may be given in the textual form of 'adjective noun' or as back-of-the-book index entry 'noun, adjective'. For instance, the lexicon entry *endlicher Automat* (finite state automaton) leads to the following 6 inflected forms: *endlichem Automaten*, *endlicher Automat*, *endlicher Automaten*, *endlichen Automaten*, *endliche Automat*, *endliche Automate*.

2.2.4. Related work

As far as term structures in French are concerned, Daille (2003) gives an overview that provided a base for our own analysis of multi-word terms structures. This classification was adapted and extended according to our potential glossary entries.

Jacquemin (2001) developed FASTR, a system for identifying morphological and syntactical term variants for French and English where also minor lexical modifications may take place. We did not use this system mainly for two reasons: we also had to treat German and the creation of lexical variants was of minor importance for us too.

In her contrastive study, Savary (2008) discusses different approaches of computational inflection regarding multi-word units. She emphasizes the lexical and sometimes idiosyncratic nature of multi-word expressions that may lead to problems for simple rule-based syntactic systems. However, our small-scale evaluation presented in the next section does not indicate severe problems for our approach.

3. Evaluation

In this section, we present results of our tools derived from two small-scale evaluations.

3.1.1. French

A development set with 400 and a test set with 50 glossary entries were taken randomly from *EuroVoc*⁸,

⁸ <http://eurovoc.europa.eu/drupal>

the EU's multilingual thesaurus. Table 1 shows the results for both data sets. Parsing failures were due to unknown vocabulary entries such as abbreviations (e.g. *CEC*, *P et T*) or compounds (e.g. *désoxyribonucléique*, *spéctrométrie*). Surprisingly, quite common French words like *jetable* and *environnemental* (appeared 5 times in the development set) were not covered by the lexicon. To alleviate the problem of missing vocabulary, additional open resources may be exploited⁹. Wrong parses were caused by ambiguities between nouns and adjectives.

3.1.2. German

50 German multi-word terms were selected randomly from the preferred terms in *EuroVoc*. Without the unknown word guesser, the generation of inflected variants fails for 10 terms, resulting in an accuracy of 80%. Applying the unknown word guesser for nouns allows a correct generation in 5 cases, thus giving an accuracy of 90%. 2 cases are due to unknown short nouns (the guesser requires a minimal length), 2 cases are due to unknown adjectives, and 1 case originated from an implementation error concerning adjectival nouns as *Beamter* (civil servant).

4. Conclusions

We have built a practical morphological generation service for French and German terms based on linguistically motivated processing. For multi-word terms, more constructions can be easily added through modifications of the syntactic term grammar.

In order to achieve a higher lexical coverage, other resources can be integrated. In our French system, there is already an interface that allows for simple addition of new regular nouns and adjectives. For German, additional syntactic constructions for multi-word terms will be added.

In order to resolve ambiguities on the level of parts of speech within multi-token terms, a part-of-speech tagging approach is feasible. However, for that purpose a specifically trained tagger is necessary

⁹ E.g. wiktionaries (<http://fr.wiktionary.org/wiki/Wiktionnaire>), or different lexica with inflected forms such as lefff - lexique des formes fléchies du français (<http://www.labri.fr/perso/-clement/lefff>), Dictionnaire DELA fléchi du français (<http://infolingu.univ-mlv.fr>), or Lexique3 (<http://www.-lexique.org>), a lexicon with lemmata and grammatical categories.

In a future step, we plan to extract nominal groups from a syntactically annotated corpus and use that material for the training of a part-of-speech tagger.

5. Acknowledgements

The University of Zurich supported this work by IIL grant funds. Luzia Roth implemented the French part under the supervision of Simon Clematide. The implementation of the lexicographic interface in OLAT was realized by Roman Haag under the supervision of Florian Gnägi.

6. References

- Beesley, K.R., Karttunen, L. (2003): Finite-State Morphology: Xerox Tools and Techniques. CSLI Publications.
- Clematide, S. (2008): An OLIF-based open inflectional resource and yet another morphological system for German. In A. Storrer et al. (Eds.), *Text Resources And Lexical Knowledge: selected papers from the 9th Conference on Natural Language Processing, KONVENS*, Mouton de Gruyter, pp. 183-194.
- Clematide, S. (2009): A morpho-syntactic generation service for German glossary entries. In S. Clematide, M. Klenner, and M. Volk (Eds.), *Searching Answers: Festschrift in Honour of Michael Hess on the Occasion of His 60th Birthday*, Münster, Germany: Monsenstein und Vannerdat, pp. 33-43.
- Daille, B. (2003): Conceptual Structuring Through Term Variations. In *Proceedings of the ACL 2003 workshop on multiword expressions analysis acquisition and treatment*, pp. 9-16.
- Jacquemin, C. (2001): *Spotting and Discovering Terms through Natural Language Processing*. Massachusetts Institute of Technology.
- Lezius, W. (2000): Morphy - German morphology, Part-of-Speech tagging and applications. In *Proceedings of the 9th EURALEX International Congress*, Stuttgart, pp. 619-623.
- Savary, A. (2008): Computational Inflection of Multi-Word Units. A contrastive study of lexical approaches. *Linguistic Issues in Language Technology - LiLT*, 1(2).
- Zielinski, A., Simon C. (2008): Morphisto: An Open-Source Morphological Analyzer for German. In *Proceedings of the FSMNLP 2008*, pp. 177-184.

Tackling the Variation in International Location Information Data: An Approach Using Open Semantic Databases

Janine Wolf¹, Manfred Stede², Michaela Atterer¹

¹Linguistic Search Solutions R&D GmbH, Rosenstraße 2, 10178 Berlin

²Universität Potsdam, Karl-Liebknecht-Straße 24-25, 14476 Potsdam

E-mail: janine@wolf-velten.de, stede@uni-potsdam.de, michaela.atterer@lssrd.de

Abstract

International location information ranges from mere relational descriptions of places or buildings over semi-structured address-like information up to fully structured postal address data. In order to be utilized, e.g. for associating events or people with geographical information, these location descriptions have to be decomposed and the relevant semantic information units have to be identified. However, they show a high amount of variation in order, occurrence and presentation of these semantic information units. In this work we present a new approach of using a semantic database and a rule-based algorithm to tackle the variation in such data and segment semi-structured location information strings into pre-defined elements. We show that our method is highly suitable for data cleansing and classifying address data into countries, reaching an f-score of up to 97 for the segmentation task, an f-score of 91 for the labelled segmentation task, and a success rate of 99% in the classification task.

Keywords: address parsing, OpenStreetMap, address segmentation, data cleansing

1. Introduction

Databases of international location information, as maintained by most companies, often contain incomplete address data, variation in the order of elements, mixing of international conventions for address formatting or even semi-translated address parts. Moreover, the address data can be structured insufficiently or erroneously according to the database fields which makes the data unusable for further classification, querying and data cleansing tasks.

Table 1 shows a number of possible variations of the same German address.

address string	problem description
Willy-Brandt Street 1, Berlin	partial translations
#1 Willy-Brandt Street, Berlin 1000	non-standard format
Willy-Brand-Str. 1	incorrect spelling
Willy-Brandt-Str. 1, 1000 Berlin 20	politically outdated
Willy-Brandt-Str.1, Haus 1 3.Et., Zi. 101	presence of more detailed information
In der Willy-Brandt-Str in Berlin	incomplete, e.g. extracted from free text

Table 1: Examples of variation in postal addresses based on the German address **Willy-Brandt-Str. 1, 10557 Berlin**

Apart from this kind of variation we also face variation in the description of location objects such as colloquial

variations as *Big Apple* for *New York*, historical variations (*Chemnitz/Karl-Marx-Stadt*), transcription variants (*Peking/Beijing*) or translation variants (*München/Munich*).

International addresses create further variation in address data as the typical Japanese address shown in Table 2 exemplifies.

part of description string	element type
11-1	street number (mixed information: estate and building no.)
Kamitoba-hokotate-cho	city district
Minami-ku	ward of a city (town)
Kyoto	city (here: also prefecture)
601-8501	postal code

Table 2: Address elements: Japanese postal address example **11-1 Kamitoba-hokotate-cho, Minami-ku, Kyoto 601-8501**

All these variations pose major problems for data warehousing, such as deduplication, record linkage and identity matching.

In this work we propose a method which is highly suitable for data cleansing. Tests on German, Australian and Japanese data show that it is moreover suitable for classifying address data into countries.

Our approach is rule-based and uses the open geographical database *OpenStreetMap*¹ as well as country-specific rules and patterns. It is robust and easily extensible to further languages.

2. Related Work

Most work concerned with the segmentation of location information is based on statistical techniques

(Borkar et al., 2001; Agichtein & Ganti, 2004; Christen & Belacic, 2005; Christen et. Al, 2002; Peng & McCallum, 2003; Marques & Gon Calves, 2004; Cortez & De Moura, 2010).

However, as the experiments by Borkar et al. (2001) show, these methods drastically decrease in performance once confronted with a mixture of location strings from different countries. While an experiment on uniformly formatted address data from the U.S. reaches 99.6% accuracy, performance drops to 88.9% when trained and tested on addresses from mixed countries².

There are only few published approaches on rule-based systems (Appelt et al, 1992; Riloff, 1993). Rule-based systems are generally thought to be less robust to noise, harder to adapt to other languages and thus considered suitable mainly for small domains. However a comparison of a rule based and statistical system (Borkar et al., 2001) showed that rules can compete with statistical approaches especially on inhomogeneous data. Given the fact, that huge geographical databases have become available in recent years, high-quality rule-based systems can be developed for large unrestricted domains with relatively little effort and easily be extended to more languages by adding more databases for the relevant countries.

3. Location information Segmentation

Figure 1 shows the general architecture of our system.

In a preprocessing step the location information string is **tokenised** and **normalised** according to country-specific normalisation patterns (e.g. *str* becomes *straße*). Initial **grouping** is done if applicable, i.e. if indications for grouping already exist. These steps are necessary for a

later *OpenStreetMap* query because abbreviations or partial street names cannot be found in the database.

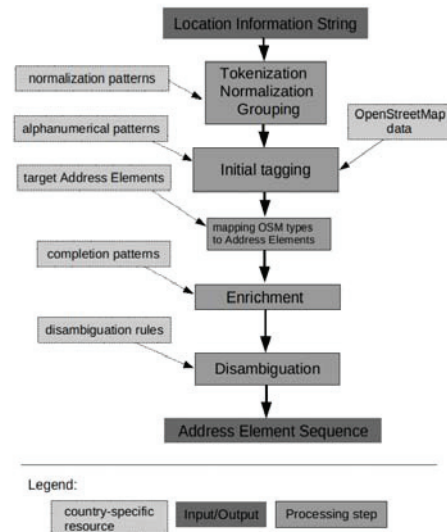


Figure 1: The system architecture

In the succeeding **tagging step** all identifiable geographical names are tagged by querying *OpenStreetMap*, allowing tagging ambiguities. Country-specific string patterns aid the tagging of elements containing numbers. One of the difficulties within this step is that address elements often consist of more than one token. The challenge lies in querying for *Oxford Street* and not erroneously tagging *Oxford* as the place name and leaving *Street* untagged. This is achieved by a longest match policy. However, all match information is preserved. Multiple queries are used to account for diacritical variation as in umlauts in German (e.g. *ü*) and parentheses as parts of official geographical names (as in *Frankfurt(Main)/Frankfurt Main*) and other non alpha-numercial marks such as hyphens.

As a result of this step, the elements are tagged with *OpenStreetMap* (OSM) internal types and not yet the address element types we are looking for. OSM types are often ambiguous. The string *Potsdam, Brandenburg* is tagged as *Potsdam (county/city/hamlet) Brandenburg (town, state)*, for instance³.

¹<http://www.openstreetmap.org>

²The accuracy measure used in this article is an overall measure of all element-wise measurements for the address elements under consideration and similar to the *labelled recall* measure used in Section 4.

³For a human reader familiar with the location it is clear, that this denotes the city of Potsdam within the state of Brandenburg, even though there is also a city called Brandenburg in the state of Brandenburg, for instance, or a hamlet called Potsdam in the state of Schleswig-Holstein.

The following step **maps OSM types to address elements**. In the *OpenStreetMap* project, every country-specific subproject, e.g. the Japanese or the German OSM project, has its own guidelines about how to tag locations according to their administrative unit status (as being a city, town or hamlet⁴). Therefore we use country specific mappings from OSM internal types to one or more of the desired target address element types we define.

The **enrichment step** provides rules for labelling address elements which have not been attributed a tag by a previous step because they were not found in the knowledge base due to spelling errors, for instance. The completion rules are of the following form:

$(type_1, type_2, \dots, type_n) \rightarrow targetAddressElement$

If for each $type_x$, $x = 1 \dots n$, for the token at index x , the respective type can be found in the list of possible types, the tokens in the sequence are grouped and labelled with the type *targetAddressElement*. Examples for language specific completion token types are found in Table 3.

A token tagged with one of these affix types indicates a (possibly still unlabelled) preceding/following location name and the token group is labelled appropriately including the marker token.

compl. type	examples	description
town_suf	ku	Suffix marking a town/ward (Japan)
station_suf	Station, Ekimae, Meieki	Word marking a train station (Japan)
village_suf	mura, son	Suffix marking a village (Japan)
city_dist_pref	Aza, Koaza	Prefix usually preceding a city district or sub-district (Japan)
street_suf	Avenue, Road	Suffix marking a street name (Australia)
state_pref	Freistaat	Prefix marking a state name (Germany)

Table 3: Completion types

Some examples of completion rules are listed in Table 4. The left hand side of the rules specifies the token type pattern, the right hand side defines the target address element. An @ means that the token at the respective

position must not have other possible types than the specified one.

completion rule	matching example
$(city_prefix, city) \rightarrow city$	Hansestadt Hamburg
$(orientation_prefix, other, street_suffix) \rightarrow street_name$	Lower Geoge Street (instead of <i>George Street</i>)
$(orientation_prefix, city)$	East Launceston
$(contains_street_suffix) \rightarrow street_name$	Ratausstraße (instead of <i>Rathausstraße</i>)
$(city, loc_suffix) \rightarrow city_district$	Berlin Mitte
$(state_prefix, state) \rightarrow state$	Freistaat Bayern
$(@city, @city) \rightarrow city$	Munich (München)
$(street_number, street_number_ext) \rightarrow street_number$	34a
$(street_number, sep_last_alphanum) \rightarrow street_number$	34 - 36

Table 4: Example completion rules

The final **disambiguation step** provides rules which decide which of the attributed types for each element is selected. In the aforementioned example, Brandenburg would thus be tagged a *state* and not a *city*.

The disambiguation rules take the form

$(leftNeighbourType, currentType, rightNeighbourType)$

where *currentType* is the target address element type of the token group under consideration. Either *rightNeighbourType* or *leftNeighbourType* may be empty (i.e. any type is allowed). If such a rule can be applied, the token group under consideration will be labelled with *currentType*.

4. Experiments

4.1. Data

We conducted our experiments using two different datasets. The first dataset was collected from the Internet, the second corpus was a company internal database. Eleven external annotators collected variations of location information data from the Internet and annotated them according to the annotation guidelines given in Wolf (2011). They collected 154 strings for German, 35 of which were used for development and the rest for testing. For Australia they collected 143 strings,

⁴A hamlet is a small town or village.

34 of which were used for development. The Japanese data were collected and annotated by the first author. 76 of the 242 data points were used for development.

The company internal database contained 57 examples for Germany, 162 examples for Australia and 56 for Japan. They were already (sometimes not correctly) attributed to 3 database fields *address*, *postal code* and *city*. To obtain a gold standard, a correct re-ordering of the elements was done manually by the first author.

4.2. Segmentation

Our first experiment consisted of correctly segmenting the internet data with our system. As a baseline we used unsophisticated systems for each language which took about 1.5 hours to program each and use patterns for postal code, a small list of endings for street names and knowledge about the typical order of address elements in the country. Our evaluation should thus reflect the superiority of a full-fledged system compared to an ad-hoc solution.

Tables 5, 6 and 7 show the evaluation results for the segmentation task for each country using f-scores based on recall and precision as computed by the PARSEVAL measures (cf. Manning & Schütze, 1996), which are suitable for evaluating systems generating bracketed structures with labels.

F-score type	baseline	system
unlabelled	87.36	96.91
labelled	70.23	91.36

Table 5: Evaluation results for German data

F-score type	baseline	system
unlabelled	68.05	95.85
labelled	64.93	86.60

Table 6: Evaluation results for Australian data

F-score type	baseline	system
unlabelled	75.45	91.80
labelled	45.47	73.50

Table 7: Evaluation results for Japanese data

The baseline systems showed above all problems with multi-token address elements (*Frankfurt (Main)*, *Bad*

Homburg) and addresses that did not conform to the standard ordering.

The full-fledged system clearly outperforms the baselines by a difference in f-score (when counting correct labels and not only correct element boundaries) of 21 points for Germany, 12 for Australia and 28 for Japan.

The contribution of the completion patterns was an increase in f-score of up to 13.03 points for the Japanese data (unlabelled) and a minimum of 0.28 for Australia (labelled).

4.3. Data cleansing

In a second experiment we tested whether the system is suitable for data cleansing. A problem already mentioned in the introduction is erroneous data structuring according to the fields of a database. By using the system for attributing address elements to the database field we could reduce the rate of elements in an incorrect database field for the company internal database by 16.77 percentage points (pp) for German, 19.31pp for Australia, and 29.84pp for Japan.

4.4. Address classification

We also conducted an experiment to find out whether the system is able to correctly guess the country of a location information string. Our testing method ignores country information (*Japan*, *Germany*, *Australia*) if present, and selects the country by computing the rate of tokens in the input which could not be classified by the system, neither by the database nor by the country specific patterns for suffixes, prefixes, special words or alphanumeric strings. As a result the system selects the country with the lowest rate of unlabelled tokens. For this experiment, we used 518 location information strings from both the Internet and the company internal data (166 for German, 271 from Australia, 81 from Japan), 99.22% of which were correctly attributed to their country.

5. Discussion and Future Work

We present a system that successfully deals with the high variability in international textual location information, by classifying the components of location strings. The implemented system is robust and easily

extensible to more countries. We tested the system with 3 countries with strongly diverging standards for the expression of location information (*Germany, Australia and Japan*). New countries can be added within a few hours, as only certain country specific files have to be edited and the corresponding *OpenStreetMap* knowledge base has to be plugged in. Most European countries are similar to Germany, and the U.S. and Canada almost identical to the Australian system, so that a large part of the world can easily be covered.

The system was shown to successfully improve the address element segmentation in a company internal database with high variation in orthography and formatting, even containing translated names.

Moreover, the system is able to almost always correctly guess the country that textual location information can be attributed to.

In future work, the system can be further improved to deal with a greater variety of typographical or transcription errors by using phonetic indexing algorithms as Soundex for English or Traphoty matching rules (Lisbach, 2010) for international languages.

6. Acknowledgements

We would like to thank all external annotators that helped gathering and annotating the test data and the LSS R&D GmbH for making a company internal address database available to us in order to test the system.

7. References

- Agichtein, E., Ganti, V. (2004): Mining Reference Tables for Automatic Text Segmentation. In KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, Seattle, WA, USA, ACM.
- Appelt, D.E., Hobbs, J.R., Bear, J., Israel, D., Tyson, M. (1992): FASTUS: A Finite-state Processor for Information Extraction from Real-world Text.
- Borkar, V.; Deshmukh, K., Sarawagi, S. (2001): Automatic segmentation of text into structured records .
- Christen, P., Belacic, D. (2005): Automated Probabilistic Address Standardisation and Verification. Australasian Data Mining Conference 2005 (AusDM05).
- Christen, P.; Churches, T., Zhu, J.X. (2002): Case-Probabilistic Name and Address Cleaning and Standardisation. The Australasian Data Mining Workshop 2002.
- Cortez, E., De Moura, E.S. (2010): ONDUX: On-Demand Unsupervised Learning for Information Extraction. In Proceedings of the 2010 international conference on Management of data (SIGMOD '10), pp. 807–818.
- Lisbach, B. (2010): Linguistisches Identity Matching. Vieweg+Teubner. ISBN 978-3-8348-9791-6. URL http://dx.doi.org/10.1007/978-3-8348-9791-6_11.
- Manning, C.D., Schütze, H. (1999): Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, Massachusetts.
- Marques, N.C., Gon Calves, S. (2004): Applying a Part-of-Speech Tagger to Postal Address Detection on the Web, 2004.
- Peng, F., McCallum, A. (2003): Accurate Information Extraction from Research Papers using Conditional Random Fields. In: Information Processing Management.
- Riloff, E. (1993): Automatically Constructing a Dictionary for Information Extraction Tasks, AAAI Press / MIT Press. pp. 811–816.
- Wolf, J. (2011): Classifying the components of textual location information. Diploma Thesis, Department für Linguistik, Universität Potsdam.

Towards Multilingual Biographical Event Extraction – Initial Thoughts on the Design of a new Annotation Scheme –

Michaela Geierhos^{*}, Jean-Leon Bouraoui[§], Patrick Watrin[§]

^{*} CIS, Ludwig-Maximilians-Universität München, Geschwister-Scholl-Platz 1, D-80539 München, Germany

[§] CENTAL, Université Catholique de Louvain, place Blaise Pascal 1, B-1348 Louvain-la-Neuve, Belgium

E-mail: micha@cis.uni-muenchen.de, mehdi.bouraoui@uclouvain.be, patrick.watrin@uclouvain.be

Abstract

Within this paper, we describe the special requirements of a semantic annotation scheme used for biographical event extraction in the framework of the European collaborative research project Biographe. This annotation scheme supports interlingual search for people due to its multilingual support covering four languages such as English, German, French and Dutch.

Keywords: biographical event extraction for interlingual people search, semantic annotation scheme

1. Introduction

In everyday life, people search is frequently used for private interests such as locating classmates and old friends, finding partners for relationships or checking someone's background.

1.1. People Search within Business Context

In a business context, *finding the right person with the appropriate skills and knowledge* is often crucial to the success of projects being undertaken (Mockus & Herbsleb, 2002). For instance, an employee may want to ascertain who worked on a particular project to find out why particular decisions were made without having to crawl through documentation (if there is any). Or, he may require a highly trained specialist to consult about a very specific problem in a particular programming language, standard, law, etc. Identifying experts may reduce costs and facilitate a better solution than could be achieved otherwise.

Possible scenarios could be the following ones:

- A personnel officer wants to find information about a person who applied for a specific position and has to collect additional career-related information about the applicant;
- A company requires a description of the state-of-the-art in some field and, therefore, wants to locate an expert this knowledge area;

- An enterprise has to set up an additional team supporting an existing group and has to find new employees with similar expertise;
- Organizers of a conference have to match submissions with reviewers;
- Job centers or even labor bureaus are interested in mapping appropriate job offers to personal data sheets.

These scenarios demonstrate that it is a real challenge within any commercial, scientific, or governmental organization to manage the expertise of employees such that experts in a particular area can be identified.

1.2. Background: The Biographe Project

A step beyond document retrieval, people search is restricted to person-related facts. The Biographe project¹ develops grammar-based analysis tools to extract person-related facts in four languages (English, German, French, and Dutch). The project received the Eurostars² label in 2009. Kick-off has been given in March 2010 and the project lasts for 24 months. The research consortium is composed of four companies and two public research departments, based in four European countries (France, Belgium, Germany, and Austria). The team creates a multipurpose people search platform able to reconstruct biographies of people. It uses all available information

¹ <http://www.biographe.org>

² <http://www.eurostars-eureka.eu>

sources such as profiles on social websites, press articles, CVs or private documents. The platform collects, extracts and structures this multilingual information in indexes and relational databases ready to be used by different task-oriented people search applications.

In this context, a semantic annotation scheme is commonly used. But conceiving such a scheme entails several technical, scientific and task-specific issues, especially when the platform is multilingual, which is still quite rare.

1.3. Multilinguality

One innovative feature of our people search platform is its multilinguality, or – to be precise – its ability to structure information, coming from the four different European languages detailed above, in a common database. By using this multilingual database, it is possible to create applications searching people through queries and documents in different languages – a feature known as interlingual search (or Cross Language Information Retrieval (CLIR)). Creating a common multilingual database allows the development of a pan-European and wholly accessible search engine offering interfaces in English and in several major European languages. Besides, this people search engine is able to handle diacritical marks such as accents (circumflex, trema, tilde, double grave accent, etc.). This apparently simple feature is very rare, due to the dominance of American search engines neglecting all accents. Accents, diacritics and not-latin symbols are very important in order to differentiate between people.

1.4. Objectives of the Paper

This extended abstract states our initial thoughts on the design of a new annotation scheme in such a specific framework. Since we cooperate with companies providing business and people search solutions, they already have established parsing technologies and our annotation scheme has to fulfill their technical requirements. Therefore, we only mention one of the main state-of-the-art schemes that is commonly used for biographical annotation tasks. We do not give a critical overview of all existing schemes because we have to develop an integrated solution and therefore try to somehow reinvent the wheel. Furthermore, we have to

discuss the particular context of multilingual annotation and finally give an example of our annotation scheme.

2. Yet another Annotation Scheme?

2.1. Linguistic Notion of Biographical Events

We define biographical events as **predicative** relations linking several arguments out of which one is an instance belonging to the argument type `<Person>`. There is no restriction on the selection of the other elements participating in a biographical relationship. However, we observed that other arguments are typically instances of the semantic classes `<Person>`, `<Location>`, `<Date>`, `<Organization>`, `<Business Sector>`, `<Subject>`, `<Profession>`, `<Award>`, etc.

- a. *John Miller* **retired** as `<Profession>senior accountant</Profession>` in `<Date>1909</Date>`.
- b. *Michael Caine* **won** the `<Award>Academy Award for Best Supporting <Profession>Actor </Profession></Award>`.
- c. *Jim Sweeney* **will also be joining** `<Organization>AmeriQuest </Organization>` as `<Profession>Vice President</Profession>`.

2.2. Events in the Information Extraction Task

One approach to defining events is used for Information Extraction (IE), being “the automatic identification of selected types of entities, relations, or events in free text” (Grishman, 2003:545). In general, information extraction tasks use surface-based patterns to identify concepts and relations between them. Patterns may be handcrafted or learned automatically, but typically include a combination of character strings, part of speech or phrasal information (Grishman, 1997). A succession of regular expressions is normally used to identify these structures; they are applied when triggered by keywords (McDonald, 1996). Most information extraction systems either use hand written extraction patterns or use a machine learning algorithm that is trained on a manually annotated corpus. Both of these approaches require massive human effort and hence prevent information extraction from becoming more widely applicable.

The problem that we are addressing is related to this traditional IE task covered by the sixth and seventh Message Understanding Conferences (MUC)³ and later

³ http://www-nlpir.nist.gov/related_projects/muc/

replaced by the Automatic Content Extraction (ACE) campaigns. According to the MUC campaigns, identifying an IE event is to extract fillers for a predefined event template. In this framework, IE events were identified by rule-based, lexicon-driven, machine learning or other systems.

2.3. The ACE Annotation Guidelines for Events

Since 1999, ACE (Automatic Content Extraction)⁴ has replaced MUC and has extended the task definition for the campaigns, including more and more scenarios. For the ACE task (Doddington et al., 2004), the participating systems are supposed to recognize several predefined semantic types of events (life, movement, transaction, business, conflict, personell, etc.) together with the constituent parts corresponding to these events (agent, object, source, target, time, location, etc.). For example, Table 1 provides an overview of the LIFE event type (with several subtypes including, BORN, DIED, etc.), together with the arguments which should be extracted for these events.

There exist approaches that identify events according to the TimeML annotation guidelines using rule-based (Sauri et al., 2005) or machine learning approaches (Bethard & Martin, 2006). The TimeML specification language was used to create the TimeBank (Pustejovsky et al., 2003) corpus.

Life event subtype	Arguments
BE-BORN	Person, Time, Place
MARRY	Person, Time, Place
DIVORCE	Person, Time, Place
INJURE	Agent, Victim, Instrument, Time, Place
DIE	Agent, Victim, Instrument, Time, Place

Table 1: An overview of ACE LIFE event subtypes

2.4. Limits of ACE annotation scheme

Since we dedicated our research to biographical events, we only address the LIFE and PERSONELL event types defined by the ACE English Annotation Guidelines for Events (Linguistic Data Consortium, 2005, p. 65 and sq.). Concerning the ACE English Annotation Guidelines for Events, the number of arguments considered as relevant

⁴ <http://projects ldc.upenn.edu/ace/>

is quite limited. For example, the BE-BORN event type disregards useful information such as the birth name, family background, or birth defects. Especially, birth names are useful to distinguish between people by identifying that, for example, *Stefani Joanne Angelina Germanotta* and *Lady Gaga* is the same person in the following context:

- d. *Lady Gaga was born as Stefani Joanne Angelina Germanotta on March 28, 1986.*

Since we need more detailed information about people, their work and occupations, we dismiss the ACE annotation standard for biographical event types. Hence we propose a more suitable one, which we present in the next sections.

3. Requirements of the Annotation Scheme

3.1. Compatible with Local Grammars

Within the Biographe Project, we focus on a linguistic description of biographical events. For example, the (born ... died ...) parentheses typically used in biographical articles help us to spot the date of birth and death in the first line of the biography. However, there are variations in expressing a lifetime period, e.g. *Jane Smith (June 1965 – September 14, 2001)*. In this case, the keywords *born* and *died* are totally missing. There are many syntactic variations in heterogeneous text expressing the same types of biographical information (e.g. birth, death) which are reduced to the basics in a structured representation. Our project partners created local grammars (Gross, 1997) using the free software tool Unitex⁵ (Paumier, 2010) in order to describe the syntactic and lexical structures of biographical information. Formally, local grammars are recursive transition networks (Woods, 1970), symbolized by graphs (cf. Figure 1).

In the framework described above, the need for named entity annotation is evident. Indeed, it relies on the accurate identification of the named entities and their corresponding relations. Consequently, it is necessary to design an annotation scheme that is capable of being integrated into the local grammar concept and can be applied to all languages provided by our system.

⁵ <http://www-igm.univ-mlv.fr/~unitex>

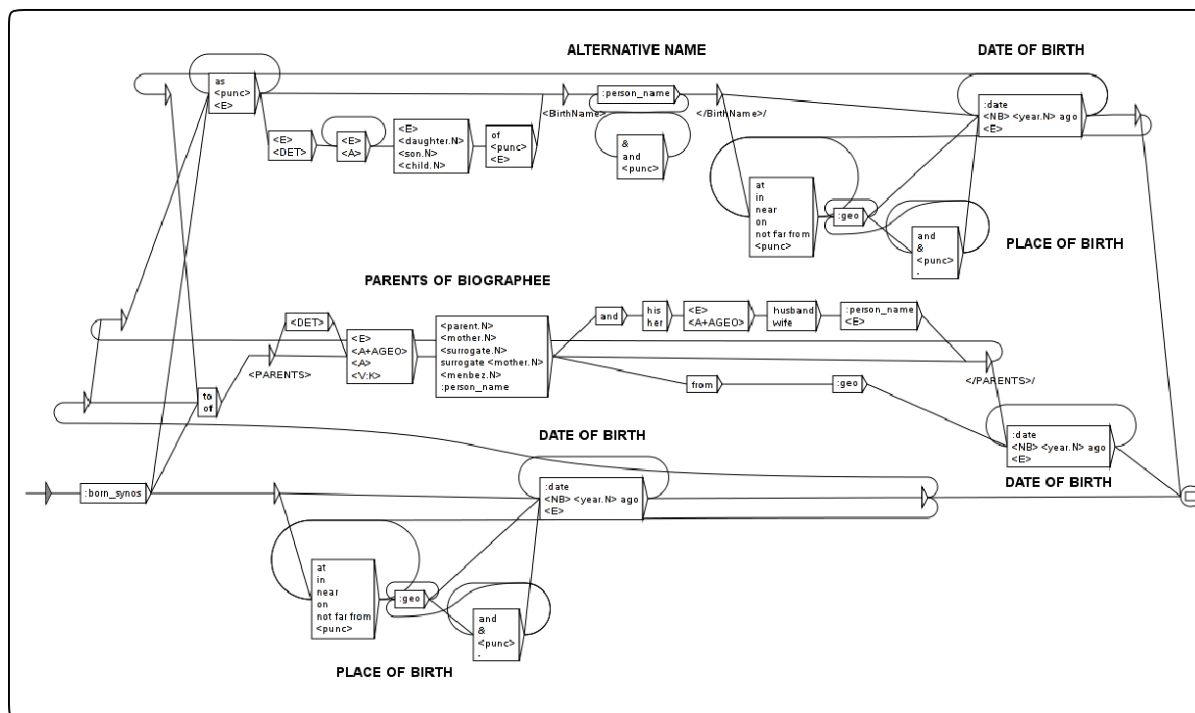


Figure 1: Local grammar for the extraction of persondata fields belonging to the event “Birth”

3.2. Definition of Annotation Units

As stated above, the scheme is used for biographical information annotation. Hence, we defined a set of named entity categories as well as the relations between them. More precisely, this scheme follows three principles:

- 1) The definition of “**entity patterns**”: they are the basis components of the annotation scheme. They benefit from the main characteristics that can be used for describing an entity; e.g. “location”, “date” ... Until now, there are 20 different entity patterns;
- 2) The next higher level is the definition of “**event patterns**”: they are composed of two or more entity patterns. In “event patterns”, entity patterns play different roles: one will always be the head of a pattern. Other optional or mandatory patterns can be attached to this head. For instance, the event pattern “awards” has as head the entity pattern “person”, which arguments are the entity patterns “domain”, “date”, and optionally another “person” if there is more than one award winner. At the same level, we also define so-called “**relation patterns**”. They build up the relationship between different entity partners in order to express the type of relation between them;

- 3) The highest level embodies the definition of “**template sets**”. They are driven from different event patterns and/or relation patterns. For example, the template set “career” comprises two event patterns: “profession” and “awards”, which themselves consist of different entity patterns, as we explained above.

3.3. Annotation Sample

Here is an instance of the use of this annotation scheme.

e. *Elio Di Rupo was born on July, 18th, 1951, at Morlanwelz, from Italian parents who arrived in Belgium in 1947.*

After applying the annotation scheme described above we get the following result (use of one event pattern and of two entity patterns):

```
<EVENT domain="biography" label="birth"><ENTITY
variable="1" anaphoric="0" category="Person">
{Elio Di Rupo,.N+comp+PERS}</ENTITY> was born
<ENTITY variable="0" anaphoric="0" category=
"date"> on {July 18th 1951,.ADV+Time+Moment}
</ENTITY> at<ENTITY variable="0" anaphoric="0"
category="Place">Morlanwelz</ENTITY></EVENT>,
from Italian {parents,.N+comp+FAMILY+IMMEDIATE}
who arrived in Belgium {in 1947,.ADV+Time+Moment
+Imprecis}.
```

3.4. Annotation Features

The sample sentence (e) annotated in Section 3.3 shows that the scheme foresees the future application of **anaphora resolution** tools. Until now, it only works with anaphoric pronouns but it is planned to extend its capabilities to more complex anaphoric terms.

There is a **{ notation}** used beside XML because the annotation scheme has to be processed by the UNITEX⁶ system (Paumier, 2010:44-46) which expects such a kind of meta-syntax in order to treat **multi-word expressions** (e.g. “July 18th 1951”) on the one hand and **assign lexico-semantic types** (e.g. ADV+Time) to text units on the other hand.

Moreover, the attribute “variable” can be assigned to the values 0 or 1 if a syntactic variability is possible for a recognized unit (e.g. “Elio de Rupo”). Since the city name given in our sample sentence (here: “Morlanwelz”) cannot change its structure, we assign 0 to the attribute “variable”. However, “Elio de Rupo” can appear another time in the text as “de Rupo” or “Elio” or “de Rupo, Elio” or “Mr de Rupo” and so on. We therefore assign 1 to the attribute “variable”.

3.5. Technical Basis

The scheme is defined in XML format. It will be applied to the text to annotate in conjunction with the use of the DBPedia ontology⁷. Remind that it is an ontology based on the extraction and the organisation of Wikipedia⁸ information. This ontology features different categories, each one corresponding to the description of a characteristic of an object or concept. These categories are linked, using the Web Ontology Language (OWL), defined by the World Wide Web Consortium (W3C)⁹.

This fine-grained ontology would be largely sufficient to cover all of the task needs. Besides, it could easily be used for producing annotations in a Resource Description Format (RDF) triple format¹⁰, also defined by the W3C. This entails that it could be easily used for conceiving and implementing a database.

Besides, such a database could be requested thanks to the

⁶ <http://www-igm.univ-mlv.fr/~unitex>

⁷ <http://dbpedia.org>

⁸ <http://www.wikipedia.org>

⁹ <http://www.w3.org/TR/owl-features>

¹⁰ <http://www.w3.org/RDF>

SPARQL language¹¹, which is a SQL like query language especially designed by the W3C to be compatible with OWL and RDF.

This solution meets all of the specifications required by the project: knowledge representation, indexation for most of the human languages (beyond English, German, French, and Dutch), updatability of the database, etc.

4. Conclusion and future works

This paper described an annotation scheme conceived and implemented in the framework of a European project. In regards to other scheme, its main advantages are the multilingual support and its generality for any named entity related task.

Our short term perspective is to evaluate its robustness, especially when automatically applied by local grammars. In future, we will adopt it to other named entity related tasks and additional natural languages.

5. Acknowledgements

This work is supported by the Eurostars Programme, a R&D initiative funded by the European Community, The Brussels Institute for Research and Innovation (INNOV^{IRIS}), and by the German Federal Ministry of Education and Research (Grant No. 01QE0902B). We express our sincere thanks to all for financing this research within the collaborative research project Biographe E!4621 (<http://www.biographe.org>).

6. References

- Bethard, S., Martin, J. (2006): Identification of event mentions and their semantic class, in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP–2006), Association for Computational Linguistics, Sydney, Australia, pp. 146-154.
- Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., Weischedel, R. (2004): The Automatic Content Extraction (ACE) Program. Tasks, Data, and Evaluation, in Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004), Canary Islands, Spain.
- Grishman, R. (1997): Information Extraction: Techniques and Challenges, in M. T. Pazzienza (ed.), Proceedings of the Information Extraction

¹¹ <http://www.w3.org/TR/rdf-sparql-query>

- International Summer School (SCIE-97), Springer-Verlag.
- Grishman, R. (2003): Information Extraction, in R. Mitkov (ed.), *The Oxford Handbook of Computational Linguistics*, Oxford University Press, pp. 545-559.
- Gross, M. (1997): The Construction of Local Grammars, in E. Roche & Y. Schabes (eds), *Finite-State Language Processing*, MIT Press, Cambridge, Massachusetts, USA: 329-354.
- Linguistic Data Consortium (2005): ACE English Annotation Guidelines for Events, Version 5.4.3 2005.07.01,
http://www ldc.upenn.edu/Projects/ACE/docs/English-Events-Guidelines_v5.4.3.pdf
- McDonald, D. (1996): Internal and External Evidence in the Identification and Semantic Categorization of Proper Names, in *Corpus Processing for Lexical Acquisition*: MIT Press, pp. 31-43.
- Mockus, A., Herbsleb, J.D. (2002): Expertise browser: a quantitative approach to identifying expertise. In *ICSE'02: Proceedings of the 24th International Conference on Software Engineering*, pp. 503–512.
- Paumier, S. (2010): Unitex User Manual 2.1,
<http://igm.univ-mlv.fr/~unitex/UnitexManual2.1.pdf>.
- Pustejovsky, J., Castaño, J, Ingria, R., Saurí, R., Gaizauskas, R., Setzer, A., Katz, G., Radev, D. (2003): TimeML: A specification language for temporal and event expressions, in *Proceedings of the International Workshop of Computational Semantics (IWCS–2003)*, Tilburg, The Netherlands.
- Saurí, R., Verhagen, M., Pustejovsky, J. (2005), Evita: A robust event recognizer for QA systems, in *Proceedings of the Joint Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP-2005)*, Vancouver, Canada, pp. 700-707.
- Woods, W. A. (1970): Transition network grammars for natural language analysis, in *Communications of the ACM*, n° 10, vol. 13, ACM, New York, NY, USA, pp. 591–606.

The *Corpus of Academic Learner English (CALE)*: A new resource for the study of lexico-grammatical variation in advanced learner varieties

Marcus Callies, Ekaterina Zaytseva

Johannes-Gutenberg-Universität Mainz, Department of English and Linguistics

Jakob-Welder-Weg 18, 55099 Mainz

E-mail: mcallies@uni-mainz.de, zaytseve@uni-mainz.de

Abstract

This paper introduces the *Corpus of Academic Learner English (CALE)*, a Language for Specific Purposes learner corpus that is currently being compiled for the quantitative and qualitative study of lexico-grammatical variation patterns in advanced learners' written academic English. CALE is designed to comprise seven academic genres produced by learners of English as a foreign language in a university setting and thus contains discipline- and genre-specific texts. The corpus will serve as an empirical basis to produce detailed case studies that examine individual (or the interplay of several) determinants of lexico-grammatical variation, e.g. semantic, structural, discourse-motivated and processing-related ones, but also those that are potentially more specific to the acquisition of L2 academic writing such as task setting, genre and writing proficiency. Another major goal is to develop a set of linguistic criteria for the assessment of advanced proficiency conceived of as "sophisticated language use in context". The research findings will be applied to teaching English for Academic Purposes by creating a web-based reference tool that will give students access to typical collocational patterns and recurring phrases used to express rhetorical functions in academic writing.

Keywords: learner English, academic writing, lexico-grammatical variation, advanced proficiency

1. Introduction

Recently, second language acquisition (SLA) research has seen an increasing interest in advanced stages of acquisition and questions of near-native competence. Corpus-based research into learner language (Learner Corpus Research, LCR) has contributed to a much clearer picture of advanced interlanguages, providing evidence that learners of various native language (L1) backgrounds have similar problems and face similar challenges on their way to near-native proficiency. Despite the growing interest in advanced proficiency, the fields of SLA and LCR are still struggling with i) a definition and clarification of the concept of "advancedness", ii) an in-depth description of ALVs, and iii) the operationalization of such a description in terms of criteria for the assessment of advancedness. In this paper, we introduce the *Corpus of Academic Learner English (CALE)*, a Language for Specific Purposes learner corpus that is currently being compiled for the quantitative and qualitative study of lexico-grammatical variation patterns in advanced learners' written academic English.

2. Corpus design and composition

Already existing learner corpora, such as the *International Corpus of Learner English* (Granger et al., 2009) include learner writing of a general argumentative, creative or literary nature, and thus not academic writing in a narrow sense. Thus, several patterns of variation that predominantly occur in academic prose (or are subject to the characteristic features of this register) are not represented at all or not frequently enough in general learner corpora. CALE is designed to comprise academic texts produced by learners of English as a foreign language (EFL) in a university setting. CALE may therefore be considered a Language for Specific Purposes learner corpus, containing discipline- and genre-specific texts (Granger & Paquot, forthcoming). Similar corpora that contain native speaker (NS) writing and may thus serve as control corpora for CALE are the *Michigan Corpus of Upper-Level Student Papers* (MICUSP, Römer & Brook O'Donnell, forthcoming) and the *British Academic Written English corpus* (BAWE, Alsop & Nesi, 2009).

CALE's seven academic text types ("genres") are written as assignments by EFL learners in university courses, see Figure 1.

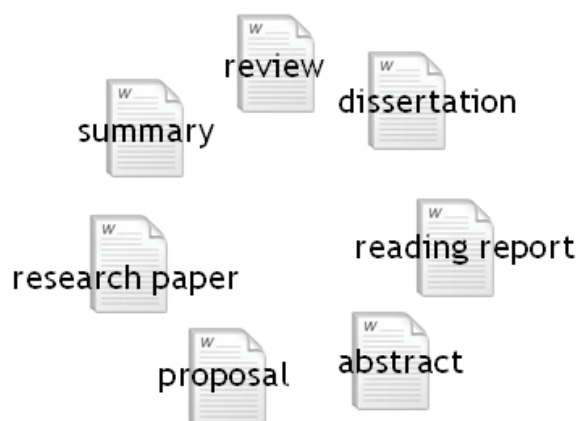


Figure 1: Academic text types in CALE

We are currently collecting texts and bio data from German, Chinese and Portuguese students, and are planning to include data from EFL learners of other L1 backgrounds to be able to draw cross-linguistic and typological comparisons as to potential L1 influence.

The text classification we have developed for CALE is comparable with the NS control corpora mentioned above, but we have created clear(er) textual profiles, adopting the situational characteristics and linguistic features identified for academic prose by Biber and Conrad (2009). A text's communicative purpose or goal serves as the main classifying principle, which helps to set apart the seven genres in terms of

- a) text's general purpose
- b) its specific purpose(s)
- c) the skills the author demonstrates, and
- d) the author's stance.

In addition, we list the major features of each text type as to

- a) structural features
- b) length, and
- c) functional features.

3. Corpus annotation

Students submit their texts in electronic form (typically in .doc, .docx or .pdf file format). Thus, some manual pre-processing of these incoming files is necessary. Extensive "non-linguistic" information (such as table of contents, list of references, tables and figures, etc.) is

deleted and substituted by placeholder tags around their headings or captions. The body of the text is then annotated for meta-textual, i.e. underlying structural features (section titles, paragraphs, quotations, examples, etc.) with the help of annotation tools. The texts are also annotated (in a file header) for metadata, i.e. learner variables such as L1, age, gender, etc. which are collected through a written questionnaire. The file header also includes metadata that pertain to each individual text such as genre, type of course and discipline the text was written in, the setting in which the text was produced etc. This information is also collected with the help of a questionnaire that accompanies each text submitted to the corpus. In the future, we also intend to implement further linguistic levels of annotation, e.g. for rhetorical function or sentence type.

4. Research program

In the following sections, we outline our research program. We adopt a variationist perspective on SLA, combining a learner corpus approach with research on interlanguage variation and near-native competence.

4.1. The study of variation in SLA research

Interlanguages (ILs) as varieties in their own right are characterized by variability even more than native languages. Research on IL-variation since the late 1970s has typically focused on beginning and intermediate learners and on variational patterns in pronunciation and morphosyntax, i.e. the (un-)successful learning of actually invariant linguistic forms and the occurrence of alternations between native and non-native equivalent forms. Such studies revealed developmental patterns, interpreted as indicators of learners' stages of acquisition, and produced evidence that IL-variation co-varies with linguistic, social/situational and psycholinguistic context, and is also subject to a variety of other factors like individual learner characteristics and biographical variables (e.g. form and length of exposure to the L2).

Since the early 2000s there has been an increasing interest in issues of sociolinguistic and sociopragmatic variation in advanced L2 learners (frequently referred to as sociolinguistic competence), e.g. learners' use of dialectal forms or pragmatic markers (mostly in L2 French, see e.g. Mougeon & Dewaele, 2004; Regan, Howard & Lemée, 2009). This has marked both a shift

from the study of beginning and intermediate to advanced learners, and a shift from the study of norm-violations to the investigation of differential knowledge as evidence of conscious awareness of (socio-)linguistic variation.

4.2. Advanced Learner Varieties (ALVs)

There is evidence that advanced learners of various language backgrounds have similar problems and face similar challenges on their way to near-native proficiency. In view of these assumed similarities, some of which will be discussed in the following, we conceive of the interlanguage of these learners as Advanced Learner Varieties (ALVs).

In a recent overview of the field, Granger (2008:269) defines advanced (written) interlanguage as "the result of a highly complex interplay of factors: developmental, teaching-induced and transfer-related, some shared by several learner populations, others more specific". According to her, typical features of ALVs are overuse of high frequency vocabulary and a limited number of prefabs, a much higher degree of personal involvement, as well as stylistic deficiencies, "often characterized by an overly spoken style or a somewhat puzzling mixture of formal and informal markers".

Moreover, advanced learners typically struggle with the acquisition of optional and/or highly L2-specific linguistic phenomena, often located at interfaces of linguistic subfields (e.g. syntax-semantics, syntax-pragmatics, see e.g. DeKeyser, 2005:7ff). As to academic writing, many of their observed difficulties are caused by a lack of understanding of the conventions of academic writing, or a lack of practice, but are not necessarily a result of interference from L1 academic conventions (McCrostie, 2008:112).

4.3. Patterns and determinants of variation in L2 academic writing

Our research program involves the study of L2 learners' acquisition of the influence of several factors on constituent order and the choice of constructional variants (e.g. genitive and dative alternation, verb-particle placement, focus constructions). One reason for this is that such variation is often located at the interfaces of linguistic subsystems, an area where advanced learners still face difficulties. Moreover, grammatical variation in L2 has not been well researched

to date and is only beginning to attract researchers' attention (Callies, 2008, 2009; Callies & Szczesniak, 2008).

There are a number of semantic, structural, discourse-motivated and processing-related determinants that influence lexico-grammatical variation whose interplay and influence on speakers' and writers' constructional choices has been widely studied in corpus-based research on L1 English. Generally speaking, in L2 English these determinants play together with several IL-specific ones such as mother tongue (L1) and proficiency level, and in (academic) writing, some further task-specific factors like imagined audience (the people to whom the text is addressed), setting, and genre add to this complex interplay of factors, see Figure 2.

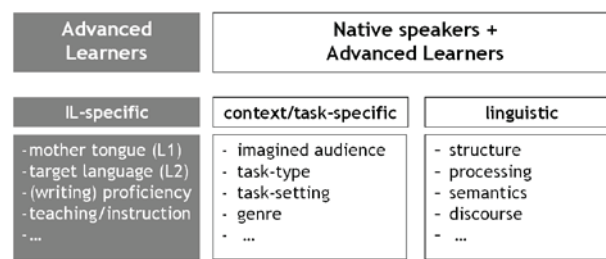


Figure 2: Determinants of variation in L1 and L2 academic writing

It is important to note at this point that differences between texts produced by L1 and L2 writers that are often attributed to the influence of the learners' L1 may in fact turn out to result from differences in task-setting (prompt, timing, access to reference works, see Ädel, 2008), and possibly task-instruction and imagined audience (see Ädel, 2006:201ff for a discussion of corpus comparability). Similarly, research findings as to learners' use of features that are more typical of speech than of academic prose have been interpreted as unawareness of register differences, but there is some evidence that the occurrence of such forms may also be caused by the influence of factors like the development of writing proficiency over time (novice writers vs. experts, see Gilquin & Paquot, 2008; Wulff & Römer, 2009), task-setting and -instruction, imagined audience and register/genre (e.g. academic vs. argumentative writing, see Zaytseva, 2011).

4.4. Case study

In this section, we provide an example of how lexico-grammatical variation plays out in L2 academic writing. In a CALE pilot study of the (non-) representation of authorship in research papers written by advanced German EFL learners, Callies (2010) examined agentivity as a determinant of lexico-grammatical variation in academic prose. He hypothesized that even advanced students were insecure about the representation of authorship due to a mixture of several reasons: conflicting advice by teachers, textbooks and style guides, the diverse conventions of different academic disciplines, students' relative unfamiliarity with academic text types and lack of linguistic resources to report events and findings without mentioning an agent. Interestingly, the study found both an overrepresentation of the first person pronouns *I* and *we*, but also an overrepresentation of the highly impersonal subject-placeholders *it* and *there* (often used in the passive voice) as default strategies to suppress the agent, see examples (1) and (2).

- (1) There are two things to be discussed in this section.
- (2) It has been shown that...

While this finding seems to be contradictory, it can be explained by a third major finding, namely the significant underrepresentation of inanimate subjects which are, according to Biber and Conrad (2009:162), preferred reporting strategies in L1 academic English, exemplified in (3) and (4).

- (3) This paper discusses...
- (4) Table 5 shows that...

Callies (2010) concluded that L2 writers have a narrower inventory of linguistic resources to report events and findings without an overt agent, and their insecurity and unfamiliarity with academic texts adds to the observed imbalanced clustering of first person pronouns, dummy-subjects and passives. The findings of this study also suggest that previous studies that frequently explain observed overrepresentations of informal, speech-like features by pointing to learners' higher degree of subjectivity and personal involvement (Granger, 2008) or unawareness of register differences (Gilquin & Paquot, 2008), may need to be supplemented by studies taking

into account a more complex interplay of factors that also includes the limited choice of alternative strategies available to L2 writers.

5. Implications for language teaching and assessment

The project we have outlined in this paper has some major implications for EFL teaching and assessment. The research findings will be used to provide recommendations for EFL teachers and learners by developing materials for teaching units in practical language courses on academic writing and English for Academic Purposes. In the long run, we plan to create a web-based reference tool that will help students look up typical collocations and recurring phrases used to express rhetorical moves/functions in academic writing (e.g. giving examples, expressing contrast, drawing conclusions etc.). This application will be geared towards students' needs and can be used as a self-study reference tool at all stages of writing an academic text. Users will be able to access information in two ways: 1) form-to-function, i.e. looking up words and phrases in an alphabetical index to see how they can express rhetorical functions, and 2) function-to-form, i.e. accessing a list of rhetorical functions to find words and phrases that are typically used to encode them.

Most importantly, the tool will present in a comparative manner structures that emerged as problematic in advanced learners' writing, for example untypical lexical co-occurrence patterns and over- or underrepresented words and phrases, side by side with those structures that typically occur in expert academic writing. This will include information on the immediate and wider context of use of single items and multi-word-units.

While the outcome is thus particularly relevant for future teachers of English, it may also be useful for students and academics in other disciplines who have to write and publish in English. Unlike in the Anglo-American education system, German secondary schools and universities do not usually provide courses in academic writing in the students' mother tongue, so that first-year students have basically no training in academic writing at all.

It has been mentioned earlier that the operationalization of a quantitatively and qualitatively well-founded description of advanced proficiency in terms of criteria

for the assessment of advancedness is still lacking. Thus, a major aim of the project is to develop a set of linguistic descriptors for the assessment of advanced proficiency. The descriptors and can-do-statements of the Common European Framework of Reference (CEFR) often appear too global and general to be of practical value for language assessment in general, and for describing advanced learners' competence as to academic writing in particular. Ortega and Byrnes (2008) discuss four ways in which advancedness has commonly been operationalised, ultimately favouring what they call "sophisticated language use in context", a construct that includes e.g. the choice among registers, repertoires and voice. This concept can serve as a basis for the development of linguistic descriptors that are characteristic of academic prose, e.g. the use of syntactic structures like inanimate subjects, phrases to express rhetorical functions (e.g. *by contrast*, *to conclude*, *in fact*), reporting verbs (*discuss*, *claim*, *suggest*, *argue*, *propose* etc.), and lexical co-occurrence patterns (e.g. *conduct*, *carry out* and *undertake* as typical verbal collocates of *experiment*, *analysis* and *research*).

6. References

- Ädel, A. (2006): Metadiscourse in L1 and L2 English. Amsterdam: Benjamins.
- Ädel, A. (2008): Involvement features in writing: do time and interaction trump register awareness? In G. Gilquin, S. Papp, & M.B. Diez-Bedmar (Eds.), *Linking up Contrastive and Learner Corpus Research*. Amsterdam: Rodopi, pp. 35-53.
- Alsop, S., Nesi, H. (2009): Issues in the development of the British Academic Written English (BAWE) corpora. *Corpora*, 4(1), pp. 71-83.
- Biber, D., S. Conrad (2009): *Register, Genre, and Style*. Cambridge: Cambridge University Press.
- Callies, M. (2008): Easy to understand but difficult to use? Raising constructions and information packaging in the advanced learner variety. In G. Gilquin, S. Papp & M.B. Diez-Bedmar (Eds.), *Linking up Contrastive and Learner Corpus Research*. Amsterdam: Rodopi, pp. 201-226.
- Callies, M. (2009): *Information Highlighting in Advanced Learner English*. Amsterdam: Benjamins.
- Callies, M. (2010): The (non-)representation of authorship in L2 academic writing. Paper presented at ICAME 31 "Corpus Linguistics and Variation in English", 26-30 May 2010, Giessen/Germany.
- Callies, M., Szczesniak, K. (2008): Argument realization, information status and syntactic weight - A learner-corpus study of the dative alternation. In P. Grommes & M. Walter (Eds.), *Fortgeschrittene Lernervarietäten. Korpuslinguistik und Zweitspracherwerbsforschung*. Tübingen: Niemeyer, pp. 165-187.
- DeKeyser, R. (2005): What makes learning second language grammar difficult? A review of issues. *Language Learning*, 55(s1), pp. 1-25.
- Gilquin, G., Paquot, M. (2008): Too chatty: Learner academic writing and register variation. *English Text Construction*, 1(1), pp. 41-61.
- Granger, S. (2008): Learner corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An international handbook*, Vol. 1. Berlin & New York: Mouton de Gruyter, pp. 259-275.
- Granger, S., Paquot, M. (forthcoming): Language for Specific Purposes learner corpora. In T.A. Upton & U. Connor (Eds.), *Language for Specific Purposes. The Encyclopedia of Applied Linguistics*. New York: Blackwell.
- Granger, S., Dagneaux, E., Meunier, F., Paquot, M. (2009): *The International Corpus of Learner English. Version 2. Handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires de Louvain.
- McCrostie, J. (2008): Writer visibility in EFL learner academic writing: A corpus-based study. *ICAME Journal*, 32, pp. 97-114.
- Mougeon, R., Dewaele, J.-M. (2004): Patterns of variation in the interlanguage of advanced second language learners. Special issue of *International Review of Applied Linguistics in Language Teaching (IRAL)*, 42(4).
- Ortega, L., Byrnes, H. (2008): The longitudinal study of advanced L2 capacities: An introduction. In L. Ortega & H. Byrnes (Eds.), *The Longitudinal Study of Advanced L2 Capacities*. New York: Routledge/Taylor & Francis, pp. 3-20.
- Regan, V., Howard, M., Lemée, I. (2009): *The Acquisition of Sociolinguistic Competence in a Study Abroad Context*. Clevedon: Multilingual Matters.
- Römer, U., Brook O'Donnell, M. (forthcoming): From student hard drive to web corpus: The design,

compilation, annotation and online distribution of MICUSP. Corpora.

Wulff, S., Römer, U. (2009): Becoming a proficient academic writer: Shifting lexical preferences in the use of the progressive. *Corpora*, 4(2), pp. 115-133.

Zaytseva, E. (2011): Register, genre, rhetorical functions: Variation in English native-speaker and learner writing. Hamburg Working Paper in Multilingualism.

From Multilingual Web-Archives to Parallel Treebanks in Five Minutes

Markus Killer, Rico Sennrich, Martin Volk

University of Zurich

Institute of Computational Linguistics, Binzmühlestrasse 14, CH-8050 Zurich, Switzerland

E-mail: markus.killer@uzh.ch, sennrich@cl.uzh.ch, volk@cl.uzh.ch

Abstract

The Tree-to-Tree (t2t) Alignment Pipe is a collection of Python scripts, generating automatically aligned parallel treebanks from multilingual web resources or existing parallel corpora. The pipe contains wrappers for a number of freely available NLP software programs. Once these third party programs have been installed and the system and corpus specific details have been updated, the pipe is designed to generate a parallel treebank with a single program call from a unix command line. We discuss alignment quality on a fully automatically processed parallel corpus.

Keywords: parallel treebank, automatic tree-to-tree alignment, TreeAligner, Text-und-Berg

1. Introduction

The process of creating parallel treebanks used to be a tedious task, involving a tremendous amount of manual annotation (see e.g. Samuelsson & Volk, 2007). Zhechev and Way (2008:1) state that "[b]ecause of this, only a few parallel treebanks exist and none are of sufficient size for productive use in any statistical MT application". Since Zhechev (2009) introduced the *Sub-Tree Aligner*, a program for the automatic generation of parallel treebanks, the feasibility of obtaining large scale annotated parallel treebanks has increased. However, the amount of preprocessing needed as well as the missing conversion of the output into a more human readable format might have kept potential users of the *Sub-Tree Aligner* at a distance. The collection of Python scripts combined in the *Tree-to-Tree Alignment Pipe (t2t-pipe)* described below takes care of all necessary pre- and postprocessing of Zhechev's *Sub-Tree Aligner*, supporting German, French and English as source and target languages. The focus of this paper is on the following two questions, both aimed at maximizing the quality of the automatic alignments:

- How big does the parallel corpus have to be in order to get satisfactory results?
- What can be said about the role of the text domain/topic of the parallel corpus?

2. Related Work

Zhechev (2009) and Koehn (2009) provide an overview of recent developments in tree-to-tree alignment, subtree alignment and the subsequent generation of parallel treebanks for use in statistical machine translation systems.

Tiedemann and Kotzé (2009) and Tiedemann (2010) propose a supervised approach to tree-to-tree alignment, requiring a small manually aligned or manually corrected treebank of at least 100 sentence pairs¹ for training purposes.

In terms of script design, the training-script for the *Moses* SMT system (Koehn, 2010b) inspired the organization of the *t2t-pipe* into several steps that can be run independently.

3. Parallel Corpora

In an ideal world, one could be inclined to take a number of parallel articles from a bilingual text collection and let the *t2t-pipe* combined with the *Sub-Tree Aligner* do the rest. Yet this is only possible if a suitable word alignment model² is available, as we will show in section 5.

¹ See <http://stp.lingfil.uu.se/~joerg/Lingua/index.html> (accessed: 21/08/11)

² All word alignment models used in this paper can be downloaded from: <http://t2t-pipe.svn.sourceforge.net/> (accessed: 21/08/11)

With the aim of collecting information on the role of corpus size and text domain/topic in creating an automatically aligned parallel treebank, the following corpora were used:

3.1. Corpus for Tree-to-Tree Alignment

A subcorpus of the Text+Berg corpus (Volk et al., 2010) consisting of four parallel articles from the Swiss Alpine Club Yearbook 1977 served as test corpus (see [TUB-4-ART] in table 1). Details on the corpus with regard to the extraction of parallel articles and sentence pairs are described in Sennrich and Volk (2010). For the purpose of this paper it is sufficient to note that the vast majority of texts can be attributed to the journalistic textual domains article/report/review with a strong topical focus on activities performed by members of the Swiss Alpine Club (climbing, hiking, trekking) and the alpine environment in general. As the corpus has been digitised from printed books it contains OCR errors.

Corpus	Lang.	Tokens	Sentence Pairs
[TUB-4-ART]	DE	21,689	1,171
	FR	25,388	(GIZA++: 1,023)
[TUB]	DE	1,617,301	92,518
	FR	1,921,583	(GIZA++: 80,698)
[EPARL]	DE	35,371,164	1,562,563
	FR	42,427,755	(GIZA++: 1,190,609)

Table 1: Parallel Corpora

[TUB-4-ART] Text+Berg Corpus 4 Articles SAC YB 1977

[TUB] Text+Berg Corpus SAC Yearbooks 1957-1982

[EPARL] Europarl Corpus 1996-2009

3.2. Corpora for Word Alignment

Additionally, we used the complete Text+Berg corpus [TUB], the Europarl corpus (Koehn, 2010a) [EPARL] and combinations of these two corpora to compute different word alignment models (see table 1 for basic corpus information). Word alignment is automatically computed through GIZA++ (Och & Ney, 2003), which implements the IBM word alignment models. For performance reasons, we set the maximum sentence length to 40 tokens³. Therefore, we used only 83% of

³ See <http://www.statmt.org/wmt11/baseline.html> (accessed: 21/08/11)

the of the [TUB] corpus and 76% of the [EPARL] corpus to estimate word alignment probabilities (see table 1 for absolute values in brackets).

We used [EPARL] to test the impact of corpus size on the results. Moreover, texts from the [EPARL] corpus belong to a completely different textual domain (parliament proceedings) and cover a wide range of political, economic and cultural topics (see Koehn, 2009:53), making it possible to use the data to figure out the role of text domain/topic in the alignment process.

4. The *t2t-pipe*

Taking an existing parallel corpus⁴ as input, the *t2t-pipe* runs through seven steps to generate automatic alignments for individual words and syntactic constituents in each parallel sentence pair. The configuration file is deliberately designed in a way that a number of different third party programs can be chosen for most of the steps, enabling easy switching between different configurations. In the brief outline of the following steps, the configuration that worked best is indicated (please refer to the *t2t-pipe* README file⁵ for details on all 12 programs used):

4.1. Steps 1-5 – Preprocessing

- 1) Extraction of Parallel Articles
- 2) Tokenization
(*Python NLTK Punkt-Tokenizer*)
Rudimentary OCR cleaning/
Fixing of word division errors
- 3) Sentence Alignment
(*Hunalign* with *dict.cc* dictionary)
- 4) Statistical Phrase Structure Parsing
(*Stanford Parser* for German,
Berkeley Parser for French)
- 5) Word Alignment
(*GIZA++* through *Moses* training script,
enhanced with *dict.cc* dictionary,
see section 4.2 for an example),
data not lower-cased

⁴ If no parallel corpus is available, the pipe includes scripts for the on-the-fly construction of a parallel corpus from the web archives of the bilingual Swiss Alpine Club magazine (German-French).

⁵ Available from: <http://t2t-pipe.svn.sourceforge.net/> (accessed: 21/08/11)

4.2. Step 6 - Tree-to-Tree Alignment

This is the most important step in a complete run of the *t2t-pipe*, as the automatic alignments are generated by Zechev's *Sub-Tree Aligner*. The process can best be described by looking at a parallel sentence pair, taken from [TUB-4-ART]:

- 1) German sentence: *Man versuche einmal einen solchen Mann abzubremesen.*
- 2) French sentence: *Essayez donc de freiner un tel homme.*⁶

- Input:

- a. Bracketed parse trees of source and target language (output of the two parsers combined into one file):

```
(ROOT (NUR (S (PIS Man) (VVFIN versuche) (ADV
einmal) (VP (NP (ART einen) (PIDAT solchen) (NN
Mann)) (VVIZU abzubremesen))) ($ . !))) \n
(ROOT (SENT (VN (V Essayez)) (ADV donc) (VPinf (P
de) (VN (V freiner)) (NP (D un) (A tel) (N
homme))) (. !)))\n\n
```

- b. Two lexical translation files generated by the *Moses* training script and *GIZA++*, enhanced using a *dict.cc* dictionary:

lex.e2f (French – German – Probability)

```
Homme Mann 1.0000000
homme Mann 1.0000000
mari Mann 1.0000000
ralentir abzubremesen 0.0666667
freiner abzubremesen 0.0666667
```

lex.f2e (German – French – Probability)

```
abzubremesen ralentir 0.0053476
abzubremesen freiner 0.0035842
Mann Homme 1.0000000
Mann homme 1.0000000
Mann mari 1.0000000
```

- Output:

Indexed bracketed parse trees of source and target language with alignment indices on a separate line (see Figure 1 for graphical alignments). In our example sentence, the *Sub-Tree Aligner* produced one wrong alignment, linking the German personal pronoun *man* to the French finite verb *essayez* (*emphasised below*):

```
(ROOT::NUR-2 (S-3 (PIS-4 Man) (VVFIN-5 versuche)
(ADV-6 einmal) (VP-7 (NP-8 (ART-9 einen) (PIDAT-10
solchen) (NN-11 Mann)) (VVIZU-12 abzubremesen))) ($.-
13 !)) \n
(ROOT::SENT-2 (VN::V-4 Essayez) (ADV-5 donc)
(VPinf-6 (P-7 de) (VN::V-9 freiner) (NP-10 (D-11
un) (A-12 tel) (N-13 homme))) (. -14 !)) \n
2 2 4 4 6 5 7 6 8 10 9 11 10 12 11 13 12 9 13 14
```

4.3. Step 7 - Conversion to TigerXML/TMX

We converted the output of Zechev's *Sub-Tree Aligner* into two language specific *TigerXML* files and an additional *XML* file containing information on node alignments. These files can be easily imported into the graphical interface of the *Stockholm TreeAligner* (Lundborg et al., 2007). Figure 1 shows the previously introduced sentence pair – including the automatically computed links – in the treebank browser perspective of the *Stockholm TreeAligner*.

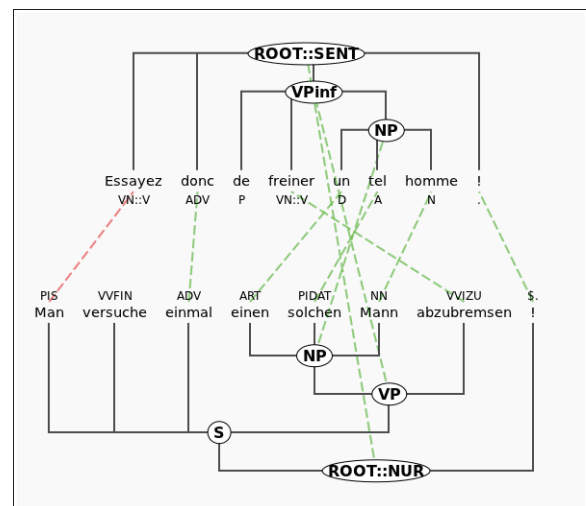


Figure 1: Automatically aligned sentence pair in *Stockholm TreeAligner*

The second supported output format is *TMX*, a format for current translation memory systems (tested with OmegaT⁷).

5. Treebank Alignment Quality

We ran six experiments (summarized in table 2) on the test corpus [TUB-4-ART] (see table 1). In each experiment, the corpus used to compute the lexical translation probabilities with *GIZA++* either differed

⁶ Sentences 1) and 2) translate roughly as: [(*Why don't*) you try to slow down a man like that (a heavy man)!]

⁷ Available from: <http://www.omegat.org> (accessed: 21/08/11)

Corpus	1 [TUB-4-ART]	2 [TUB-4-ART]	3 [EPARL]	4 [TUB]	5 [TUB-EPARL]	6 [TUB-EPARL]
Corpus Size <i>GIZA++</i>	1,023 SP	1,023 SP	1,190,609 SP	80,698 SP	258,971 SP	1,271,307 SP
In-domain (%)	100.0%	100.0%	0.0%	100.0%	31.0%	6.0%
<i>Dict.cc</i> SA/WA	NO	YES	YES	YES	YES	YES
Precision WA	57.8%	61.1%	51.3%	65.9%	69.1%	69.2%
Precision PhA	58.3%	65.4%	51.8%	81.7%	79.5%	80.4%
Precision allA	57.9%	62.1%	51.4%	69.2%	71.3%	71.7%
Correct links per SP	8.66	9.63	9.02	12.48	13.64	13.98

Table 2: Alignment precision and average number of correct links in treebank of [TUB-4-ART] corpus (1,171 sentence pairs) with respect to size, enhancement through additional lexical resources and textual domain of the corpus used to compute the lexical translation probabilities.

Precision = Correct Alignments / Suggested Alignments, SP: Sentence Pair(s) SA: Sentence Alignment, WA: Word Alignment, PhA: Phrase Alignment, allA: Word & Phrase Alignments, In-domain: domain correspondence of treebank and WA corpus

with respect to corpus size and textual domain or enhancement by external lexical resources (*dict.cc* dictionary).

We manually checked an average of 545 alignments (77% word alignments 23% phrase alignments) in 32 randomly selected sentence pairs⁸ for each of the six resulting treebanks, using the *Stockholm TreeAligner*. Our information on changes in recall is based on the absolute number of correct links in the manually checked sentence pairs (average no. of correct links = average no. of all links⁹ x precision¹⁰).

5.1. Corpus Size

Looking at the configuration outlined in section 4, three of the seven steps in the *t2t-pipe* directly depend on the corpus size (Tokenization (Dehyphenation), Sentence Alignment and Word Alignment). The analysis of the alignment quality in the resulting parallel treebank shows that roughly 1000 sentence pairs are not enough to get satisfactory results with an overall precision of 57.9% (see table 2, experiment 1). Initial tests have shown that Zhechev's *Sub-Tree Aligner* is highly

dependent on the quality of the word alignments supplied. Even though the algorithm does not directly replicate the *GIZA++* alignments:

[M]y system uses a probabilistic bilingual dictionary derived from the *GIZA++* word alignments, thus being able to side-step errors present in the original word-alignment data and to find new possible alignments that *GIZA++* had skipped for the particular sentence pair.

(Zhechev, 2009:73)

We employed two measures to increase the precision of the alignments:

- 1) We enhanced the lexical translation probabilities computed by *GIZA++* by extracting all 1-to-1 word translations from the freely available *dict.cc* dictionary (DE-FR), leading to a substantial increase in precision (+ 4.2%) and in recall (+ 0.97 correct links per sentence pair).
- 2) Step-by-step, we increased the corpus size, making use of all available resources. In experiment 3 it becomes clear that a huge increase of corpus size alone is no guarantee for better alignment results: When we use the 1,190,609 sentence pair [EPARL] corpus on its own, the recall drops by 0.61 correct

⁸ This number proved to be sufficient to include at least 100 Phrase Alignments in the sample. The identity of the treebank was masked for the manual evaluation.

⁹ computed by *Sub-Tree Aligner* for the whole treebank

¹⁰ computed from manually checked sentence pairs

links per sentence pair and the precision by 10.7% compared to experiment 2. However, increasing the size of the [TUB] corpus from 1,023 to 80,698 sentence pairs as a basis for the word alignment model leads to the biggest leap in the experiment sequence in both precision (+ 7.1%) and recall (+2.85 correct links per sentence pair) compared to experiment 2.

5.2. Domain/Topic Specific Content

The data collected in table 2 suggests that when using the unsupervised approach proposed by Zhechev (2009) the domain of the corpus used to compute the lexical translation probabilities seems to be of great importance. In experiment 3, we observe the poorest precision of all experiments with the second biggest corpus [EPARL]. Apart from a few common lexical items (e.g. *mountain, valley, river, ...*) there is hardly any overlap in terms of textual domain/topic (see section 3) and the [TUB- 4-ART] corpus itself was not used to compute lexical probabilities in experiment 3 (hence the 0% correspondence between the two corpora).

Comparing these results to the supervised approach by Tiedemann and Kotz  (2009), there seems to be an important difference, as they observe "only a slight drop in performance when training on a different textual domain" (204). The main reason for this might be that in the supervised approach the program trains phrase alignments from manually aligned training data (relatively domain/topic independent), whereas in the unsupervised approach the parallel corpus is used to compute lexical translation probabilities (heavily dependent on domain/topic).

5.3. The Right Balance of Corpus Size and Domain/Topic Specific Content

Bearing this difference of the two approaches in mind, it is not surprising that balancing (in terms of textual domain/topic - experiment 5) or expanding (maximising corpus size - experiment 6) the word alignment model affects the results in a different way:

When using a better model for estimating lexical probabilities (more data: Europarl+SMULTRON) the performance

improves only slightly to about 58.64% [F-Score compared to 57.57%]

(Tiedemann & Kotz , 2009:204)

In the unsupervised approach (used in the *t2t-pipe*) however, the use of a better word alignment model [TUB-EPARL] increases the recall by another 1.16 and 1.50 correct links per sentence pair, respectively (experiments 5/6), compared to the largest corpus with a 100% domain correspondence (experiment 4). For phrase alignments, we achieved a precision of roughly 80% from a corpus size of approx. 80,000 sentence pairs of the same domain (experiments 4-6). The maximum precision of word alignments in this set-up (data not being lower-cased) seems to be around 70% from a corpus size of about 250,000 sentence pairs, while the recall can still be slightly increased by supplying more and more data to estimate lexical probabilities. As long as there is a solid basis of several 10,000 sentence pairs belonging to the same textual domain as the parallel corpus to be aligned, expanding the corpus used to compute lexical probabilities with material of another textual domain does not seem to harm the results but can still help to increase overall precision and recall by a small margin.

6. Conclusion and Outlook

We designed the *t2t-pipe* considering the following areas of application:

- 1) Assisting human annotators of a parallel treebank by supplying good alignment suggestions: The results discussed in section 5 have shown that this can be achieved by employing a large enough parallel corpus of approx. 250,000 sentence pairs with data of the same textual domain. If the corpus is not big enough, the results can be improved by adding language material of a completely different textual domain. We achieved an overall precision of 71.7% (approx. 80% for phrase alignments). Using a corpus of 500-1,000 sentence pairs (a common size for human annotated parallel treebanks) or a word alignment model trained solely on a different textual domain does not lead to reasonable automatic alignments. However, if there already is a suitable word alignment model for a specific text

domain/topic, the generation of a brand new treebank is just five minutes away.

- 2) Visualisation/manual evaluation of the results of different components of a tree-based SMT system (e.g. Parsing, Word/Phrase Alignment): The data collected and analysed in section 5 is one possible application of the *t2t-pipe* in this category.
- 3) As a by-product, the *t2t-pipe* produces phrase alignments for translation memory systems: With a corpus of approx. 80,000 sentence pairs, the precision of the alignments is around 80%. These alignments can be manually checked and a new *TMX* file can be easily generated from the corrected alignment data.

In future versions of the program, the two approaches presented by Zhechev (2009) and Tiedemann and Kotz e (2009) could be combined. We see additional potential for improvement in using lower-cased data and a corpus free of OCR errors for word and subtree alignment.

7. References

- Koehn, P. (2009): Statistical Machine Translation. Cambridge: Cambridge University Press.
- Koehn, P. (2010a): European Parliament Proceedings Parallel Corpus 1996-2009. Release v5. TXT-Format. Description in: Europarl: A Parallel Corpus for Statistical Machine Translation, Philipp Koehn, MT Summit 2005. URL: <http://www.statmt.org/europarl>.
- Koehn, P. (2010b): MOSES. Statistical Machine Translation System. User Manual and Code Guide, November. URL: <http://www.statmt.org/moses/manual/manual.pdf>.
- Lundborg J., Marek T., Mettler M., Volk, M. (2007): Using the Stockholm TreeAligner. In Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories (TLT'06). Bergen, Norway: Northern European Association for Language Technology, pp. 73–78.
- Och, F. J., Ney, H. (2003): A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics 29, pp. 19–51.
- Samuelsson, Y., Volk, M. (2007): Alignment Tools for Parallel Treebanks. In Proceedings of the GLDV Fr uhjahrstagung, T ubingen, Germany.
- Sennrich R., Volk, M. (2010): MT-based Sentence Alignment for OCR-generated Parallel Texts. In Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010).
- Tiedemann J., Kotz e, G. (2009): Building a Large Machine-Aligned Parallel Treebank. In Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT'08). Milano, Italy: EDUCatt: pp. 197–208.
- Tiedemann J. (2010): Lingua-Align: An Experimental Toolbox for Automatic Tree-to-Tree Alignment. In Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'2010), Valetta, Malta.
- Volk, M., Bubenhofer, N., Althaus A., Bangerter, M., Marek T., Ruef, B. (2010): Text+Berg-Korpus (Pre-Release 118+ Digitale Edition Die Alpen 1957-1982). XML-Format, May. Digitale Edition des Jahrbuch des SAC 1864-1923 und Die Alpen 1925-1995. URL: <http://www.textberg.ch>.
- Zhechev V., Way, A. (2008): Automatic Generation of Parallel Treebanks. In Proceedings of the 22nd International Conference on Computational Linguistics. Manchester, UK: pp. 1105–1112.
- Zhechev, V. (2009): Automatic Generation of Parallel Treebanks. An Efficient Unsupervised System. Dissertation, School of Computing, Dublin City University.

Querying multilevel annotation and alignment for detecting grammatical valence divergencies

Oliver Čulo

FTSK, Universität Mainz

An der Hochschule 2, 76726 Germersheim

E-mail: culo@uni-mainz.de

Abstract

The valence concept has been used in machine translation as well as didactics on order to build up valence dictionaries for the respective uses. Most valence dictionaries have been built up manually, but given the growing number of parallel resources, it would be desirable to automatically exploit them as basis for building up bilingual valence dictionaries. The present contribution reports on a pilot study on a German-English parallel corpus. In this study, patterns of verb plus grammatical functions were extracted from parallel sentences. The paper reports on some of the basic findings of this extraction, regarding divergencies both in valence patterns as well as syntactic realisations of the predicate, i.e. the verb. These findings set the agenda for further research, which should focus on how to detect semantic shifts of valence carriers in translation and how this affects valence.

Keywords: valence, valence extraction, parallel corpora, translation

1. Introduction

The concept of valence (Tesnière, 1959) has been endorsed in multilingual research domains in various ways. Various machine translation systems use some notion of valence in the core of their analysis and transfer structures (see relevant descriptions e.g. for EUROTRA (Steiner, Schmidt & Zelinsky-Wibbelt, 1988), METAL (Gebrowsers, 1988), Verbmobil (Emele et al., 2000) or TectoMT (Žabokrtský, Ptáček & Pajas 2008)). For didactic purposes, various bilingual valence dictionaries have been compiled (D. Rall, Rall, & Zorrilla, 1980; Engel & Savin, 1983; Bianco, 1996; Simon-Vandenbergen, Taeldeman & Willems 1996).

Most of the valence resources mentioned are based on manually compiled valence dictionaries. Nowadays, as ever more and larger parallel corpus resources are available, it is desirable to exploit these in order to gain more data for bilingual valence dictionary creation. There have been various attempts at extracting bilingual valence dictionaries from parallel corpora. In some cases, the extraction process is tackled from a high-level semantic level, as in the case of bilingual frame semantic dictionaries (Boas, 2002; 2005). Other

approaches choose a syntactic annotation, as in the case of the Prague Czech-English Dependency Treebank (Čmejrek et al., 2004). In both cases, the semantic or „deep“ dependency (or *tectogrammatical*, see (Sgall, Hajičová & Panevová, 1986)) annotation abstracts away from syntactic variation, making the extraction task somewhat less complex. In the course of the FUSE-project (Cyrus, 2006), predicate-argument annotation and alignment between German and English sentences serves as basis for the study of both syntactic and semantic valence divergencies. Padó (2007) investigates the (frame) semantic dimension of valence divergencies. In the former case, the annotation is very specifically tailored to the project itself, making the methods harder to reproduce when applied to other corpora. In the latter study, the level of investigation again abstracts away from syntactic variation.

The study presented here focusses on grammatical differences in valence pattern between German and English. Both for the detection and description of differences, top-level grammatical function like subject, direct object etc. are used. This follows the tradition of using grammatical functions rather than syntactic

categories as e.g. in the previously listed bilingual valence dictionaries. Grammatical functions abstract away from syntactic variation but as compared to e.g. the tectogrammatical approach of (Čmejrek et al., 2004), no deep annotation is needed in order to retrieve grammatical functions of a sentence.

The corpus used in the study is annotated and aligned on multiple linguistic levels, but not with a specific focus on valence. Also, the method of querying multiple annotation and alignment levels at once is outlined. On top of that, valence divergencies are discussed with respect to factors like contrastive differences, register or translation properties and strategies.

2. Study setup

2.1. The corpus

The corpus used in the study was built to investigate contrastive commonalities and differences between English and German as well as peculiarities in translations. It consists of English originals (EO), their German translations (GTrans) as well as German originals (GO) and their English translations (ETrans). Both translation directions are represented in eight registers with at least 10 texts totalling 31,250 words per register. In the present paper, examples are taken from the registers SHARE (corporate communications), SPEECH (political speeches) and FICTION (fictional texts). Altogether, the corpus comprises one million words. Additionally, register-neutral reference corpora are included for German and English including 2,000 word samples from 17 registers.

All texts are annotated with part-of-speech information using the TnT tagger (Brants, 2000), morphology using MPRO (Maas, Rösener & Theofilidis, 2009), and grammatical functions and chunk categories, manually annotated with MMAX2 (Müller & Strube, 2006).

Furthermore, all texts are aligned on word level using GIZA++ (Och & Ney, 2003), on chunk level indirectly by mapping the grammatical functions onto each other, on clause level manually again using MMAX2, and on sentence level using the WinAlign component of the

Trados Translator's Workbench (Heyn, 1996) with additional manual correction.

2.2. A format independent API for multilevel queries

The API designed for the corpus is made up of three parts. On top, there is the interface, containing control methods with basic read/write and iteration calls for the corpus. Under the hood, a package called CoReTool is used to represent linguistic structures in stratified layers, and the parallel structures (e.g. aligned words, sentences, etc.) as sets of pairs. The intermediate level handles the XML-based data format of the corpus. Queries are mainly written using the format-independent CoReTool data structures and are thus re-usable for other corpora as well. The layers dealing with corpus management and format handling can, in theory, be exchanged depending on the corpus used. This stratificational approach is a major difference between this corpus API and other APIs, where programming data structures and underlying data format are more closely linked.

Fundamental within CoReTool is the notion of TEXT. A CORPUS is made up of an ordered collection of TEXTS, which again is made up of an ordered collection of SENTENCES, which again is made up of an ordered collection of TOKENS. This structure is so to speak the backbone of CoReTool and the minimum of data that we expect in a corpus. In addition, a CORPUS can be divided into REGISTERS which also relate to collections of TEXTS (from the CORPUS). Likewise, a SENTENCE can contain CLAUSES or CHUNKS which relate to the TOKENS of the SENTENCE. For each of these sub-units of a text (including TOKENS), it is possible to have aligned counterparts. Every single alignment is represented as a pair; so if unit U is aligned with U' and U'' , there will be two pairs $\langle U, U' \rangle$ and $\langle U, U'' \rangle$.

The CoReTool Java package uses simple data structures like ordered lists to organize the linguistic content it represents. In addition, a couple of basic methods for calculating statistics – e.g. numbers on chunk categories or grammatical functions – are included. The package so far lacks a proper backend-enabled design, so that IO methods could be easily plugged in on demand. Also,

```

for every wordPair in wordPairs

  slWord := getS1Word(wordPair)
  tlWord := getTlWord(wordPair)
  slChunk := getChunkForWord(slWord)
  tlChunk := getChunkForWord(tlWord)

  if (not mappable(getGramFunc(slChunk), getGramFunc(tlChunk))
  then markCrossingLine(slWord, tlWord, slChunk, tlChunk)
  end if

end for

```

Figure 1: Pseudo-Code of the query for crossing lines between grammatical functions and words

the linguistic representation of CoRETool is currently restricted to syntactic structures. However, the need to extend the package with further functionalities, e.g. in to be able to operate with semantic annotation as well, may or will hopefully soon be rendered unnecessary by latest developments of query tools like e.g. ANNIS2¹.

2.3. Querying for *empty links* and *crossing lines*

Two concepts are used to detect instances of valence divergencies. These concepts are based on well-known concepts from translation studies. Elements which have no alignment exhibit an *empty link*. Such 0:1-equivalents have been described e.g. by Koller (2001). Elements which are aligned, but which are embedded in higher units that are not aligned, result in *crossing lines*. This would e.g. be the case for two aligned words which are embedded in different grammatical functions. Crossing lines relate to the concept of shifts (in the given example a shift in grammatical function) as described e.g. by Catford (1965).

The corpus is queried for empty links and crossing lines using the CoRETool package. Empty links can be detected by simply querying one alignment level. For crossing lines, querying combinations of both annotation and alignment levels is necessary. A query for a shift in function requires (1) going through pairs of aligned words, (2) for each pair: getting the chunks the aligned words are embedded in, and (3) checking the mapping of these chunks, i.e. check whether the grammatical

functions they've been assigned are compatible (cf. figure 1). As in this study setup the same set of grammatical functions was used for German and English, mapping was straightforward.

3. Divergencies in valence patterns for grammatical functions

The ideal situation for valence extraction from parallel corpora would be that of sentence pairs with equivalent verbs at their core and perfectly matching syntactic patterns. Minor shifts, e.g. in the type of grammatical functions governed by the verb, can easily be accounted for. However, besides differences in realisation of arguments, there may also be differences in the realisation of the predicate. Such a typical shift is the *head switch*, in examples like *Ich schwimme gern* – *I like swimming*, where the German adverb *gern* ‘willingly, with pleasure’ becomes the full verb *like* in English. As we will see, there may be other factors for different kinds of shifts in the verb. We will be looking at more semantically/pragmatically triggered shifts, for a more syntactic investigation especially of shifts in the realisation of the predicate, e.g. support verb constructions versus full verbs, see (Čulo, 2010).

Probably the simplest case for a valence divergency on the level of grammatical functions is that of differences in the kinds of grammatical function as which an argument is realised. Compare, for instance, the sentence pair in figure 2, with the English original on top and the German translation at the bottom, and let us focus on the phrase “*Most admired Company in*

¹<http://www.sfb632.uni-potsdam.de/d1/annis/>

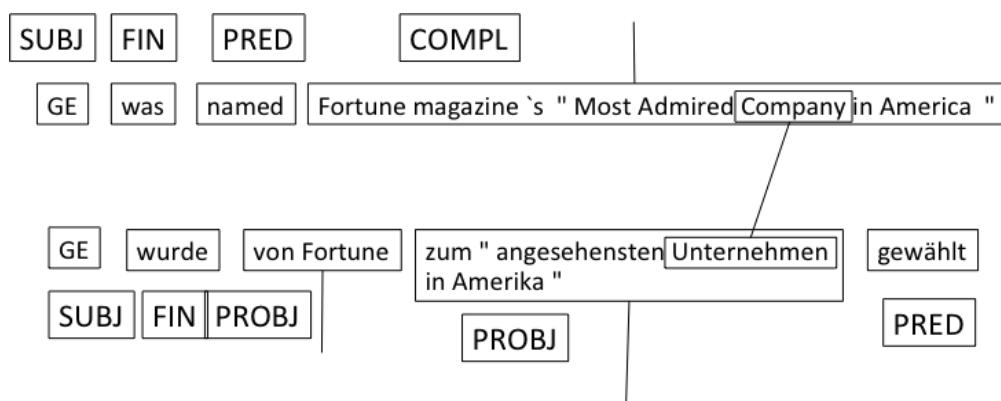


Figure 2: A crossing line for the words Company and Unternehmen and the grammatical functions COMPL and PROBJ

America". This phrase is embedded in a predicative complement (tag: COMPL) in English, as is governed by verbs like *name*, *appoint*, *elect* etc. The COMPL function has no equivalent in German, resulting in an empty link (indicated by the vertical lines with only linked to only one box). In order to understand, though, what is happening in that case, one has to evaluate the links from within the phrase: the word *Company*, for instance, is aligned with the equivalent word *Unternehmen* which is, however, embedded in a prepositional object (PROBJ) in German. The cause for this shift lies in a contrastive difference in the valence patterns of a whole class of verbs (namely the APPOINT class, following Levin (1993)). But, as there currently is no semantic annotation present in the corpus, there is no automatic way of linking the verb sense to this particular

shift. We will come back to this point when discussing the last example.

A similar shift from COMPL to a different function is shown in figure 3. Here, however, the shift is not triggered by the fact that two equivalent verbs have different valence patterns, but by a change of the main verb which does not match known concepts like head switches.

	<i>be</i> → <i>sein</i>	<i>be</i> → <i>sein</i>
E2G_SHARE	37 % (126)	63 % (215)
E2G_FICTION	45 % (138)	54 % (168)
E2G_SPEECH	60 % (224)	40 % (147)

Table 1: Proportions of *be* translated as either *sein* or with a different verb than *sein*

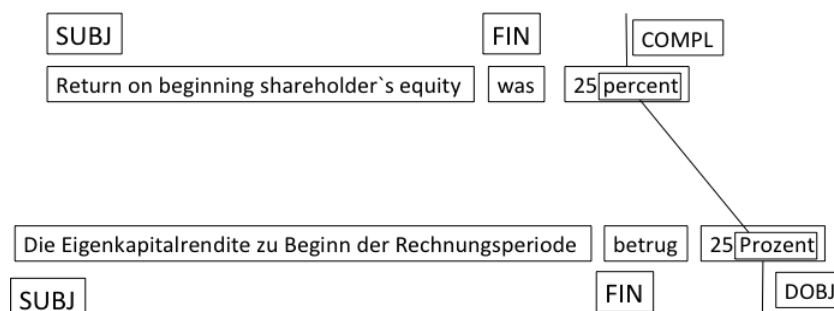


Figure 3: From English copular verb to German full verb

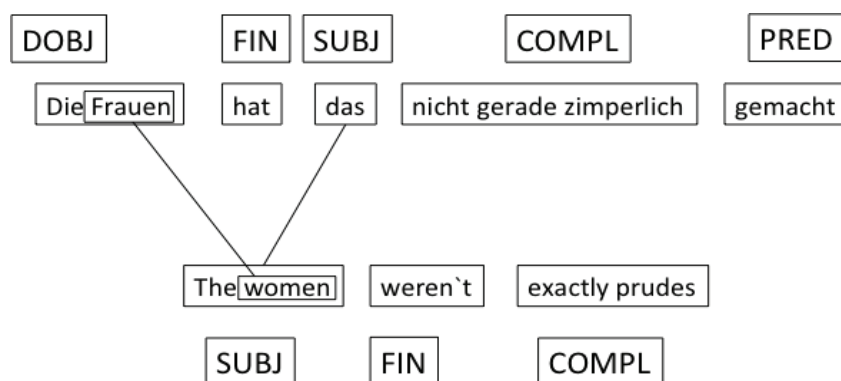


Figure 4: Multiple shifts as a result of translation strategies

The English copular verb *be* is translated with the transitive verb *betragen* in German. This particular kind of verb shift can be observed very often in the register SHARE, as shown in table 1. The reason for this lies in differences in style between English and German SHARE texts: English uses a more colloquial style where German puts rather formulaic expressions, using more full verbs than copular verbs.

Many of the shifts found in translations can be attributed to translation strategies as described e.g. by (Vinay & Darbelnet, 1958) for French and English. An example of a modulation can be seen in figure 4. Here, what can be described by looking at the surface realisation, is that the word order from the German original has been kept in the English translation, probably to preserve the stress which is put on the phrase *Die Frauen* 'the women'. But, while in German the first constituent is a direct object, this order of grammatical functions cannot be easily reproduced in English. A possible solution, as presented in the given example, is to shift the direct object to another function, here: the subject. In the given example, the verb is shifted, too, from transitive *gemacht* 'made' to the copular *weren't*. One could hypothesise that this happens in order to adapt to the different configuration of functions and their semantic content. However, in order to really explain the more complex cases of multiple shifts in one sentence, further data / annotations may be needed.

If, for instance, we add frame semantic annotation, we may be able to describe the shift of the verb with relation to shifts in semantic content. In the example in figure 4, one could annotate the first sentence with the *Cause_change* frame (with *das* as *Cause* and *Die Frauen* as *Entity*), the second one with the *state_of_entity* frame. The English sentence could thus be interpreted as a translation of only a partial component of the sense of the original sentence: the English translation focusses on the outcome of the *Cause_change* process in the German original, giving more stress to the *Entity* (*the women*) in the *State_of_entity* by placing it to the sentence initial position. How to deal with such shifts – whether to include them in an extraction process or not – remains a matter of discussion. Data from process-based translation experiments may prove helpful for shedding light on the reasons for such a “partial” translation.

4. Conclusion and outlook

As has been shown, empty links and crossing lines have proven to be reliable indicators for detecting and in some cases a basis for describing differences in grammatical valence patterns. Furthermore, it has been shown that annotation and alignment on multiple levels can be used for studying valence divergencies and possibly for extracting bilingual valence dictionaries, without resorting to an annotation scheme specialised on these purposes only.

Future work shall concentrate on a broader categorisation of valence divergencies with respect to more factors than those listed in this paper. In order to be able to link verb senses and certain types of shifts, the next step is to add (frame) semantic annotation to the corpus. Also, the purely product based data presented here could be complemented by process-based studies in the future, which should yield a more sound explanation of shifts as depicted in figure 4.

5. References

- Bianco, M. T. (1996): Valenzlexikon deutsch-italienisch. Deutsch im Kontrast 17. Heidelberg: Julius Groos.
- Boas, H. C. (2002): Bilingual FrameNet dictionaries for machine translation. In Proceedings of the third international conference on language resources and evaluation, 4:1364-1371. Las Palmas, Spanien.
- (2005): Semantic frames as interlingual representations for multilingual lexical databases. International Journal of Lexicography 4, no. 18: 445-478.
- Catford, J. C. (1965): A linguistic theory of translation. an essay in applied linguistics. Oxford: Oxford University Press.
- Čmejrek, M., Cuřín, J., Havelka, J., Hajič, J., Kubon. V. (2004): Prague Czech-English dependency treebank: syntactically annotated resources for machine translation. In Proceedings of LREC 2004, 5:1597-1600. Lisbon, Portugal.
- Čulo, O. (2010): Valency, translation and the syntactic realisation of the predicate. In D. Vitaš and C. Krstev, Proceedings of the 29th International Conference on Lexis and Grammar (LGC), 73-82. Belgrade, Serbia.
- Cyrus, L. (2006): Building a resource for studying translation shifts. In Proceedings of LREC 2006.
- Emele, M. C., Dorna, M., Lüdeling, A., Zinsmeister, H., Rohrer, C. (2000): Semantic-based transfer. In W. Wahlster (ed.), *Verbmobil*, 359-376. Artificial intelligence. Berlin ; Heidelberg [u.a.]: Springer.
- Engel, U., Savin, E. (1983): Valenzlexikon deutsch-rumänisch. Deutsch im Kontrast 3. Heidelberg: Julius Groos.
- Gebrowsers, R. (1988): Valency and MT: recent developments in the METAL system. In Proceedings of the second conference on applied natural language processing, 168-175.
- Koller, W. (2001): Einführung in die Übersetzungswissenschaft. Narr Studienbücher. Tübingen: Gunter Narr.
- Levin, B. (1993): English verb classes and alternations. The University Chicago Press.
- Padó, S. (2007): Translational equivalence and cross-lingual parallelism: the case of framenet frames. In Proceedings of the nodalida workshop on building frame semantics resources for scandinavian and baltic languages. Tartu, Estonia.
- Rall, D., Rall, M., Zorrilla, O. (1980): Diccionario de valencias verbales: aleman-español. Tübingen: Gunter Narr.
- Sgall, P., Hajičová, E., Panevová, J. (1986): The meaning of the sentence in its semantic and pragmatic aspects. Springer Netherland.
- Simon-Vandenberg, A.-M., Taldeman, J., Willems, D. (eds) (1996): Aspects of contrastive verb valency. *Studia Germanica Gandensia* 40.
- Steiner, E., Schmidt, P., Zelinsky-Wibbelt. C. (1988): From syntax to semantics: insights from machine translation. London: Francis Pinter.
- Tesnière, L. (1959): *Éléments de syntaxe structurale*. Paris: Klincksieck.
- Vinay, J.-P., Darbelnet, J. (1958): *Stylistique comparée du français et de l'anglais. Méthode de translation*. Paris: Didier.
- Žabokrtský, Z., Ptáček, J., Pajas. P. (2008). TectoMT: highly modular MT system with tectogrammatcs used as transfer layer. In Proceedings of WMT 2008.

SPIGA - A Multilingual News Aggregator

Leonhard Hennig[†], Danuta Ploch[†], Daniel Prawdzik[§], Benjamin Armbruster[§], Christoph Büscher[§], Ernesto William De Luca[†], Holger Düwiger[§], Sahin Albayrak[†]

[†]DAI-Labor, TU Berlin

Berlin, Germany

E-mail: {leonhard.hennig,danuta.ploch,ernesto.deluca,sahin.albayrak}@dai-labor.de,

{daniel.prawdzik,benjamin.armbruster,christoph.buescher,holger.duwiger}@neofonie.de

[§]Neofonie GmbH

Berlin, Germany

Abstract

News aggregation web sites collect and group news articles from a multitude of sources in order to help users navigate and consume large amounts of news material. In this context, Topic Detection and Tracking (TDT) methods address the challenges of identifying new events in streams of news articles, and of threading together related articles. We propose a novel model for a multilingual news aggregator that groups together news articles in different languages, and thus allows users to get an overview of important events and their reception in different countries. Our model combines a vector space model representation of documents based on a multilingual lexicon of Wikipedia-derived concepts with named entity disambiguation and multilingual clustering methods for TDT. We describe an implementation of our approach on a large-scale, real-life data stream of English and German newswire sources, and present an evaluation of the Named Entity Disambiguation module, which achieves state-of-the-art performance on a German and an English evaluation dataset.

Keywords: topic detection and tracking, named entity disambiguation, multilingual clustering, news personalization

1. Introduction

News aggregation web sites such as Google News¹ and Yahoo! News² collect and group news articles from a multitude of sources in order to help users navigate and consume large amounts of news material. Such systems allow users to stay informed on current events, and to follow a news story as it evolves over time. In this context, an event is defined as something that happens at a specific time and place (Fiscus & Doddington, 2002), e.g. “the earthquake that struck Japan on March 11th, 2011”.

Topic Detection and Tracking (TDT) methods address two main challenges of such systems: The detection of new events (topics) and the tracking of articles related to a known topic in newswire streams (Allan, 2002). Addressing these tasks typically requires a comparison of text models. In topic tracking, the comparison is between a document and a topic, which is often represented as a centroid vector of the topic’s documents. Topic detection compares a document to all known topics, to decide if the

document is about a novel topic. Text models are often based on the Vector Space Model, or are represented as language models (Larkey, 2004).

Going one step further, multilingual news aggregation enables users to get an overview of the press coverage of an event in different countries and languages, and has been a part of TDT evaluations since 1999 (Wayne, 2000). For multilingual TDT, topic and document comparisons require the use of multilingual text models, or alternatively the translation of documents (Larkey, 2004). Previous research has typically used machine translation to convert stories to a base language (Wayne, 2000). Machine-translated documents, however, are of lower quality than human-translated documents, and full-fledged machine translation of complete documents is costly in terms of required models and linguistic tools (Larkey, 2004). Moreover, real-life TDT systems have to filter large amounts of new documents as they arrive over time, and thus require the use of efficient, scalable approaches.

As news stories typically revolve around people, places, and other named entities, Shah et al. (2006) show that using concepts, such as named entities and topical

¹<http://news.google.com>

²<http://news.yahoo.com>

keywords, rather than all words for vector representations can lead to a higher TDT performance. While there are many ways to extract concepts from documents, Wikipedia has gained much interest recently as a lexical resource (Mihalcea, 2007), as it covers concepts from a wide range of domains and is freely available in many languages. Furthermore, Wikipedia’s inter-language links can be used to translate multilingual concepts. However, previous research in multilingual TDT has not attempted to utilize Wikipedia as a resource for concept extraction and translation.

Representing documents as concept vectors raises the additional challenge of dealing with natural language ambiguities, such as ambiguous name mentions and the use of synonyms (Cucerzan, 2007). For example, the name mention ‘Jordan’ may refer to several different persons, a river, and a country. As these phenomena lower the quality of vector representations, it is necessary to resolve ambiguous name mentions against their correct real-world referent. This task is known as Named Entity Disambiguation (NED) (Bunescu & Pasca, 2006). State-of-the-art approaches to NED employ supervised machine learning algorithms to combine features based on document context knowledge with entity information stored in an encyclopedic knowledge base (KB) (Bunescu & Pasca, 2006; Zhang et al., 2010). Common features include popularity (Dredze et al., 2010), similarity metrics exploring Wikipedia’s concept relations (Han & Zhao, 2009), and string similarity. In current research, NED has mainly been considered as an isolated task (Ji & Grishman, 2011), and has not yet been applied in the context of TDT.

The contributions of this paper are twofold: We propose a novel model for a multilingual news aggregator that combines Wikipedia-based concept extraction, named entity disambiguation, and multilingual TDT (Section 2). Our model is based on a representation of documents and topics as vectors of concepts. This choice of representation, combined with concept translation, enables the application of a wide range of well-known TDT algorithms regardless of the language of the input documents, and leads to efficient and scalable implementations. We also describe an implementation of our model on a large-scale, multilingual news stream. Furthermore, we extend our NED algorithm previously proposed in (Ploch, 2010) to a German KB, and present

an evaluation of the Named Entity Disambiguation module on a newly-created German dataset (Section 3).

2. Multilingual News Aggregation Model

Our approach to multilingual TDT is schematically outlined in Figure 1. For each news article, we successively perform language-dependent concept extraction (Section 2.1), NED (Section 2.2) and multilingual TDT (Section 2.3). In addition, we outline an algorithm for news personalization in Section 2.4. Finally, we give details of the implementation of our model in Section 2.5, and describe a user interface for the presentation of news stories in Section 2.6.

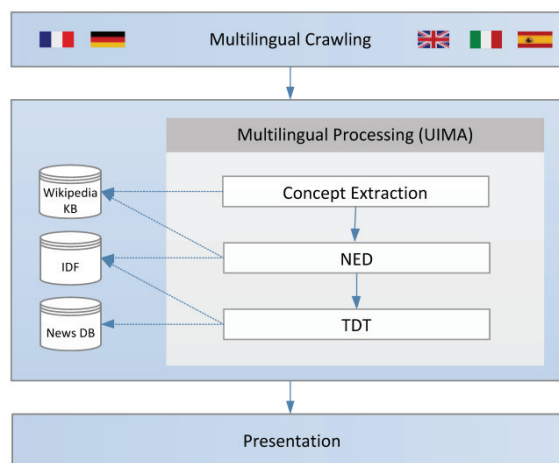


Figure 1: Multilingual News Aggregation Model

2.1. Concept extraction

We create a lexicon of terms, phrases and named entities by collecting titles, internal anchor texts, and redirects, from Wikipedia articles. The use of Wikipedia as the basis of our lexicon allows us to construct concept vectors for news articles in different languages, and facilitates the creation of new lexicons. We utilize the inter-language tables of Wikipedia to create a mapping between concepts in different languages. In the final lexicon, each concept is represented by an image, which is used to uniquely identify the concept, and a list of linguistic variants (inflected forms, synonyms and abbreviations). For example, the concept ‘Jordan (Country)’ may be referred to by ‘Jordan’, ‘Urdun’, or ‘Hashemite Kingdom of Jordan’.

After concept extraction, each news article is represented as a weighted bag-of-concepts. All other words contained

in the document are discarded. We weight concepts using a variant of the traditional tf.idf-weighting scheme (Allan, 2005). The document frequency is calculated over a sliding time window in order to better reflect the changing significance of terms in a dynamic collection of news articles:

$$w(c_i, d_j) = \frac{n(c_i, d_j)}{n(c_i, d_j) + 0.5 + 1.5 \times |d_j| / |\bar{d}|} \times \frac{\log((|D| + 0.5) / n_D(c_i))}{\log(1 + |D|)},$$

where $w(c_i, d_j)$ is the weight of concept i in document j , D is the collection of documents, $n(c_i, d_j)$ is the frequency of concept i in document j and $n_D(c_i)$ is the number of documents containing c_i .

2.2. Multilingual Named Entity Disambiguation

The concept vector of a document may initially encompass ambiguous concepts, and in particular ambiguous name mentions. If a document contains e.g. the name mention ‘Michael Jordan’ the real-world referent might be the famous basketball player, but also the researcher in machine learning known under this name. The same document may also refer to ‘Air Jordan’, which is a synonymous name for the basketball player. In both cases the challenge is to figure out the correct meaning of the name mention for clearly constructing the concept vector of the document.

Our approach to NED is based on our earlier work described in (Ploch, 2010), which we extend here to a German KB. We disambiguate name mentions found in a text by utilizing an encyclopedic reference knowledge base (KB) to link a name mention to at most one entry in the KB (Bunescu & Pasca, 2006). Furthermore, we also determine if a name mention refers to an entity not covered by the KB, which is known as Out-of-KB detection (Dredze et al., 2010). This may occur for less popular but still newsworthy entities with no corresponding KB entry. Especially challenging is the disambiguation of common names, like for instance ‘Paul Smith’, of unknown entities sharing their name with a popular namesake.

Our approach to NED is based on the observation that entities in texts co-occur with other entities. We therefore utilize the entities surrounding an ambiguous name for their resolution. On the basis of Wikipedia’s internal link

graph we create a reference KB containing for each entity its known surface forms (i.e. name variants) and its links to other entities and concepts (Wikipedia articles).

Given a name mention identified in a document, the candidate selection component retrieves a set of candidate entities from the KB, using a fuzzy, weighted search on index fields storing article titles, redirect titles, and name variants. We cast NED as a supervised classification task and train two Support Vector Machine (SVM) classifiers (Vapnik, 1995). The first classifier ranks the candidate KB entities for a given surface form. Subsequently, the second classifier determines whether the surface form refers to an Out-of-KB entity. Besides calculating well-known NED features like the bag-of-words similarity, the popularity of an entity given a specific surface form and the string similarity (baseline feature set), we implement features that exploit Wikipedia’s link graph. To this end, we represent the document context of an ambiguous entity and each candidate as a vector of links that are associated with the candidate entities in our KB, and compute several similarity features using the resulting bag-of-links vectors. The full approach is described in more detail in (Ploch, 2010).

2.3. Multilingual Topic Detection and Tracking

Given the disambiguated concept vector representation of a document, we employ a hierarchical agglomerative clustering approach for TDT. The centroid vector of a topic is created by averaging the concept weights of the documents assigned to that topic. The clustering algorithm then compares a new document to the centroid vectors of existing topics using a combination of the two vectors’ cosine similarity and a time-dependent penalty. The time factor is included to prefer assigning new documents to more recent events, and to limit the infinite growth of old events (Nallapati et al., 2004). If a document’s similarity to all clusters is lower than a predefined threshold, we assume that this document deals with a new event, and starts a new cluster.

In order to cluster documents from different languages, we utilize the inter-language mappings and translate the concept vectors to a single language. Thus, the document concept vectors as well as the cluster centroid vectors share a common space of concepts, to which we can apply our clustering approach.

2.4. News Personalization

The Personal News Agent (PNA) enables the user to personalize the news stream to match her information need. We define a user profile as a weighted vector u consisting of components u^+ and u^- , which represent the concepts that a user is interested respectively not interested in. We include u^- to allow for a more fine-grained control of news selection. Similar to the centroid vectors of document clusters, this approach enables a language-independent representation of a user's information needs.

The process of identifying relevant news articles is performed analogously to the TDT algorithm described in the previous section. The relevance of a new document with respect to the user profile is calculated as the cosine similarity of the document's concept vector and u . Documents with a similarity higher than a predefined threshold are assumed to match a user's information need, and presented to the user.

2.5. System Implementation

Our implementation of the approach described in the previous sections consists of three main components, and is shown in Figure 1. We used a crawler that collects news articles and associated metadata from approximately 1400 German and English newswire sources. The news articles are processed in a pipeline based on the Apache UIMA framework³. Events and the news articles associated with them are presented to the user via a web interface. The system is geared towards large-scale processing of newswire streams in near real-time. It processes approximately 70.000 news articles per day, and manages up to 200.000 event clusters over a time span of four weeks.

The current system processes English and German news, using a lexicon of 1.5 and 1.1 million concepts respectively, and is planned to include French, Italian and Spanish news sources. The usable intersection between the German and English lexicons amounts to 700K concepts. Concepts are identified in text with a longest-matching substring strategy (Gusfield, 1999). The concept weighting uses a time span of 4 weeks to determine document frequency.

Our implementation of the NED module utilizes

classifier models trained on the TAC-KBP 2009 dataset and a German dataset (see Section 3), both of which are based on newswire documents.

The TDT component's parameters, such as cluster similarity thresholds and time penalty values, are currently tuned manually based on an analysis of the clusters produced by the algorithm. We utilize the concept set of the German Wikipedia as the basis for translating the concept vectors of English news articles. In addition, concept types are weighted differently, as for example places and person names are more helpful than general topics to detect events in news streams.

For the news personalization component, the creation of a user profile is based on the selection of news articles by the user according to her interests. Concept vectors are extracted from user-selected articles as described in Section 2.1. The concept vectors are then merged and weighted to create a centroid vector u , with concepts having a negative weight representing the component u^- . The news personalization module uses a slightly different weighting scheme than the TDT component, assigning a higher weight to general topics (e.g. elections, tax cuts) than to named entities.

2.6. User Interface

We present events and news articles to users via a web interface. The interface includes a start page giving an overview of the most important events in several news categories, as well as pages for each category. Given the large amount of news stories published every day, our system implements several methods to rank event clusters for presentation to the user. These include measures based on cluster novelty, size, and hotness. The hotness measure is calculated as a weighted combination of a cluster's total growth since its creation time, and its recent growth in a sliding time window. For our system, we determined the weights experimentally over a range of settings. This approach ensures that breaking news are presented first both on the start page and on category pages. In addition, we implement a filtering strategy for news articles to provide users with an in-depth, diversity-oriented overview of each event, instead of merely listing an event's news articles in order of their age. Figure 2 shows the overview page of an example event, displaying the event's lead article as well as two earlier news articles in German and English.

³ Apache UIMA– Unstructured Information Management Architecture (<http://uima.apache.org/>)

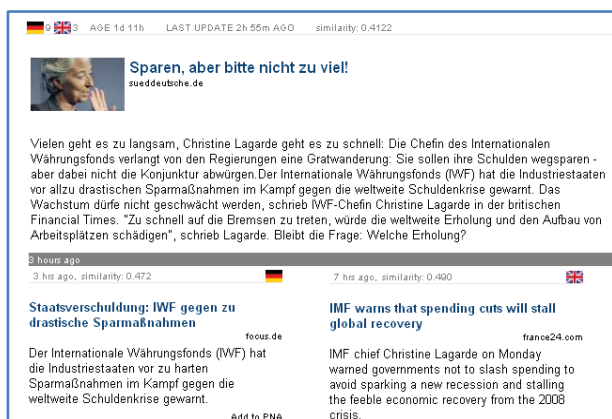


Figure 2: A sample multilingual news cluster

3. Evaluation of NED

We evaluate the quality of our NED approach on two datasets to examine how its performance compares to other state-of-the-art systems, and which accuracy it achieves for different languages.

The first dataset is the TAC-KBP 2009 dataset for English (Simpson et al., 2009). It consists of 3,904 queries (name mention-document pairs) with 57% queries for Out-of-KB entities. The KB queries are divided into 69% queries for organizations and 15% queries for persons and geopolitical entities each. In addition to the English NED dataset we created a German dataset with 2,359 queries. This dataset consists of 30% Out-of-KB queries and 70% KB queries, where 46% of the queries relate to organizations, 27% to persons and 24% to geopolitical entities. 3% are of an unknown type ‘UKN’.

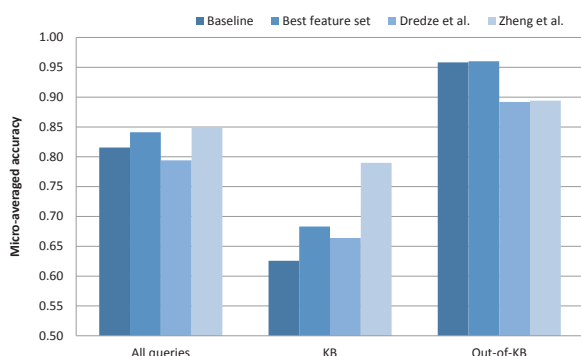


Figure 3: Micro-averaged accuracy of different approaches to English NED for the TAC-KBP 2009 dataset on all, KB and Out-of-KB queries.

For both datasets, we perform 10-fold cross-validation by training the SVM classifiers on 90% of the queries and

testing on the remaining 10%. Results reported in this paper are then averaged across the test folds. We utilize the official TAC-KBP 2009 evaluation measure of micro-averaged accuracy, which is computed as the fraction of correctly answered queries.

Figure 3 and Figure 4 show the micro-averaged accuracies for all, KB and Out-Of-KB queries. As shown in Figure 3 for the English dataset, our best feature set improves the accuracy of the baseline model by 2.7%, and achieves a micro-averaged accuracy of 0.84. Regarding other systems tested on the same dataset (Dredze et al., 2010; Zheng et al., 2010), our results compare favorably. In particular, the detection of Out-of-KB entities outperforms that of other systems. The experiments confirm our assumption that co-occurring entities and their relations are suitable for NED. Similar results are obtained for the German dataset, as shown in Figure 4. The overall accuracy of 0.77 on this dataset is slightly lower than for the TAC 2009 dataset. Again, the accuracy for Out-of-KB queries is higher than the disambiguation accuracy for KB queries, but compared to TAC 2009 the results are more balanced.

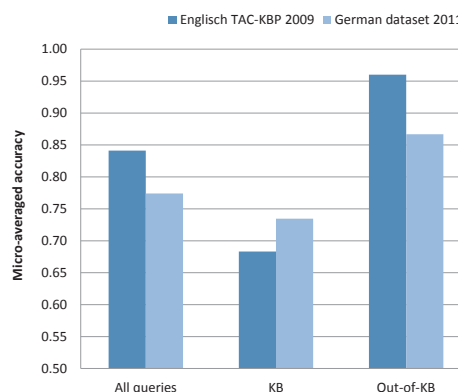


Figure 4: Comparison of micro-averaged NED accuracy on the English TAC-KBP 2009 and the German dataset.

4. Conclusions

We described a model for a multilingual news aggregator which combines Wikipedia-based concept extraction, named entity disambiguation and multilingual TDT to detect and track events in multilingual news streams. Our approach exploits Wikipedia as a large-scale, multilingual knowledge source both for representing documents as concept vectors and for resolving ambiguous named entities. We also described a

fully-operational implementation of our approach on a real-life, large scale multilingual news stream. Finally, we presented an evaluation of the Named Entity Disambiguation module on a German and an English dataset. Our approach achieves state-of-the-art results on the TAC-KBP 2009 dataset, and shows similar performance on a German dataset.

In future work, we plan to evaluate the Topic Detection and Tracking component using the TDT 3 dataset (Wayne, 2000), in order to verify the validity of our overall approach. We also plan to evaluate the effect of NED on the performance of the TDT algorithm.

Furthermore, we intend to include more languages to provide a pan-European overview of news events. This will raise additional challenges related to the mapping of concepts in different languages, the disambiguation of named entities, and the clustering strategies applicable to the resulting vector representation, since many Wikipedia versions are often significantly smaller than the English one. For example, we plan to extend our link-based NED approach by exploiting cross-lingual information.

5. Acknowledgments

The authors wish to express their thanks to the Neofonie GmbH team who strongly contributed to this work. The project SPIGA is funded by the Federal Ministry of Economics and Technology (BMWi).

6. References

- Allan, J. (2002): Introduction to topic detection and tracking. In: *Topic detection and tracking*, pp. 1–16. Kluwer Academic Publishers.
- Allan, J., Harding, S., Fisher, D., Bolivar, A., Guzman-Lara, S., Amstutz, P. (2005): Taking topic detection from evaluation to practice. In: *Proc. of HICSS '05*.
- Bunescu, R., Pasca, M. (2006): Using encyclopedic knowledge for named entity disambiguation. In: *Proc. of EACL-06*, pp. 9–16.
- Cucerzan, S. (2007): Large-Scale named entity disambiguation based on Wikipedia data. In: *Proc. of EMNLP-CoNLL'07*, pp. 708–716.
- Dredze, M., McNamee, P., Rao, D., Gerber, A., Finin, T. (2010): Entity disambiguation for knowledge base population. In: *Proc. of Coling 2010*, pp. 277–285.
- Fiscus, J., Doddington G. (2002): Topic detection and tracking evaluation overview. In: *Topic detection and tracking*, pp. 17–31. Kluwer Academic Publishers.
- Gusfield, D. (1999): *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge University Press.
- Han, X., Zhao, J. (2009): Named entity disambiguation by leveraging wikipedia semantic knowledge. In: *Proc. of CIKM 2009*, pp. 215–224.
- Ji, H., Grishman, R. (2011): Knowledge Base Population: Successful Approaches and Challenges. In: *Proc. of ACL 2011*, pp. 1148–1158.
- Larkey, L.S., Feng, F., Connell, M., Lavrenko, V. (2004): Language-specific models in multilingual topic tracking. In: *Proc. of SIGIR '04*, pp. 402–409.
- Mihalcea, R., Csomai, A. (2007): Wikify!: linking documents to encyclopedic knowledge. In: *Proc. of CIKM '07*, pp. 233–242.
- Nallapati, R., Feng, A., Peng, F., Allan, J. (2004): Event threading within news topics. In: *Proc. of CIKM 2004*, pp. 446–453.
- Ploch, D. (2011): Exploring Entity Relations for Named Entity Disambiguation. In: *Proc. of ACL 2011*, pp. 18–23.
- Shah, C., Croft, W., Jensen, D. (2006): Representing documents with named entities for story link detection (SLD). In: *Proc. of CIKM '06*, pp. 868–869.
- Simpson, H., Strassel, S., Parker, R., McNamee, P. (2009): Wikipedia and the web of confusable entities: Experience from entity linking query creation for TAC 2009 knowledge base population. In: *Proc. of LREC '10*.
- Vapnik, V.N. (1995): *The nature of statistical learning theory*. Springer-Verlag, New York, NY, USA.
- Wayne, C. (2000): Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation. In: *Proc. of LREC '00*.
- Zhang, W., Su, J., Lim, C., Tan W., Wang, T. (2010): Entity linking leveraging automatically generated annotation. In: *Proc. of Coling 2010*, pp. 1290–1298.
- Zheng, Z., Li, F., Huang, M., Zhu, X. (2010): Learning to link entities with knowledge base. In: *Proc. of NAACL-HLT '10*, pp. 483–491.

From Historic Books to Annotated XML: Building a Large Multilingual Diachronic Corpus

Magdalena Jitca, Rico Sennrich, Martin Volk

Institute of Computational Linguistics, University of Zurich

Binzmühlestrasse 14, 8050 Zürich

E-mail: mjitca, sennrich, volk @ifi.uzh.ch

Abstract

This paper introduces our approach towards annotating a large heritage corpus, which spans over 100 years of alpine literature. The corpus consists of over 16.000 articles from the yearbooks of the Swiss Alpine Club, 60% of which represent German texts, 38% French, 1% Italian and the remaining 1% Swiss German and Romansh. The present work describes the inherent difficulties in processing a multilingual corpus by referring to the most challenging annotation phases such as article identification, correction of optical character recognition (OCR) errors, tokenization, and language identification. The paper aims to raise awareness for the efforts in building and annotating multilingual corpora rather than to evaluate each individual annotation phase.

Keywords: multilingual corpora, cultural heritage, corpus annotation, text digitization

1. Introduction

In the project Text+Berg¹ we are digitizing publications of the Alpine clubs from various European countries, which consist mainly of reports on the following topics: mountain expeditions, the Alpine culture, the flora, fauna and geology of the mountains.

The resulting corpus is a valuable knowledge base to study the changes in all these areas. Moreover, it enables the quantitative analysis of diachronic language changes as well as the study of typical language structures, linguistic topoi, and figures of speech in the mountaineering domain.

This paper describes the particularities of our corpus and gives an overview of the annotation process. It presents the most interesting challenges that our multilingual corpus brought up, such as text structure identification, optical character recognition (OCR), tokenization, and language identification. We focus on how the multilingual nature of the text collection poses new problems in apparently trivial processing steps (e.g. tokenization).

¹ See www.textberg.ch

2. The Text+Berg Corpus

The focus of the Text+Berg project is to digitize the yearbooks of the Swiss Alpine Club from 1864 until today. The resulting corpus contains texts which focus on conquering and understanding the mountains and covers a wide variety of text genres such as expedition reports, (popular) scientific papers, book reviews, etc.

The corpus is multilingual and contains articles in German (some also in Swiss German), French, Italian and even Romansh. Initially, the yearbooks contained mostly German articles and few in French. Since 1957 the books appeared in parallel German and French versions (with some Italian articles), summing up to a total of 53 parallel editions German-French and 90 additional multilingual yearbooks. The corpus contains 16.000 articles, 60% of which represent German texts, 38% French, 1% Italian and the remaining 1% Swiss German and Romansh. This brings our corpus to 35,75 million words extracted from almost 87.000 book pages, 10% of which representing parallel texts. This feature of the corpus allows for interesting cross-language comparisons and has been used as training material for Statistical Machine Translation systems (Sennrich & Volk, 2010).

3. The Annotation Phases

This section introduces our pipeline for processing and annotating the Text+Berg corpus. More specifically, the input consists of HTML files containing the scanned yearbooks (for yearbooks in paper format), as they are exported by the OCR software. We work with two state-of-the-art OCR programs (Abbyy FineReader 7 and OmniPage 17) in order to convert the scan images into text and then export the files in HTML format. Our processing pipeline takes them through ten consecutive stages: 1) HTML cleanup, 2) structure reducing, 3) OCR merging, 4) article identification, 5) parallel book combination, 6) tokenization, 7) correction of OCR errors, 8) named entity recognition, 9) Part of Speech (POS) tagging and 10) additional lemmatization for German. The final output consists of XML documents which mark the article structure (title, author), as well as sentence boundaries, tokens, named entities (restricted to mountain, glacier and cabin names), POS tags and lemmas. Our document processing approach is similar to other annotation pipelines, such as GATE (Cunningham et al., 2002), but it is customized for our alpine corpus. In terms of space complexity, the annotated output files require almost three times more storage space than the input HTML files and 2,3 times more space than the tokenized XML files, respectively.

In the following subsections we expand on the processing stages that are especially challenging for a multilingual corpus.

3.1. Article Identification

The identification of articles in the text is performed during the fourth processing stage. The text is annotated conforming to an XML schema which marks the article boundaries (start, end), its title and author, paragraphs, page breaks, footnotes and captions. Some of the text structure information can be checked against the table of contents (ToC) and table of figures (where available), which are manually corrected in order to have a clean database of all articles in the corpus. Another relevant resource for the article boundary identification is the page mapping file that is automatically generated in the second stage, which relates the number printed on the original book page with the page number assigned during scanning. The process of matching entries from

the table of contents to the article headers in the books is not trivial, as it requires that the article title, the author name(s) and the page number in the book are correctly recognized. We allow small variations and OCR errors, as long as they are below a specific threshold (usually a maximum deviation of 20% of characters is allowed). For example, the string *K/albard -Eine Reise in die Eiszeit.* will be considered a match for the ToC entry *Svalbard - Eine Reise in die Eiszeit*, although not all their characters coincide.

Proper text structuring relies on the accurate identification of layout elements such as article boundaries, graphics and captions, headers and footnotes. Over the 145 years the layout of the yearbooks has changed significantly. Therefore we had to adapt different processing steps for all the various designs. The particularities of these layouts have been discussed in (Volk et al., 2010a).

The yearbooks since 1996 are a collection of monthly editions and their pagination is no longer continuous (it starts over every month). This change affects the page mapping process, which performs well only when page numbers are monotonically increasing. Moreover, article boundaries are hard to determine when a single page contains several small articles and not all of them specify their author's name. These particularities are also reflected in the layout, as the header lines (where existing) no longer contain information about author or title, but about the article genre. Under these circumstances, we still achieved a percentage of 80% identified articles for these new yearbooks, a value comparable to the overall percentage of the corpus.

3.2. Correction of OCR Errors

The correction process aims to detect and overcome the errors introduced by the OCR systems and is carried out in two different stages of the annotation process. The first revision is done in the third stage (OCR merging), where the input is still raw text, with no additional information about either the structure or the language of the articles. At this stage we combine the output of our two OCR systems. The algorithm computes the alignments in a page-level comparison of the input files provided by each system and searches the Longest Common Subsequence in a n-character window. In case

of mismatch, the system disambiguates among the different candidates and selects the word with the highest probability in that context (computed based on the word's frequency in the Text+Berg corpus). The implemented algorithm and the evaluation results are thoroughly discussed in (Volk et al., 2010b).

OCR-merging is a worthwhile approach since there are many situations where one system can fix the other's errors. Our experience has shown that Abbyy FineReader performs the better OCR, with over 99% accuracy (Volk et al., 2010b). But there are also cases where it fails to provide the correct output, whereas OmniPage provides the right one. For example, the sequence *Cependant, les cartes disponibles sont squivent approximatives* (English: However, the available maps are often approximate) is provided by FineReader. The system has introduced the spelling mistake *squivent*, which doesn't appear in the output of the second system (here *souvent*). This triggers the replacement of the non-word *squivent* with the correct version *souvent*.

During the seventh annotation stage, after tokenization, we correct errors caused by graphemic similarities. The automatic correction is performed at the word-level by pattern matching over sequences of characters. In order to achieve this, we have compiled lists of common error patterns and their possible replacements. For example, a word-initial 'R' is often misinterpreted as 'K', resulting in words such as *Kedaktion* instead of *Redaktion* (English: editorial office). For each tentative replacement we check against the word frequency list in order to decide whether a candidate word appears in the corpus more frequently than the original or the other possible replacement candidates. In this case, *Redaktion* has 1127 occurrences in the corpus, whereas *Kedaktion* only 9. Reynaert (2008) describes a similar statistical approach for both historical and contemporary texts.

As the yearbooks until 1957 contained articles written in several languages, we have used a single word frequency dictionary for all of them (German, French and Italian). The dictionary has been built from the Text+Berg corpus and thus contains all the encountered word types and their corresponding frequencies, computed over the same corpus. The interesting aspect about this dictionary is its reliability, in spite of being trained with noisy data (text containing OCR-errors).

Correctly spelled words will typically have a higher frequency than the ones containing OCR errors. The list contains predominantly German words due to the high percentage of German articles in the first 90 yearbooks, thus the frequency of German words is usually higher than that of French words. This can lead to wrong substitution choices, such as a German word in a French sentence (e.g. *Neu* (approx. 4400 hits) instead of *lieu* (approx. 3000 hits)). Therefore we have decided to create a separate frequency dictionary for French words, which is used only for the monolingual French editions.

3.3. Tokenization

In this stage the paragraphs of the text are split into sentences and words, respectively. Tokenization is considered to be a straightforward problem that can be solved by applying a simple strategy such as split on all non-alphanumeric characters (e.g. spaces, punctuation marks). Studies have shown, however, that this is not a trivial issue when dealing with hyphenated compound words or other combinations of letters and special characters (e.g. apostrophes, slashes, periods etc.). He and Kayaalp (2006) present a comparative study of several tokenizers for English, showing that their output varies widely even for the same input language. We would expect a similar performance from a general purpose tokenizer dealing with several languages.

We will exemplify the language-specific issues with the use of apostrophes. In many languages, they are used for contractions between different parts of speech, such as verb + personal pronoun *es* in German (e.g. *hab's* → *habe* + *es*) or determiner and noun in French or Italian (e.g. *l'abri* → *le* + *abri*). On the other hand, in old German written until 1900, like in modern English, it can also express possession (e.g. *Goldschmied's*, *Theobald's*, *Mozart's*). Under these circumstances, which is the desired tokenization, before or after the apostrophe? The answer is language-dependent and this underlies our approach towards tokenization.

We use a two-step tokenization and perform the language recognition in between. The advantage of this approach is that we can deliver a language-specific tokenization of any input text (given that it is written in the supported languages). In the first step we carry out a rough tokenization of the text and then identify sentence

boundaries. Once this is achieved, we can proceed to the language identification, which will be discussed in section 3.4.

Afterwards we do another round of tokenization focused on word-level, where the language-specific rules come into play. We have implemented a set of heuristic rules in order to deal with special characters in a multilingual context, such as abbreviations, apostrophes or hyphens. For example, each acronym whose letters are separated by periods (e.g. C.A.S. or A.A.C.Z.) is considered a single token, if it is listed in our abbreviations dictionary. A German apostrophe is split from the preceding word (e.g. *geht's* → *geht* + *'s*), whereas in French and Italian it remains with the first word (e.g. *dell'aqua* → *dell'* + *aqua*, *l'eau* → *l'* + *eau*). Besides, we have compiled a small set of French apostrophe words which shouldn't be separated at all (e.g. *aujourd'hui*).

Disambiguation for hyphens occurring in the middle of a word is performed by means of the general word frequency dictionary. For example, if *nordouest* has 14 hits and *nord-ouest* 957 hits, we conclude that the hyphen is part of the compound and thus *nord-ouest* should be regarded as a single token. On the other hand, hyphens marking line breaks may also appear in the middle, like in the word *rou-te*. In this case, the hyphenated word appears 3 times in the dictionary, whereas the one without, *route*, 6335 times. Therefore the hyphen will be removed from the word.

3.4. Language Identification

The accuracy of the language identification is crucial for the automatic text analysis performed during the annotation process, such as tokenization, part-of-speech tagging, lemmatization or named entity identification. Therefore we perform a fine-grained analysis, at sentence level. We work with a statistical language identifier² based on the approach presented in (Dunning, 1994). The module uses two classifiers: one to distinguish between German, French, English and Italian and another one in order to discriminate between Italian and Romansh. In case the identified language is German, a further analysis based on the frequency dictionary is being carried out in order to decide whether or not it is Swiss German (CH-DE). This dictionary

² <http://search.cpan.org/dist/Lingua-Ident/Ident.pm>

contains frequently used Swiss German dialect words which do not have homographs in standard German. Whenever a sentence contains more than 10% dialect words from this list, the language of the sentence is set to CH-DE.

However, the statistical language identification is not reliable for very short sentences. In order to achieve higher accuracy, we apply the heuristic rule that only sentences longer than 40 characters are fed to the language identifier. All the others are assigned the language of the article, as it appears in the ToC. The correctness of this decision relies on the fact that all ToC files are proofed manually, so that we do not introduce noisy data.

Table 1 gives an overview of the distribution of the identified languages in the articles from the Text+Berg corpus. We present here only the composition of German and French articles, as they represent the great majority of our corpus (approximately 98%). The values are not 100% accurate, as they are automatically computed by means of statistical methods. However, they mirror the global tendencies of the corpus that over 95% of the sentences in an article are in the language of the article, a conclusion which corresponds to our expectations. An interesting finding is the percentage variation of foreign sentences. For example, German sentences are two times more frequent in French articles than the French sentences in German articles (in percentage terms). One reason for this is the fact that some French articles are translated from German and preserve the original bibliographical references, captions or footnotes. Other sources of language mixture are quotations and direct speech, aspects which can be encountered in both German and French articles.

3.5. Linguistic Processing

In the last two annotation stages we perform some linguistic processing, namely lemmatization and part of speech tagging. The markup is done by the TreeTagger³. For our corpus, we have applied the standard configuration files for German, English and Italian. In the case of French we adopted a different approach, and we have trained our own parameter files based on the Le Monde-Treebank (Abeillé, 2003).

³ www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger

Article language	Number of sentences per language						
	de	en	fr	it	rm	ch-de	total
DE	1.166.141	1035	11.607	1481	1490	799	1.182.553
FR	12.392	607	670.599	1187	1277	2	686.064

Table 1: The language distribution of the sentences in the Text+Berg corpus

```

<book id="1901_mul">
  <article n="5">
    <tocEntry title="Altes und Neues aus dem Säntisgebiet"
      author="C. Egloff" lang="de" category="Freie Fahrten"/>
    <div>
      <s n="5-88" lang="de">
        <w n="5-88-1" pos="APPRART" lemma="im">Im</w>
        <w n="5-88-2" pos="NN" lemma="Eiltempi">Eiltempo</w>
        <w n="5-88-3" pos="VVFIN" lemma="gehen">ging</w>
        <w n="5-88-4" pos="PPER" lemma="es">'s</w>
        <w n="5-88-5" pos="PTKVZ" lemma="weiter">weiter</w>
        <w n="5-88-6" pos="$.," lemma=",">,</w>
        <w n="5-88-7" pos="ADV" lemma="erst">erst</w>
        <w n="5-88-8" pos="ADJD" lemma="kletternd">kletternd</w>
        <w n="5-88-9" pos="$.," lemma=",">,</w>
        <w n="5-88-10" pos="ADV" lemma="dann">dann</w>
        <w n="5-88-11" pos="APPR" lemma="über">über</w>
        <w n="5-88-12" pos="NN" lemma="Felstrümmen">Felstrümmen</w>
        <w n="5-88-13" pos="KON" lemma="und">und</w>
        <w n="5-88-14" pos="NN" lemma="Schneefeld">Schneefelder</w>
        <w n="5-88-15" pos="$.," lemma=",">,</w>
      </s>
    </div>
  </article>
</book>

```

Figure 1: An annotation snippet

Romansh is not yet supported due to the lack of a sufficiently large annotated corpus for training the corresponding parameter file. Figure 1 shows a sample output: an annotated sentence in XML format.

The TreeTagger assigns only lemmas for word forms that it knows (that have been encountered during the training). This results in a substantial number of word forms with unknown lemmas. Therefore we use an additional lemmatization tool, in order to increase the coverage of lemmatization. This approach has been implemented for German only because of its large number of compounds.

We use the system Gertwol⁴ to insert missing German lemmas. Towards this goal we collect all word form types from the corpus and have Gertwol analyse them. If the TreeTagger does not assign a lemma to a word, whereas Gertwol provides an appropriate alternative, we choose the output of the latter system. This has resulted in approximately 700.000 additional lemmas, 80% percent of which represent noun lemmas, 15% adjectives and the remaining 5% other parts of speech.

After performing this step, the remaining unknown

⁴<http://www2.lingsoft.fi/cgi-bin/gertwol>

lemmas are mostly names and words containing OCR errors. We are interested in extending this strategy for French and Italian, in order to further increase the coverage of the annotation.

4. Tools for Accessing the Corpus

The Text+Berg corpus can be accessed through several search systems. For example, we have stored our annotated corpus in the Corpus Query Workbench (Christ, 1994), which allows us to browse it via a web interface⁵. The queries follow the POSIX EGREP syntax for regular expressions. The advantage of this system is that it provides more precise results than usual search engines (which perform a full text search) due to our detailed annotations. For example, it is possible to query for all mountain names ending in *horn* that were mentioned before 1900. Moreover, it is also possible to restrict queries to particular languages or POS tags.

In addition, we have built a tool for word alignment searches in our parallel corpus⁶. Given a German search term, the tool displays all hits in the German part of the corpus together with the corresponding French sentences with the aligned word(s) highlighted. Other than being a word alignment visualization tool, it also serves as bilingual concordance tool to find mountaineering terminology in usage examples. In this way it is easy to determine the appropriate translation for words like *Haken* (English: hook) or *Steigeisen* (English: crampon). Moreover, it enables a consistent view of the possible translations of ambiguous words as *Kiefer* (English: jaw, pine) or *Mönch* (English: monk, mountain name). Figure 2 depicts the output of the system for the word *Leiter*, which can either refer to leader or ladder.

⁵Access to the CQW is password-protected. See <http://www.textberg.ch/index.php?id=4&lang=en> for registration.

⁶<http://kitt.ifi.uzh.ch/kitt/alignsearch/>

1959, article 8 Colin Wyatt: <i>Bergfahrt durch Nepal</i>	Eine zweite Leiter führte uns durch eine Deckenöffnung aufs Dach , von wo man das Dorf und das Tal überblickte .	Une autre échelle nous conduisit par un trou du plafond sur le toit en terrasse dominant le village et la vallée .
1959, article 41 G.O. Dyhrenfurth: <i>Himalaya-Chronik 1958</i>	Leiter war Gurdial Singh , Honorary Local Secretary des Himalayan Club in Dehra Dun .	Le chef en était Gourdiyal Singh , secrétaire local de l' Himalayan Club à Dehra Dun .
1984, article 4 Lorenz Seiler, Roland Radlinger: <i>Situationswahrnehmung und Angstentstehung im Bergsport</i>	Innerhalb der Gruppe nimmt der Leiter eine besondere Stellung ein :	A l' intérieur du groupe , le moniteur occupe une place particulière :
1985, article 14 Trevor Braham: <i>Himalaya-Chronik 1984</i>	Zwei der Briten , der Leiter und M. Fowler stiegen weiter bis ca. 7000 m auf , bevor sie umkehrten .	Deux des grimpeurs , le chef de l' expédition et M. Fowler , ont continué l' ascension jusqu' à 7000 mètres environ , où ils ont fait demi-tour .

Figure 2: Different translations of the German word *Leiter* in the Text+Berg corpus

5. Conclusion

In this paper we have given an overview of the annotation workflow of the Text+Berg corpus. The pipeline is capable of processing multilingual documents and dealing with both diachronic varieties in language and noisy data (OCR errors). The flexible architecture of the pipeline allows us to extend the corpus with more alpine literature and to process it in a similar manner, with little overhead.

We have provided insights into the multilingual challenges in the annotation process, such as OCR correction, tokenization or language identification. We intend to further reduce the number of OCR errors by launching a crowd correction wiki page, where the members of the Swiss Alpine Club will be able to correct such mistakes. Regarding linguistic processing, we will continue investing efforts in improving the quality of the existing annotation tools with language-specific resources (e.g. frequency dictionaries, additional lemmatizers). We will also work on improving the language models for Romansh and Swiss German dialects, in order to increase the reliability of the language identifier.

6. References

- Abeillé, A., Clément, L., Toussenet, F. (2003): Building a Treebank for French. In Building and Using Parsed Corpora, Text, Speech and Language Technology(20), pp. 65–187.
- Christ, O. (1994): The IMS Corpus Workbench Technical Manual. Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.
- Cunningham, H., Maynard, D., Bontcheva, K. (2002): GATE: A framework and graphical development environment for robust NLP tools and applications. In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics.
- Dunning, T. (1994): Statistical identification of language. Technical Report MCCA-94-273, New Mexico State University.
- He, Y., Kayaalp, M. (2006): A comparison of 13 tokenizers on MEDLINE. Technical Report LHNCBC-TR-2006-003, The Lister Hill National Center for Biomedical Communications.
- Reynaert, M. (2008): Non-interactive OCR post-correction for giga-scale digitization projects. In A. Gelbukh (Ed.), Proceedings of the Computational Linguistics and Intelligent Text Processing 9th International Conference, Lecture Notes in Computer Science. Berlin, Springer, pp. 617–630.
- Sennrich, R., Volk, M. (2010): MT-based sentence alignment for OCR-generated parallel texts. In Proceedings of AMTA. Denver.
- Volk, M., Bubenhofer, N., Althaus, A., Bangerter, M., Furrer, L., Ruef, B. (2010a): Challenges in building a multilingual alpine heritage corpus. In Proceedings of the Seventh international conference on Language Resources and Evaluation (LREC).
- Volk, M., Marek, T., Sennrich, R. (2010b): Reducing OCR errors by combining two OCR systems. In Proceedings of the ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010).

Visualizing Dependency Structures

Chris Culy, Verena Lyding, Henrik Dittmann

European Academy of Bozen/Bolzano

viale Druso 1, 39100 Bolzano, Italy

E-mail: chris@chrisculy.net, verena.lyding@eurac.edu, henrik.dittmann@eurac.edu

Abstract

In this paper we present an advanced visualization tool specialized for the presentation and interactive analysis of language structure, namely dependency structures. Extended Linguistic Dependency Diagrams (xLDDs) is a flexible tool that provides for the visual presentation of dependency structures and connected information according to the users' preferences. We will explain how xLDD makes use of visual variables like color, shape and position to display different aspects of the data. We will provide details on the technical background and discuss issues with the conversion of dependency structures from different dependency banks. Insights from a small user study will be presented and we will discuss future directions and application contexts for xLDD.

Keywords: dependency structures, dependency diagrams, visualization

1. Introduction

Dependency banks, and hence dependency structures, are becoming ever more widely available for different languages and are popular for a range of applications, from theoretical and applied linguistics research to pedagogy in linguistics and language learning (cf. e.g. the VISL project¹; (Hajič et al., 2001; Nivre et al., 2007)). In this context, also a number of (usually static) visualizations of dependency structures have been presented (Gerdes & Kahane, 2009; Nivre et al., 2006). Generally, visualizations of language and linguistic information (“LInfoVis”, from “Linguistic Information Visualization”) are becoming more widespread (see (Rohrdantz et al., 2010) for an overview), but visualizations targeted specifically at linguists and their informational needs are still not very common. Current attempts to visualize language data are usually either visually very simple or linguistically uninformed, and often very much bound to a specific application context. We are trying to improve this situation with a series of advanced LInfoVis tools. In this paper, we present Extended Linguistic Dependency Diagrams (xLDDs), an example of a LInfoVis tool which combines advanced visualization techniques with linguistic knowledge to create a new kind of interactive dependency diagram. This tool can be easily adapted for a variety of uses in a

variety of environments and can be used with a range of dependency structure formats.

2. Dependency Structures and Dependency Diagrams

We will distinguish between **dependency structures**, which are mathematical objects (graphs), and **dependency diagrams**, which are visual representations of dependency structures. Unfortunately, the linguistics literature does not always maintain this distinction, but it is an important one, since the same dependency structure can have many different visual representations (see e.g. ANNIS² for multiple visual representations of the same structure).

While there is no standard, or even general agreement, about what information should or should not be included in a dependency structure, essentially dependency structures are directed (usually acyclic) graphs that indicate binary head-dependent relations between parts of a sentence (see (Hudson, 1984) for early examples of dependency structures). We will call a dependency structure **basic** if it consists only of the tokens of the sentence and the relations between them, without any additional information. However, almost all dependency structures have more information than just relations between tokens (e.g. often there is lemma or part of speech (POS) information associated with the tokens).

¹ <http://visl.sdu.dk/visl/en/parsing/automatic/dependency.php>

² <http://www.sfb632.uni-potsdam.de/~dl/annis>



Figure 1: Basic linearized xLDD with color coding of parts of speech and dependency types; TiGerDB 8046, structure as in (Boyd et al., 2007)

We will refer to these dependency structures as **advanced**.

We will call a dependency diagram **linearized** if it shows the tokens of the sentence in their typical presentation direction (e.g. left to right for German, right to left for Arabic). Basic dependency structures allow for basic diagrams only, as the information to visualize is restricted to tokens and dependency relations. Figure 1 shows an advanced dependency diagram of an advanced dependency structure, in that it includes a variety of information, including POS information in addition to the tokens and dependency relations. The dependency relations are indicated by directed arcs between the tokens, and the directions of the arrows follow the EAGLES³ recommendation of having the arrow pointing towards the head.

It goes beyond the presentation of a typical linearized diagram in the use of color and in the positioning of the arcs. The POS of the words are encoded by colored nodes and tokens, and hovering over a token shows a tooltip with the POS type, as in Figure 1 NN (noun) for the word “Absage”. Color is also used to distinguish different dependency relations, blue arcs indicate verb–object relations, red arcs indicate verb-subject relations, green is used for modifier relations, gray for determiner-noun relations and black for the root dependency. Furthermore, the positioning of the arcs above and below the text visually separates subject and object relations (arcs below text) from any other type of relation (arcs above text). The example in Figure 1 is based on Boyd et al.’s (2007) reanalysis (to Decca-XML format)⁴ of sentences from the Tiger Dependency Bank (TiGerDB) (Brants et al., 2002). We will have more to say about it shortly.

³ <http://www.ilc.cnr.it/EAGLES96/segsasg1/node44.html>

⁴ We would like to thank Adriane Boyd and Detmar Meurers for kindly providing us with the data they describe in (Boyd et al., 2007).

3. Extended Linguistic Dependency Diagrams

3.1. Visual encoding of information

One of the key ideas of information visualization is that we can use different visual features to encode different aspects of the information being visualized. Dependency structures, especially advanced dependency structures, provide lots of information that we can represent in various ways. xLDDs use three main visual properties to encode information in addition to the basic token and dependency information: position, color and size. These three visual variables are **preattentive**, meaning that we perceive strong differences without having to search for them actively. Information that is encoded in this way stands out among the other information present in the diagram and hence is much easier to locate and identify by the user. For example, in Figure 1 we can immediately find the verbal argument relations by the position of their arcs below the text, and the subject relation by its red color.

Position is used in two ways. First, we can position the arcs above or below the text, using any kind of property, simple or calculated, to determine which arcs are below and which above (as in Figure 1). The second use of position is that of the vertical placement of tokens. By varying the standard vertical placement of tokens (i.e. not all on the same horizontal line) we can also encode certain kinds of information, as e.g. in Figure 2, where words that are split into several tokens are placed one level below the other text. This example shows an alternative reanalysis of the sentence from Figure 1, here based on By’s (2009) reanalysis of sentences from the TiGerDB. By made different choices from Boyd et al. He did not include POS information, and he split compound nouns into several tokens. Hence, we are provided with

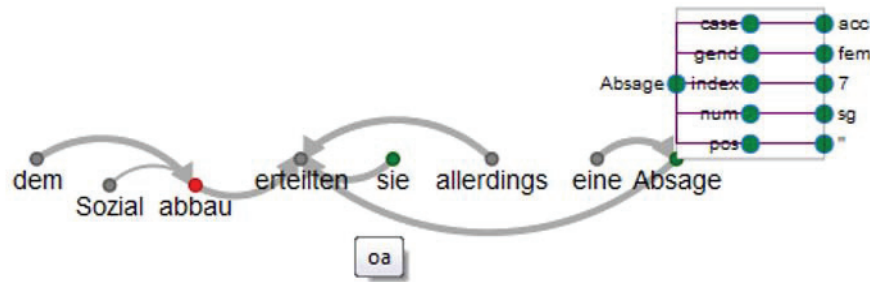


Figure 2: Advanced xLDD, encoding by levels words that are split into multiple tokens; TiGerDB 8046, structure as in (By, 2009)

different information for the visualization. While in Figure 1 color is used on nodes and tokens to encode token-related information and on arcs to encode information on the dependency relations, in Figure 2 the coloring of nodes indicates the linear position of subject/object nodes relative to their heads: subject/object nodes left of their head are colored in red, right of their head in green, between multiple heads in yellow. Nodes of other relations are colored gray.

The visual feature size is employed in xLDDs in form of the thickness of lines of arcs. In Figure 2 it is used to distinguish arcs between sub-words (thin) and any other arc (thick). As in Figure 1, arcs of subject and object relations are placed below the text and others above.

There are several other visual aspects that we could use to encode information. We could, for example, also use the size or style/font of the text, or the shape of the nodes corresponding to the tokens to encode other information. All of these visual encodings in xLDDs, especially the preattentive ones, help (potentially) the user see patterns more quickly and more accurately than a monochrome, uniformly positioned dependency diagram.

3.2. Visual presentation and interaction

Another major hallmark of contemporary visualizations is their adjustability and interactivity. Some aspects of the visualization may not encode information but can be modified to improve readability, or cater to the individual user's preferences. These include curvature and style of arcs, positioning of words, text size, shape of arrow heads. More or less circular arcs, staggered words, and smaller text size help to create compact displays that fit more information on the screen, which can be an advantage for displaying long sentences. Note that the same visual property, e.g. arc width, may either be facultative (when

it doesn't vary within one xLDD diagram) or may be used to encode information (when it does vary). Which setup is most helpful depends on the data to be visualized as well as on the user. Giving the user the flexibility to set those variables, besides setting variables for the visual encoding, is a major benefit of xLDD.

In addition, by interacting with the visualization the user can get more information about the underlying data than can be seen in a static diagram. In the case of xLDDs, the application can provide different kinds of information in response to actions aimed at different parts of the diagram, for example clicking on a token, or its corresponding node, or moving the mouse over an arc or token. In Figure 2, we see that hovering over an arc brings up a tooltip with its relation type (here *oa* (direct object) between "Absage" and "erteilten"). Double-clicking on the node for "Absage", shows token-related information, that is case, number, gender and index information, but no POS information, since it is not available in the underlying data. Since this information does not involve two tokens, it is not represented via arcs in the main diagram. It would also be possible to interactively suppress information, e.g. eliminating all arcs except the ones of interest. As with the visual features, which kinds of interaction serve what kinds of information depends on the particular application, the particular data, as well as on user preferences.

3.3. Architecture and technical details

xLDD is implemented in JavaScript, using the Protovis toolkit (Bostock & Heer, 2009). We have created a simple JSON exchange format for dependency structures (JSDS). Input dependency structures, whether from a fixed local source or from a dynamic web service, are converted into

JSDS before being visualized by the xLDD framework. The xLDD framework contains an extensible visual encoding and interactive component, which allow the application developer complete control over what kinds of information are visually encoded and how, and similarly, what kinds of interactions there are. xLDD is thus intended as a tool that will be incorporated into a website or web application.

Unfortunately, not all dependency structures contain the tokens of the source sentence or their order. Dependency structures following the example of the PARC 700 (King et al., 2003), for example, do not. These structures cannot be visualized as linearized dependency diagrams since they lack the relevant information, and since xLDDs are necessarily linearized, structures of this type cannot be visualized using xLDD. However, often these non-linearizable structures can be converted into linearizable ones. In fact, both of the presented examples are based on the TiGerDB, which does not contain the original tokens, following the model of the PARC 700. In both cases, the original dependency structures have been reanalyzed by other researchers to include the original token and token order information, cf. (Boyd et al., 2007) for Figure 1 and (By, 2009) for Figure 2. However, these conversions to a linearizable form are not trivial, and cannot necessarily be fully automated. An additional point is that the two conversions make different decisions about things like tokenization and POS, and so the resulting dependency structures are different from each other as well as from the original structures.

Thus, in order for a dependency structure to be usable by xLDD, it must meet two conditions: first it must be linearizable (or converted to a linearizable form), and second it must be converted to the JSDS exchange format. Regarding the required exchange format, we have already written converters to JSDS for the CoNLL 2007 Dependency Parsing format⁵, as well as for By's formats and the Decca-XML format (Boyd et al., 2007). Our target format (JSDS format) is quite simple, so that converters for other (linearizable) formats to JSDS (e.g. MALT-XML⁶) would be easy to write.

4. User evaluation, future directions, and conclusion

In other work (Culy et al., 2011), we report on an evaluation study that we did of an earlier version of xLDD. Two usability tests plus the collection of subsequent evaluative feedback were carried out with four subjects with linguistics and language didactics background. For testing the use of the different visual features in xLDD the subjects were asked to find specified dependency relations in nine different xLDD displays (e.g. with and without the coloring of arcs, with different types of leveled and staggered text, etc.). In the tests the users' reactions to xLDD (thinking aloud) and their performance (time and errors for task completion) were recorded. In general, users preferred visual cues over text-based indications (e.g. details in the pop-up window for each lemma) for solving the given tasks. They found color-coding and placement of the arcs to be very useful, with vertical positioning of the text somewhat less so. They also would have preferred to have some control over the visual encodings, which was not possible in the test situation, but is integrated into some of the current sample applications of xLDD in response to the users' requests. Since users did not understand what, if anything, was being encoded by vertical positioning, giving them control over the vertical positioning might have made it more useful. The main negative reaction was to problems with overlapping arrows and text, especially when the figure is zoomed out (i.e. gets smaller). Back on the positive side, there was consensus that xLDD would be useful in language learning and teaching.

Finally, there are issues about how to visualize mismatches between the dependency structure and the original sentence (which are also issues for linearization). One case is that of punctuation, which may not be included in the dependency structure, but which is in the original sentence. While we might visualize only the dependency structure proper, it seems useful for some applications (e.g. language learning) to include the original punctuation.

A second case is that of null elements of a sentence that are included in some dependency structures, e.g. the TiGerDB. For example, the dependency structure for "Was nicht zur Politik wird, hat keinen Zweck."

⁵ <http://nextens.uvt.nl/depparse-wiki/DataFormat>

⁶ <http://w3.msi.vxu.se/~nivre/research/MaltXML.html>

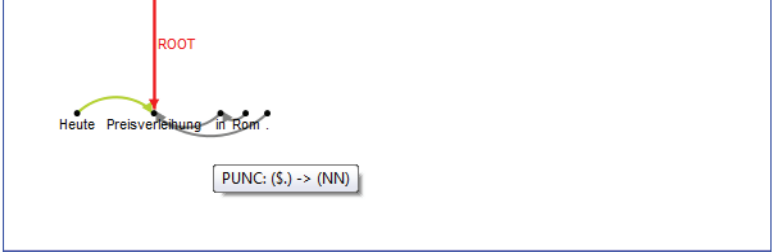
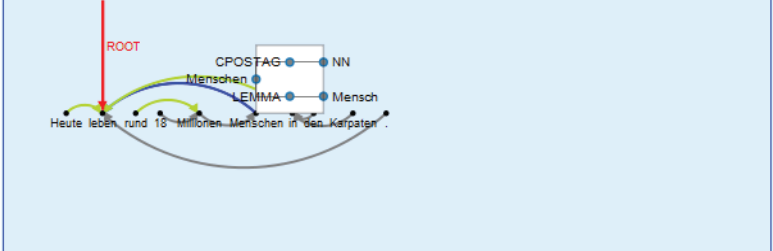
Diagram	Sentence
	Heute Preisverleihung in Rom .
	Heute leben rund 18 Millionen Menschen in den Karpaten .

Figure 3: Presentation of corpus query results in the prototype thumbnails application; sentences matching the query (here “Heute” in corpus of German press releases) are presented as small xLDDs side by side with plain text.

(TiGerDB 8247) has a null subject of “hat”. Since these null elements are not visible parts of the original sentence (no token is representing them), it is not clear how to visualize them. A similar question arises in dealing with multiple information contained within a single token. By (2009) and Boyd et al. (2007) make different decisions in how they handle these cases. For example, By (2009) inserts a null token following “zur” in the same example, corresponding to an empty determiner “der” (dative form of “die”) in the original Tiger structure, but Boyd et al. (2007) do not. This underscores our earlier comment that there is no agreement about the nature of dependency structures. A related issue has to do with abstract nodes, nodes which correspond to a syntactic category rather than to a null token. For example, the dependency diagram in TiGerDB for “Dazu bedarf es Kompetenz und eines gewissen Apparates.” (TiGerDB 8020) contains a node “coord” which is the head of a “coord_form” dependency with “und” as the dependent. “coord” is also the head of two “cj” dependencies with “Apparat” and “Kompetenz” as the dependents. Since “coord_form” is not a token in the sentence, it is not clear how to visualize it and its relations.

A third visualization issue is where tokenization does not agree with orthographic boundaries (e.g. compounds in Tiger, where the compounds are separate elements in the original and in (By, 2009), but not in (Boyd et al., 2007)). We have done some preliminary experiments concerning

these mismatches, but we plan on testing a wider range of examples. Finally, we can point out that all of these mismatches arise from ideas about dependency structures that vary from the idea of representing relations between words.

In addition to addressing the functional difficulties evident in the evaluation, we have created a series of examples and prototype applications using xLDD that also take into account some of the other results of the evaluation. Several of the examples allow the user to specify which linguistic properties are encoded by which visual variables. While we can give the user full control over these encodings, often it is sufficient to use simple specifications of arc position and/or color of the arcs or tokens. Using too many visual variables is just as confusing as using none, or even more so. The specific choices of visual encodings depend on what the user is interested in – there is no single best encoding that encompasses all tasks and interests.

One of the prototypes is an interactive diagram constructor for an on-line textbook. Given a sentence, the student can specify the relations among tokens, and the diagram will be constructed incrementally. It can also be verified against a correct diagram provided by the instructor. A second prototype combines a corpus query engine with xLDD. The search results (obtained via a web service) are presented as a table of the sentences and small versions of the diagrams (as shown in Figure 3). All

these small diagrams can (simultaneously) have their visual encodings adjusted, and on clicking on any of them a larger version of that diagram is presented. These two prototypes underline the point that xLDD is a **component** which can be customized and used in any number of ways, and we hope that it will be adopted and adapted by others (e.g. in the context of CLARIN⁷).

In sum, xLDD is a new way of visualizing dependency structures, which incorporates advanced visualization techniques and provides flexibility for customizing the visualization. Color and position are used to encode information which is omitted or difficult to see in other dependency diagrams. Interaction provides even more opportunities to efficiently explore the structure. The preliminary results of a small-scale user study are promising, and give indications about what needs to be focused on for integration into specialized applications.

5. References

- Bostock, M., Heer, J. (2009): Protovis: A Graphical Toolkit for Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 15(6), pp. 1121–1128.
- Boyd, A., Dickinson, M., Meurers, D. (2007): On representing dependency relations – Insights from converting the German TiGerDB. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories (TLT 2007, Bergen, Norway)*, pp. 31–42.
- Brants, S., Dipper, S., Hansen, S., Lezius, W., Smith, G. (2002): The TIGER Treebank. In *Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT 2002, Sozopol, Bulgaria)*, pp. 24–41.
- Buchholz, S., Marsi, E. (2006): CoNLL-X Shared Task on Multilingual Dependency Parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X, New York City, NY, USA)*, pp. 149–164.
- By, T. (2009): The TiGer Dependency Bank in Prolog format. In *Proceedings of Recent Advances in Intelligent Information Systems (IIS'09, Warsaw, Poland)*, pp. 119–129.
- Culy, C., Lyding, V., Dittmann, H. (2011): xLDD: Extended Linguistic Dependency Diagrams. In *Information Visualization: Proceedings of the 15th International Conference on Information Visualization (IV 2011, London, UK)*, pp. 164–169.
- Gerdes, K., Kahane, S. (2009): Speaking in Piles: Paradigmatic annotation of French spoken corpus. In *Proceedings of the Corpus Linguistics Conference (CL2009, Liverpool, UK)*.
- Hajič, J., Vidová Hladká, B., Pajas, P. (2001): The Prague Dependency Treebank: Annotation Structure and Support. In *Proceedings of the IRCS Workshop on Linguistic Databases (Philadelphia, PA, USA)*, pp. 105–114.
- Hudson, R. (1984): *English Word Grammar*. London: Blackwell.
- King, T.H., Crouch, R., Riezler, S., Dalrymple, M., Kaplan, R.M. (2003): The PARC 700 Dependency Bank. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03, Budapest, Hungary)*.
- Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riel, S., Yuret, D. (2007): The CoNLL 2007 Shared Task on Dependency Parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007 (Prague, Czech Republic)*, pp. 915–932.
- Nivre, J., Hall, J., Nilsson, J. (2006): Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006, Genoa, Italy)*, pp. 2216–2219.
- Rohrdantz, C., Koch, S., Jochim, C., Heyer, G., Scheuermann, G., Ertl, T., Schütze, H., Keim, D.A. (2010): Visuelle Textanalyse. *Informatik-Spektrum*, 33(6), pp. 601–611.

⁷European Research Infrastructure CLARIN, <http://www.clarin.eu>

A functional database framework for querying very large multi-layer corpora

Roman Schneider

Institut für deutsche Sprache
R5 6-13, D-68161 Mannheim
schneider@ids-mannheim.de

Abstract

Linguistic query systems are special purpose IR applications. We present a novel state-of-the-art approach for the efficient exploitation of very large linguistic corpora, combining the advantages of relational database management systems (RDBMS) with the functional MapReduce programming model. Our implementation uses the German DEREKO reference corpus with multi-layer linguistic annotations and several types of text-specific metadata, but the proposed strategy is language-independent and adaptable to large-scale multilingual corpora.

Keywords: corpus storage, multi-layer corpora, corpus retrieval, database systems

1. Introduction

In recent years, the quantitative examination of natural language phenomena has become one of the predominant paradigms within (computational) linguistics. Both fundamental research on the basic principles of human language, as well as the development of speech and language technology, increasingly rely on the empirical verification of assumptions, rules, and theories. More data are better data (Church, & Mercer, 1993): Consequently, we notice a growing number of national initiatives related to the building of large representative datasets for contemporary world languages. Besides written (and sometimes spoken) language samples, these corpora usually contain vast collections of morphosyntactic, phonetic, semantic etc. annotations, plus text- or corpus-specific metadata. The downside of this trend is obvious: Even with specialized applications, our ability to store linguistic data is often bigger than the ability to process all this data.

A lot of essential work towards the querying of linguistic corpora goes into data representation, integration of different annotation systems, and the formulation of query languages (e.g., Rehm et al., 2008; Zeldes et al., 2009; Kepser, Mönnich & Morawietz,

2010). But the scaling problem still remains: As we go beyond corpus sizes of some million words, and at the same time increase the number of annotation systems and search keys, query costs rise disproportionately. This is due to the fact that unlike traditional IR systems, corpus retrieval systems not only have to deal with the “horizontal” representation of textual data, but with heterogeneous metadata on all levels of linguistic description. And, of course, the exploration of inter-relationships between annotations becomes more and more challenging as the number of annotation systems increases. Given this context, we present a novel approach to scale up to billion-word corpora, using the example of the multi-layer annotated German Reference Corpus DEREKO.

2. The Data

The German Reference Corpus DEREKO currently comprises more than four billion words and constitutes the largest linguistically motivated collection of contemporary German. It contains fictional, scientific, and newspaper texts – as well as several other text types – and is annotated morphosyntactically with three competing systems (Connexor, Xerox, TreeTagger). The automated enrichment with additional metadata is underway.

	1 Mio	10 Mio	100 Mio	1000 Mio	4000 Mio
rare (<i>traurige/isoliert/trauen</i>)	0,03s	0,16s	0,3s	0,6s	0,9s
low (<i>langsam/lesen/verfügt</i>)	0,29s	0,35s	0,37s	0,65s	1s
mid (<i>ohne/nun/uns</i>)	0,3s	0,4s	0,5s	1,2s	3,8s
high (<i>nicht/ist/dem</i>)	0,31s	0,47s	1,49s	10,47s	38,7s
top (<i>der/die/und</i>)	0,4s	0,87s	4,67s	47,06s	301s

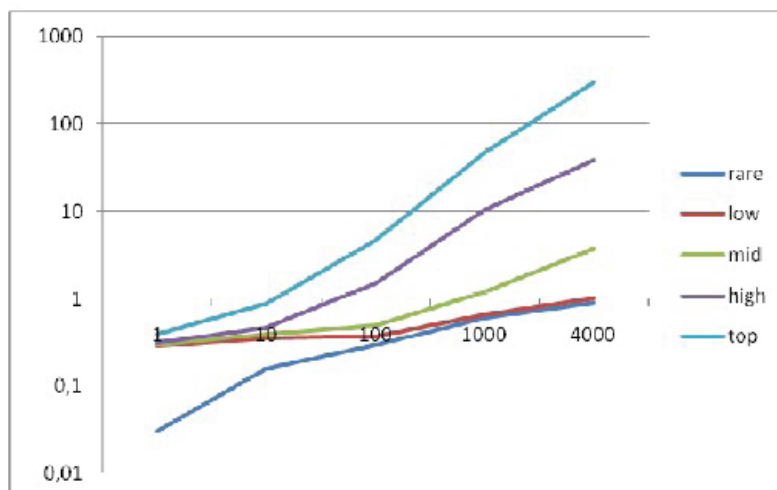


Figure 1: Response times for nested SQL queries with three search keys (logarithmic scaled axis)

3. Existing Approaches

We empirically evaluated the most prominent existing querying approaches, and contrasted them with our functional model (the full paper will contain our detailed series of measurements). Given the reasonable assumptions that XML/SGML-based markup languages are more suitable for data exchange than for efficient storing and retrieval, and that traditional file-based data storage is less robust and powerful than database management systems, we focused on the following strategies:

- i. In-Memory Search: Due to the fact that a computer's main memory is still the fastest form of data storage, there are attempts to implement in-memory databases even for considerably large corpora (Pomikálek, Rychlý & Kilgarriff, 2009). These indexless systems perform well for unparsed texts, but are strongly limited in terms of storage size and therefore cannot deal with data-intensive multi-layer annotations.
- ii. N-Gram Tables: In order to overcome physical limitations, newer approaches use database

management systems and decompose sequences of strings into indexed n-gram tables (Davies, 2005). This allows queries over a limited number of search expressions, but space requirements for increasing values of n are enormous. Sentence-external queries with regular expressions or NOT-queries – both are crucial for comprehensive linguistic exploration – cannot use the n-gram-based indexes and thus perform rather poor.

- iii. Advanced SQL: Another strategy is to make use of the relational power of sub-queries and joins within a RDBMS. Chiarcos et al. (2008) use an intermediate language between query formulation and database backend; Bird et al. (2005) present an algorithm for the direct translation of linguistic queries into SQL. This approach uses absolute word positions, and therefore allows proximity queries without limitation of word distances. But again, even with the aid of the integrated cost-based optimizer (CBO), response times for increasing numbers of search keys become extremely long. We evaluated the proposed strategy on 1, 10, 100, 1000,

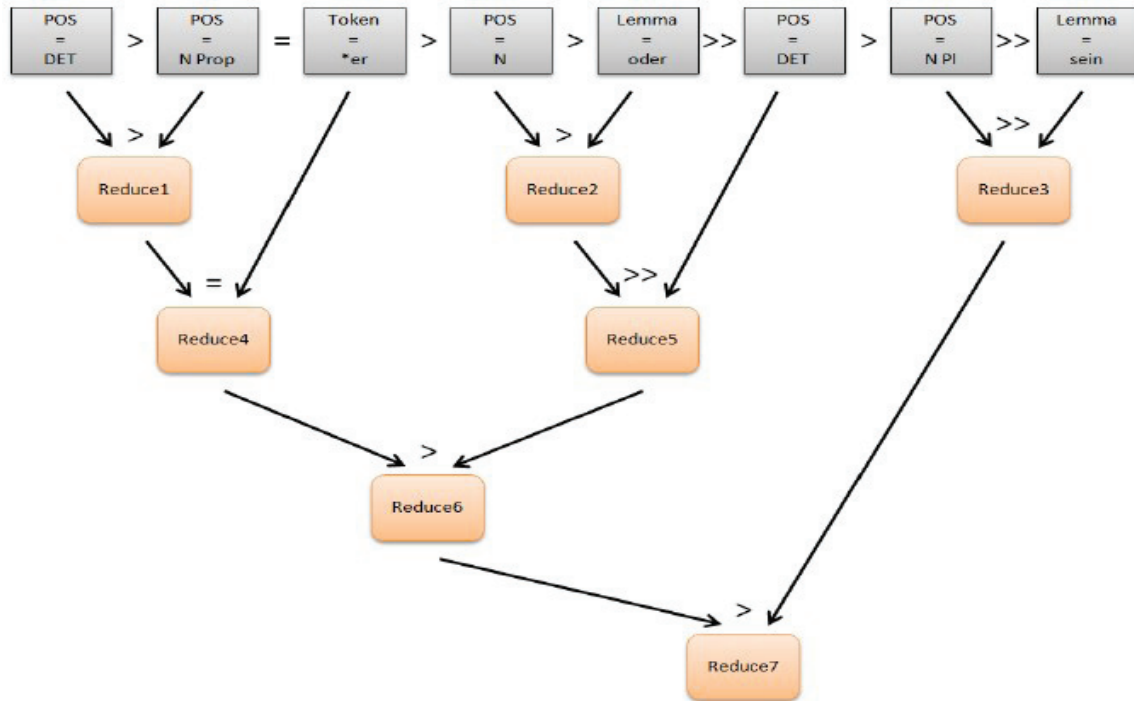


Figure 2: MapReduce processes for a concatenated query with eight search keys

iv. and 4000 million word corpora with rare-, low-, mid-, high-, and top-level search keys and found out that concatenated queries soon exceed the capability of our reference server because nested loops generate an immense workload. Figure 1 shows the response times in seconds for the query “select count(t1.co_sentenceid) from tb_token t1, (select co_id, co_sentenceid from tb_token where co_token=token1) t3, (select co_id, co_sentenceid from tb_token where co_token = token2) t2 where co_token = token3 and t1.co_sentenceid = t2.co_sentenceid and t1.co_sentenceid = t3.co_sentenceid and t1.co_id > t2.co_id and t2.co_id > t3.co_id;”, using three search keys on identical metadata types and a single-column index. This query simply counts the number of sentences that contain three specified tokens (token1, token2, token3) in a fixed order. Compared to a similar query on the 4000 Mio corpus with one search key (5s for a top-level search) or two search keys (56s), the increase of response time is obviously disproportional (301s). It gets remarkably less performant for searches on different metadata types (token, lemma, part-of-speech etc.) using multi-

column indexes. Furthermore, by adding text-specific metadata restrictions like text type or publication year, this querying strategy produces response times of several hours and thereby becomes fully unacceptable for real-time applications.

4. Design and Implementation

As our evaluation shows, existing approaches do not handle queries with complex metadata on very large datasets sufficiently. In order to overcome bottlenecks, we propose a strategy that allows the distribution of data and processor-intensive computation over several processor cores – or even cluster of machines – and facilitates the partition of complex queries at runtime into independent single queries that can be executed in parallel. It is based on two presuppositions:

- i. Mature relational DBMS can be used effectively to maintain parsed texts and linguistic metadata. We intensively evaluated different types of tables (heap tables, partitioned tables, index organized tables) as well as different index types (B-tree, bitmap, concatenated, functional) for the distributed storing and retrieval of linguistic data.

Figure 3: Web-based retrieval form with our sample query

ii. The MapReduce programming model supports distributed programming and tackles large-data problems. Though MapReduce is already in use in a wide range of data-intensive applications (Lin & Dyer, 2010), its principle of “divide and conquer” has not been employed for corpus retrieval yet.

In order to prove the feasibility of our approach, we implemented our corpus storage and retrieval framework on a commodity low-end server (quad-core microprocessor with 2.67 GHz clock rate, 16GB RAM). For the reliable measurement of query execution times, and especially to avoid caching effects, we always used a cold-started 64-bit database engine.)

Figure 2 illustrates the map/reduce processes for a complex query, using eight distinct search keys on

different metadata types: Find all sentences containing a determiner immediately followed by a proper noun ending on “er”, immediately followed by a noun, immediately followed by the lemma “oder”, followed by a determiner (any distance), immediately followed by a plural noun, followed by the lemma “sein” (any distance). Within a “map” step, the original query is partitioned into eight separate key-value pairs. Keys represent linguistic units (position, token, lemma, part-of-speech, etc.), values may be the actual content. Thus, we can simulate regular expressions (a feature that is often demanded for advanced corpus retrieval systems, but difficult to implement for very large datasets).

The queries can be processed in parallel and pass their results (sentence/position) to temporary tables. The

subsequent “reduce” processes filter out inappropriate results step by step. Usually, this cannot be executed in parallel, because each reduction produces the basis for the next step. But our framework, implemented with the help of stored procedures within the RDBMS, overcomes this restriction by dividing the process tree into multiple sub-trees. The reduce processes for each sub-tree are scheduled simultaneously, and aggregate their results after they are finished. So the seven reduce steps of our example can be executed within only four parallel stages.

Our concatenated sample query with eight multi-type search keys on a four billion word corpus took less than four minutes, compared with several hours when employing SQL joins as in 3 (iii). The parallel MapReduce framework is invoked by an extensible web-based retrieval form (see figure 3) and stores the search results within the RDBMS, thus making it easy to reuse them for further statistical processing. Additional metadata restrictions (genre, topic, location, date) are translated into separate map processes and reduced/merged in parallel to the main search.

5. Summary

The results of our study demonstrate that the joining of relational DBMS technology with a functional/parallel computing framework like MapReduce combines the best of both worlds for linguistically motivated large-scale corpus retrieval. On our reference server, it clearly outperforms other existing approaches. For the future, we plan some scheduling refinements of our parallel framework, as well as support for additional levels of linguistic description and metadata types.

6. References

- Church, K., Mercer, R. (1993): Introduction to the Special Issue on Computational Linguistics Using Large Corpora. *Computational Linguistics* 19:1, pp. 1-24.
- Rehm, G., Schonefeld, O., Witt, A., Chiarcos, C., Lehmborg, T. (2008): A Web-Platform for Preserving, Exploring, Visualising and Querying Linguistic Corpora and other Resources. *Procesamiento del Lenguaje Natural* 41, pp. 155-162.
- Zeldes, A., Ritz, J., Lüdeling, A., Chiarcos, C. (2009): ANNIS: A Search Tool for Multi-Layer Annotated Corpora. *Proceedings of Corpus Linguistics 2009*. July 20-23, Liverpool, UK.
- Kepser, S., Mönnich, U., Morawietz, F. (2010): Regular Query Techniques for XML-Documents. Metzger, D., Witt, A. (Eds): *Linguistic modeling of information and Markup Languages*, Springer, pp. 249-266.
- Pomikálek, J., Rychlý, P., Kilgarriff, A. (2009): Scaling to Billion-plus Word Corpora. *Advances in Computational Linguistics* 41, pp. 3-13.
- Davies, M. (2005): The advantage of using relational databases for large corpora. *International Journal of Corpus Linguistics* 10 (3), pp. 307-334.
- Chiarcos, C., Dipper, S., Götze, M., Leser, U., Lüdeling, A., Ritz, J., Stede, M. (2008): A Flexible Framework for Integrating Annotations from Different Tools and Tag Sets. *Traitement Automatique des Langues* 49(2), pp. 271-293.
- Bird, S., Chen, Y., Davidson, S., Lee, H., Zhen, Y. (2005): Extending Xpath to Support Linguistic Queries. *Workshop on Programming Language Technologies for XML (Plan-X)*.
- Lin, J., Dyer, C. (2010): Data-Intensive Text Processing with MapReduce. *Morgan & Claypool Synthesis Lectures on Human Language Technologies*.

Hybrid Machine Translation for German in taraXÜ: Can translation costs be decreased without degrading quality?

Aljoscha Burchardt, Christian Federmann, Hans Uszkoreit

DFKI Language Technology Lab

Saarbrücken & Berlin, Germany

E-mail: {burchardt,cfedermann,uszkoreit}@dfki.de

Abstract

A breakthrough in Machine Translation is only possible if human translators are taken into the loop. While mechanisms for automatic evaluation and scoring such as BLEU have enabled fast development of systems, these systems have to be used in practice to get feedback for improvement and fine-tuning. However, it is not clear if and how systems can meet quality requirements in real-world, industrial translation scenarios. taraXÜ paves the way for wide usage of hybrid machine translation for German. In a joint consortium of research and industry partners, taraXÜ integrates human translators into the development process from the very beginning in a post-editing scenario collecting feedback for improvement of its core translation engines and selection mechanism. taraXÜ also performs pioneering work by integrating languages like Czech, Chinese, or Russian, that are not well studied to-date.

Keywords: Hybrid Machine Translation, Human Evaluation, Post-Editing

1. Introduction

Machine Translation (MT) is a prime application of Language Technology. Research on Rule-Based MT (RBMT) goes back the early days of Artificial Intelligence in the 1960s and some systems have reached a high level of sophistication (e.g. Schwall & Thurmair, 1997; Alonso & Thurmair, 2003). Since the mid 1990, Statistical MT (SMT) has become the prevalent paradigm in the research community (e.g. Koehn et al., 2007; Li et al., 2010). In the translation and localization industry, Translation Memory Systems (TMS) are used to support human translators by making informed suggestions for recurrent material that has to be translated.

As human translators can no longer satisfy the constantly raising translation need, important questions that need to be investigated are:

- 1) How good is MT quality today, especially for translation from and to German?
- 2) Which paradigm is the most promising one?
- 3) Can MT aid human translators and can it help to reduce translation costs without sacrificing quality?

These questions are not easy to answer and it is clear that research on the matter is needed. The quality of MT output cannot be objectively assessed in a once-and-for-all measure (see e.g. Callison-Burch et al., 2006) and it also strongly depends on the nature of the

input material. Various MT paradigms have different strengths and shortcomings, not only regarding quality. For example, RBMT allows for a good control of the overall translation process, but setting up and maintaining such a system is very costly as it requires trained specialists. SMT is cheap, but it requires huge amounts of compute power and training data, which can make it difficult to include new languages and domains. TMS can produce human quality, but are limited in coverage due to their underlying design. Finally, the question of how human translators can optimally be supported in their translation workflow has largely been untouched.

Machine Translation for German The number of available mono- and bi-lingual resources for German is quite high. In the “EuroMatrix”¹ which collects resources, corpora, and systems for a large number of language pairs, German ranges on the third place behind English and French. Still, only little research has been focused on MT for language pairs including German, especially for translation tasks to and from languages other than English.

¹ <http://www.euromatrixplus.net/matrix/>

Source: Empfehlung für die zweite Lesung Piecyk (A5-0232/2000)
 Translation: Recommendation for the second reading Piecyk (A5-0232/2000)

Reset (Ctrl-Alt-R)

Please check the two most severe error classes which apply for the shown sentence.

- Missing content word(s)
- Content word(s) wrong in meaning
- Wrong functional word(s)
- Incorrect word form(s)
- Incorrect word order
- Incorrect punctuation
- Other error

Submit (Ctrl-Alt-S)

Whenever extra commenting is necessary, put your comments here...

Figure 1: Error classification interface used within taraxÜ .

This paper reports on taraxÜ², which aims to address the aforementioned questions in a consortium consisting of partners from both research and industry. taraxÜ takes the selection from hybrid MT results including RBMT, TMS, and SMT as the first part of its analytic process. Then a self-calibration³ component applies, extended by controlled language technology and human post-processing to match real-world translation concerns. A novelty in this project is that human translators are integrated into the development process from the very beginning: Within several human evaluation rounds, the automatic selection and calibration mechanisms will be refined and iteratively improved. This paper focuses on hybrid translation (Section 2) and the large-scale human evaluation rounds in taraxÜ (Section 3). In the conclusion and outlook (Section 4), ongoing and future research is sketched.

2. Hybrid Machine Translation

Hybrid MT is a recent trend (e.g. Federmann et al., 2009; Chen et al., 2009) for leveraging the quality of MT. Based on the observation that different MT systems often have complementary strengths and weaknesses, different methods for hybridization are investigated that aim to “fuse” an improved translation out of the good parts of several translation candidates.

Complementary Errors Typical difficulties for SMT are morphology, sentence structure, long-range re-ordering, and missing words, while strengths are disambiguation and lexical choice.

RBMT systems are typically strong in morphology, sentence structure, have the ability to handle long-range phenomena, and also ensure completeness of the resulting translation. Weaknesses arise from parsing errors and wrong lexical choice. The following examples illustrate the complementary nature of such systems’ errors.

- 1) **Source:** Then, in the afternoon, the visit will culminate in a grand ceremony, at which Obama will receive the prestigious award.
- 2) **RBMT**⁴: Dann wird der Besuch am Nachmittag in einer großartigen Zeremonie gipfeln, an der Obama die berühmte Belohnung bekommen wird.
- 3) **SMT**⁵: Dann am Nachmittag des Besuchs in beeindruckende Zeremonie mündet, wo Obama den angesehenen Preis erhalten werden.

As you can see in the translation of Example 1), the RBMT system generated a complete sentence, yet with a wrong lexical choice for award. The SMT system on the other hand generated the right reading, but made morphological errors and did not generate a complete German sentence. In the translation of Example 4), a parsing error in the analysis phase of the RBMT system led to an almost unreadable result while the SMT decoder gener-

² <http://taraxu.dfki.de/>

³ Due to limited space, this won’t be discussed herein.

⁴ System used: Lucy MT (Alonso & Thurmair, 2003)

⁵ System used: phrase-based Moses (Koehn et al., 2007)

Source:	Empfehlung für die zweite Lesung Piecyk (A5-0232/2000)
System C:	Recommendation for second reading of the Piecyk report (A5-0232 / 2000)
Please do a minimal post-correction for the selected sentence.	
Recommendation for second reading of the Piecyk report (A5 - 0232 / 2000)	
Submit	(Ctrl-Alt-S) Reset (Ctrl-Alt-R)

Figure 2: Post-editing interface used within taraXÜ.

ated a generally intelligible translation, yet with stylistic and formal deficits.

- 4) **Source:** Right after hearing about it, he described it as a “challenge to take action.”
- 5) **RBMT:** Nachdem er richtig davon gehört hatte, bezeichnete er es als eine “Herausforderung, um Aktion auszuführen.”
- 6) **SMT:** Gleich nach Anhörung darüber, beschrieb er es als eine “Herausforderung, Maßnahmen zu ergreifen.”

Hybrid combination can hence lead to better overall translations.

A Human-centric Hybrid Approach In contrast to other hybrid approaches; taraXÜ is in the first place designed to support human post-editing, e.g., in a translation agency. Two different modes have to be handled by the project’s selection mechanism:

- **Human post-editing:** Select the sentence that is easiest to post-edit and have the user edit it.
- **Standalone MT:** Select the overall best translation and present it to the user.

For the translation of 4), the best selection in Standalone MT mode would probably be 6), which is a useful translation, e.g., for information gisting. In Human post-editing mode, 5) would be a better selection as it can relatively quickly be transformed into 7), which is a human-quality translation.

- 7) **Human edit of 5):** Gleich, nachdem er davon gehört hatte, bezeichnete er es als eine “Herausforderung, zu handeln.”

One goal of taraXÜ is the design and implementation of such a novel selection mechanism; however this is still work in progress and will be described elsewhere. Apart from properties of the source sentence (domain, complexity, etc.) and the different translations (grammatical

correctness, sentence length, etc.), the selection mechanism will also take into account “metadata” of the various systems involved such as runtime, number of out-of-vocabulary-warnings, number of different readings generated, etc.

One industry partner in the project consortium provides modules for language checking that will not only be used in the selection mechanism, but also in pre-processing of the input. Starting from the observation that many translation problems arise from problematic input, another goal of taraXÜ is to develop automatic methods for pre-processing input before it is sent to MT translation engines.

3. Large-Scale Human Evaluation

Several large-scale human evaluation rounds are foreseen within the duration of taraXÜ, mainly for the calibration of both the selection mechanism as well as the pre-editing steps, but also for measuring the time needed for post-editing, and for getting a detailed error classification on the translation output from the various MT systems under investigation. The evaluation rounds are performed by external Language Service Providers that usually offer human translation services and hence are considered to act as non-biased experts.

Evaluation Procedure The language pairs that will be implemented and tested during the runtime of taraXÜ are listed in Table 1.

German	↔	English French Japanese Russian Spanish
English	↔	Chinese Czech

Table 1: Language pairs treated in taraXÜ.

We use an extended version of the browser-based evaluation tool Appraise (Federmann, 2010) to collect human judgments on the translation quality of the various systems under investigation in taraXÜ. A screen-shot of the error classification interface can be seen in Figure 1, the post-editing view is presented in Figure 2.

Pilot Evaluation Round The first (pilot) evaluation round of taraXÜ includes the language pairs EN→DE, DE→EN, and ES→DE. The corpus size per language pair is about 2,000 sentences, the data taken mainly from previous WMT shared tasks, but also extracted from freely available technical documentation. Two evaluation tasks will be performed by the human annotators, mirroring the two modi of our selection mechanism:

- 1) In the first task, the annotators have to rank the output of four different MT systems depending on their translation quality. In a subsequent step, they are asked to classify the two main types of errors (if any) of the chosen best translation. We use a subset of the error types suggested by (Vilar et al., 2006), as shown in Figure 1.
- 2) The second task for the human annotators in the first evaluation round is selecting the translation that is easiest to post-edit and to perform the editing. Only a minimal post-editing should be performed.

Some very first results of the ongoing examination of the first human evaluation round are shown in Table 2. The top of the table shows the over-all ranking among the four listed systems, bold face indicates the best system. Below are the results for translation from Spanish and English into German, respectively. On the bottom of the table, overall results on selected corpora are shown from the news domain (1,030 sentences from the WMT-2010 news test set of Callison-Burch et al. (2010), sub-sampled proportionally to each one of its documents) and from the technical documentation of the OpenOffice project.

One observation is that the systems' ranks are comparably close except for Trados, which is not a proper MT system. The very good result of Trados on the news corpora requires further investigation. A noticeable result is that Google performs worst on the WMT corpus although the data should—in principle—have been available online for training; this will also require some more detailed inspection. The latter might, however, explain the good performance of the web-based system on the OpenOffice corpus.

	Lucy	Moses	Trados	Google
Overall	2.00	2.38	3.74	1.86
DE-EN	2.01	2.46	3.80	1.73
ES-DE	1.85	2.42	3.72	1.99
EN-DE	2.12	2.28	3.71	1.89
WMT10	2.52	2.59	2.21	2.69
OpenOffice	1.72	2.77	3.95	1.56

Table 2: First human ranking results, as the average rank of each system in each task.

4. Conclusions and Outlook

In this paper, we have argued and shown evidence that a human-centric hybrid approach to Machine Translation is a promising way of integrating this technology into industrial translation workflows. Even in this early stage, taraXÜ has generated positive feedback and raised interest, especially on the side of the industry partners. We reported early results from the first (pilot) evaluation of taraXÜ, including language pairs EN→DE, DE→EN, and ES→DE. After analyzing the results of this pilot, further evaluation rounds will iteratively extend the numbers of languages covered and include questions related to topics such as controlled language, error types, and the effect of different subject domains. In the presentation of this paper, we will include a more detailed discussion of the first evaluation results.

5. Acknowledgements

This work has partly been developed within the taraXÜ project financed by TSB Technologiestiftung Berlin – Zukunftsfonds Berlin, co-financed by the European Union – European fund for regional development. This work was also supported by the EuroMatrixPlus project (IST-231720) that is funded by the European Community under the Seventh Framework Programme for Research and Technological Development.

6. References

- Alonso, J. A., Thurmair, G. (2003): The compendium translator system. In Proceedings of the Ninth Machine Translation Summit.
- Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocki, M., Zaidan, O. (2010): Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, pp. 17–53, Uppsala, Sweden. Association for Computational Linguistics. Revised August 2010.
- Callison-Burch, C., Osborne, M., Koehn, P. (2006): Re-evaluating the role of bleu in machine translation research. In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, pp. 249–256.
- Chen, Y., Jellinghaus, M., Eisele, A., Zhang, Y., Hunsicker, S., Theison, S., Federmann, C., Uszkoreit, H. (2009): Combining multi-engine translations with Moses. In Proceedings of the Fourth Workshop on Statistical Machine Translation, pp. 42–46, Athens, Greece. Association for Computational Linguistics.
- Federmann, C. (2010): Appraise: An open-source toolkit for manual phrase-based evaluation of translations. In Proceedings of the Seventh conference on International Language Resources and Evaluation. European Language Resources Association (ELRA).
- Federmann, C., Theison, S., Eisele, A., Uszkoreit, H., Chen, Y., Jellinghaus, M., Hunsicker, S. (2009): Translation combination using factored word substitution. In Proceedings of the Fourth Workshop on Statistical Machine Translation, pp. 70–74, Athens, Greece. Association for Computational Linguistics.
- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C. J., Bojar, O., Constantin, A., Herbst, E. (2007): Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pp. 177–180, Prague, Czech Republic. Association for Computational Linguistics.
- Li, Z., Callison-Burch, C., Dyer, C., Ganitkevitch, J., Irvine, A., Khudanpur, S., Schwartz, L., Thornton, W., Wang, Z., Weese, J., Zaidan, O. (2010): Joshua 2.0: A toolkit for parsing-based machine translation with syntax, semirings, discriminative training and other goodies. In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, pp. 133–137, Uppsala, Sweden. Association for Computational Linguistics.
- Papineni, K., Roukos, S., Ward, T., Zhu, W.-J. (2001): Bleu: a method for automatic evaluation of machine translation. IBM Research Report RC22176 (W0109-022), IBM.
- Schwall, U., Thurmair, G. (1997): From metal to t1: systems and components for machine translation applications. In Proceedings of the Sixth Machine Translation Summit, pp. 180–190.
- Vilar, D., Xu, J., D’Haro, L. F., and Ney, H. (2006): Error Analysis of Machine Translation Output. In International Conference on Language Resources and Evaluation, pp. 697–702, Genoa, Italy.

Annotation of Explicit and Implicit Discourse Relations in the TüBa-D/Z Treebank

Anna Gastel, Sabrina Schulze, Yannick Versley, Erhard Hinrichs

SFB 833, Universität Tübingen

E-mail: (yannick.versley|erhard.hinrichs|sabrina.schulze)@uni-tuebingen.de, anna.gastel@student.uni-tuebingen.de,

Abstract

We report on an effort to add annotation for discourse relations, discourse structure, and topic segmentation to a subset of the texts of the Tübingen Treebank of Written German (TüBa-D/Z), which will allow the study of discourse relations and discourse structure in the context of the other information currently present in the corpus (including syntax, referential annotation, and named entities). This paper motivates the design decisions taken in the context of existing annotation schemes for RST, SDRT or the Penn Discourse Treebank, provides an overview over the annotation scheme and presents the result of an agreement study. In the agreement study, we use the notion of *inter-adjudicator agreement* to show that the task of discourse annotation, while challenging in principle, can be successfully solved when using appropriate heuristics.

Keywords: discourse, annotation, text segmentation, agreement

1. Introduction

Discourse information has been proven useful for a number of tasks, including summarization (Schilder, 2002) and information extraction (Somasundaran et al., 2009). While coreference corpora exist for many languages, and in large and very large sizes (frequently over one million words), the annotation of discourse structure and discourse relations has only recently gained the interest of the community at large.

Many of the existing corpora containing discourse structure and/or discourse relations are tightly bound to existing discourse theories such as Rhetorical Structure Theory (RST, Mann & Thompson, 1988) or Segmented Discourse Representation Theory (Asher, 1993), or subscribe to a fundament of coherence relations while avoiding assumptions about discourse structure (Hobbs, 1985; Wolf & Gibson, 2005).

While annotation guidelines for corpora such as the RST Discourse Treebank (Carlson et al., 2003; see Stede 2004, and van der Vlieth et al., 2011 for German and Dutch corpora, respectively, following these guidelines), an SDRT corpus (Hunter et al., 2007), or the Penn

Discourse Treebank (PDTB, Prasad et al., 2007; see Al-Saif & Markert, 2010 for an effort towards an Arabic counterpart) generally agree on the idea of discourse relations between discourse segments, they do differ in other important aspects: RST (in particular, Carlson & Marcu, 2001) and the SDRT guidelines of (Reese et al., 2007) start from **elementary discourse units** (EDUs) that form the lowest level of a hierarchical structure; the PDTB's guidelines avoid the notion of discourse units, elementary or not, by asking annotators to mark **connective arguments** which may, but do not have to, coincide with syntactic or larger units, and do not need to form a hierarchy.

In terms of the relation inventory, the most important desideratum consists in reconciling descriptive adequacy for the linguistic phenomena involved with an inventory size that can still be annotated reliably. This problem is solved in different ways: The RST guidelines contain a coarse level of 16 relation classes, which are further specified into 78 relations which are organized by **nuclearity** (where mononuclear relations put greater weight on one of the units, the nucleus, whereas

CONTINGENCY [28.8%]	EXPANSION [43.6%]
Causal [20.5%]	Elaboration [23.6%]
(c)Result-Cause (5.9%)	(s)Restatement (10.9%)
(c)Result-Enable (4.7%)	(s)Instance (3.4%)
(c)Result-Epistemic (0.4%)	(s)InstanceV (1.0%)
(c)Result-Speechact (0.4%)	(s)Background (9.1%)
(s)Explanation-Cause (6.6%)	Interpretation [4.2%]
(s)Explanation-Enable (1.2%)	(s)Summary (1.0%)
(s)Explanation-Epistemic (1.1%)	(s)Commentary (3.3%)
(s)Explanation-Speechact (0.6%)	Continuation [6.8%]
Conditional [3.0%]	(c)Continuation (6.4%)
(c)Consequence (2.1%)	TEMPORAL [14.35%]
(c)Alternation (0.5%)	(c)Narration (9.3%)
(c)Condition (0.5%)	(s)Precondition (2.4%)
Denial [5.6%]	COMPARISON [11.1%]
(c)ConcessionC (4.0%)	(c)Parallel (3.3%)
(s)Concession (2.0%)	(c)ParallelV (1.1%)
(s)Anti-Explanation (0.5%)	(c)Contrast (7.0%)
	REPORTING [9.5%]
	(s)Attribution (4.2%)
	(s)Source (6.0%)

Table 1: Taxonomy of discourse relations with corpus frequencies

multinuclear relations connect units that are equally important); Reese et al's guidelines for SDRT annotation do not posit any larger categories among their 14 relations, but organize them by a distinction between **coordinating** and **subordinating** relations (cf. Asher & Vieu, 2005; this distinction vaguely corresponds to RST's notion of nuclearity), as well as by **veridicality** (where a relation is veridical if the larger unit containing it cannot be asserted without also asserting the truth of the relation arguments). The PDTB, in contrast, contains 30 relations which are organized into a taxonomy with 16 relations at the middle level and 4 relatively coarse top-level classes (Temporal, Contingency, Comparison, Expansion).

For someone aiming to annotate a corpus with discourse structure, the choice is not easy: The Penn Discourse Treebank carefully avoids any strong commitments to the ideas it uses as a backdrop (such as Webber 2004; Knott et al., 2001), treating the annotation more like a collection of examples that can be mined to verify aspects of the theory; Al-Saif and Markert (2010), for their work on PDTB-style annotation of Arabic discourse, found it necessary to drastically simplify the annotation scheme (from 30 to 12 relations) in order to yield a feasible scheme for their annotation of explicit discourse connectives.

Rhetorical Structure Theory, the most mature of the models for an annotation scheme, has also drawn a commensurate amount of (oftentimes valid) criticism:

The most important one is that RST defines its relations in terms of speaker intentions, which yields good descriptive adequacy (given an appropriate inventory of relations), but fares less well for cognitive plausibility (cf. the overview of critiques in Taboada & Mann, 2006), with Sanders and Spooren (1999) claiming that RST lacks a separation between **intentions**, which are defined in terms of speaker and hearer, and their goals (as is customary in RST), and **coherence relations**, which connect two propositions. In a similar vein, Stede (2008) puts forward the claim that RST's notion of nuclearity encompasses criteria on different linguistic levels that are not always in agreement with each other. Despite SDRT's focus on coherence relations and its strong theoretical commitment on coherence relations and their role in structuring the text, attempts to realize these principles in a general scheme for the discourse annotation of text have been few and far in-between, with the unpublished corpus of Hunter et al (2007) being the most notable example.

Hierarchical structuring of discourse is a well-established concept, not only because it reflects the principles that have been successful in structural accounts of syntax (see Polanyi & Scha, 1983; Grosz & Sidner, 1986, or Webber, 1991, *inter alia*), but also because it allows us to formulate well-formedness (coherence) constraints, as well as accessibility (Webber, 1991) in terms of local configurations.

While such a tree structure is classically motivated through intentional notions (the *discourse segment purposes* of Grosz & Sidner, 1986), the notion of *question under discussion* has been used in information structure to explain intonational focus in terms of (a hierarchy of) question under discussion (van Kuppevelt, 1995; Roberts, 1996; Büring 2003; also Polanyi et al., 2003 for a related proposal). It also allows to couch well-formedness in terms of valid sub-questions (for subordination) or being (non-exhaustive) answers to a common question (for coordination; cf. Txurruka, 2003). Hence, we have, in addition to object-level relations (part-of, causality), an additional level of relations such as *Contrast* which are explainable in terms of information-structural notions, and which yet fulfill the intuition (made explicit by Roberts, 1996) that at any given point in discourse, interlocutors have a common notion of the discourse structure. This level is distinct from the upper-level structure that is the result of conscious structuring of the writer (possibly following genre-specific rules). As an example, some of the very general RST relations such as *Motivation* or *Preparation* are only explainable in terms of writer intentions and conscious text structuring, which may or may not be transparent to the average recipient.

Our own annotation scheme reflects van Kuppevelt's and Roberts' intuitions about a shared structure in discourse: We found it important to keep a backbone of explicit hierarchical structure, as in RST's annotation scheme, but also to avoid vague relations between large text segments, which are often genre-specific or the (sometimes idiosyncratic) result of intentional text structuring by the author. The PDTB successfully uses the metaphor of **implicit connectives** to limit discourse relations to connective-argument-sized pieces; in our case, we reconcile an explicit notion of (shallow) hierarchy with a focus on coherence relations by dividing the text into topically coherent stretches (as discussed, e.g., by Hearst, 1997), which we call **topic segments**, and annotate hierarchical discourse structure (using SDRT's notion of co- and subordinating discourse relations) inside these topic segments.

In the following text, section 2 gives more details on the corpus and on the annotation scheme, whereas section 3 presents an experiment to establish the reliability of our

scheme using an inter-annotator agreement study. Section 4 presents and summarizes our findings.

2. Corpus and Annotation Scheme

As a textual basis for the corpus, we selected newspaper articles from the syntactically and referentially annotated TüBa-D/Z corpus (Telljohann et al., 2009), with the current version totalling 919 sentences in 31 articles, or about 29.6 sentences/article (against 20.6 sentences/article on average in the complete TüBa-D/Z, which also includes very brief newswire-style reports), and altogether 1159 discourse relations and 103 topic segments (or about 9 sentences per topic segment).

The relation inventory, and the distribution of different relation types, is presented in Table 1. From the starting point of the coordinating and subordinating discourse relations in Reese et al., we found it necessary to introduce finer distinctions in some places to ensure either consistency with a related effort on annotating explicit connectives (adding new relations such as *Result-enable* which corresponds to the *Weak-Result* relation proposed by Bras et al., 2006, for SDRT), but also the distinction between **Contrast** and **Concession** which is found in both the Penn Discourse Treebank and the RST annotation guidelines, but not Reese et al.'s proposal.

The resulting 28 relations can be grouped into 8 medium-level and 5 upper-level relation types by considering properties such as **basic operation** (causal vs. additive vs. temporal, with referential as a new group to account for elaborative relations) and **symmetry** as proposed by Sanders et al (1992); the resulting higher-level types of discourse relations have much in common with the top-level taxonomic categories of the Penn Discourse Treebank with a small number of exceptions (the PDTB subsumes the non-symmetrical *Concession* relation under the label *Comparison* whereas we follow Sanders et al. in assuming a **causal** source of coherence for *Concession* and an **additive** source of coherence for the symmetrical *Contrast* relation; Our **Reporting** group includes the *Attribution* and *Source* relations that Hunter et al. use in accounting for reported facts, whereas the Penn Discourse Treebank, unlike RST and SDRT, treats attribution as an issue that is orthogonal to discourse structure).

The hierarchical organization of relations according to basic operation does not differentiate between additional properties such as coordination/subordination or veridicality. Examples (1) and (2) serve to illustrate this distinction:¹

- (1) a) Private Unternehmen dürfen die Telefonbücher der Telekom-Tochter DeTeMedien nicht ohne deren Erlaubnis zur Herstellung einer Telefonauskunfts-CDs verwenden.
b) Die beklagten Unternehmen müssen den Vertrieb der Info-CDs sofort einstellen.

Result-Cause(1a,1b)

- (2) a) Taxifahrer sind als Kolumnenthema eigentlich tabu,
b) weil sie als "weiche Angriffsziele" gelten.

Explanation-Cause(2a,2b)

When the situation specified in Arg1(1a) is interpreted as the cause of the situation specified in Arg2 (1b), the relation between those two arguments is labeled Result-Cause. Both arguments are necessary for coherence, so they are coordinated. The second example is labeled Explanation-Cause, because the situation specified in Arg1(2a) is interpreted as the result of the situation specified in Arg2 (2b). The situation in (2a) contains the main information while the situation in (2b) contributes background information. With subordinating relations, Arg2 ('further information') is always subordinated to Arg1 ('main information'), independently of surface order, as you can see in the following two examples:

- (3) a) Zwei Ex-Mafiosi behaupten zudem,
b) von dem Mordauftrag Andreottis gewußt zu haben.

Attribution(3a,3b)

- (4) a) Nach Angaben von Polizeipräsident Hagen Saberschinsky
b) haben Polizeibeamte einen ihrer Kollegen angezeigt.

Source(4b,4a)

In example (3) the main information is situated in Arg1: It is relevant for the coherence of the text to know that two mobsters testified knowing about the murder contract of Andreotti, which makes them important witnesses in the murder charges against Andreotti.

¹TüBa-D/Z sentences 2563/2564, 7482/7483

Therefore Arg2 is subordinated to Arg1. In example (4) the main information, namely that police officers press charges against one of their colleagues, is given by (4b). Therefore, 4b is the Arg1 of a *Source* relation, as it is more important to know about the complaint itself than to know where the information came from, and 4a is subordinated under 4b (cf. Hunter et al., 2007).

Table 1 contains all discourse relations. Numbers in square brackets represent the distribution of the overall class. Numbers in parentheses represent the distribution of the single relation.

In the table, coordinating relations are marked with a small 'c' in front of the relation and subordinating relations are marked with a small 's'.

3. An experiment on inter-annotator and inter-adjudicator agreement

For any annotation scheme that ventures into the domain of semantic and/or pragmatic distinctions, reliability is an issue that needs to be addressed explicitly in order to maintain the predictability of the annotated data (or, equivalently, the predictive power of conclusions from that data).

Regarding the agreement on discourse relations, Marcu et al. (1999) determined κ values between $\kappa=0.54$ (Brown corpus) and $\kappa=0.62$ (MUC) for fine-grained RST relations and between $\kappa=0.59$ (Brown) and $\kappa=0.66$ (MUC) for coarser-grained relations. In their reliability study with the Penn Discourse Treebank, Prasad et al. (2008) determined agreement values between 80% (finest level) and 94% (coarsest level with 4 relation types), but did not report any chance-corrected values. Al-Saif and Markert (2010) report values of $\kappa=0.57$ for their PDTB-inspired connective scheme, saying that most disagreements are due to highly ambiguous connectives such as *w/and*, which can receive one of several relations. In a study on their Dutch RST corpus, van der Vlieth et al. (2011) found an inter-annotator agreement of $\kappa=0.57$. To the best of our knowledge, no agreement figures have been published on the RST-based Potsdam Commentary Corpus (Stede, 2004) or any other German corpus with discourse relation annotation.

In the regular annotation process of our corpus, two annotators create EDU segmentation, topic segments,

and discourse relations independently from each other; in a second step, the results from both annotators are compared and a coherent gold-standard annotation is created after discussing the goodness-of-fit of respective partial analyses to the text and the applicability of linguistic tests. In order to account for the complete annotation process including the revision step, we follow Burchardt et al. (2006) and separately report *inter-annotator* agreement, which is determined after the initial annotation, and *inter-adjudicator* agreement, which is determined after an additional adjudication step. The adjudication step is carried out by two adjudicators based on the original set of annotations, but is performed by each adjudicator independently from the other.

In the case where multiple relations were annotated between the same EDU ranges (for example, a temporal *Narration* relation in addition to a *Result-Cause* relation from the Contingency group), we counted the annotations as matching whenever the complete set of relations (i.e. $\{Narration, Result-Cause\}$ in the example) is the same across annotators.

In a sample of three documents that we used for our agreement study, we found that annotators agreed on 49 relations spans, with the comparison yielding an agreement value of $\kappa=0.55$ for individual relations, and $\kappa=0.65$ for the middle level of the taxonomy (eight relation types).

For the inter-adjudicator task, we found an agreement on 82 relation spans, among which relation agreement was at $\kappa=0.83$ for individual relations, and $\kappa=0.85$ for the middle level of the taxonomy, or a reduction of disagreements of about 57%.

4. Discussion and Conclusion

In this article, we have presented the annotation scheme we use to annotate discourse relations of complete texts in a subset of the TüBa-D/Z corpus, and reported the results of an agreement study using these guidelines and relation inventory. While the raw inter-annotator agreement is on a similar level as other annotation efforts with a similar scope, we found that a subsequent adjudication step introduces a rather substantial reduction in disagreements (between adjudicated versions that were obtained independently of each

other), which suggests that a large part of the (raw) disagreement is due to the sheer complexity of the task and should not be taken as indicating the infeasibility of discourse structure (and discourse relation) annotation in general.

The public availability of a corpus with discourse relation annotation in combination with the syntactic and referential annotation from the main TüBa-D/Z corpus will also allow it to provide an empirical evaluation of theories concerning the interface between syntax and discourse, such as D-LTAG (Webber, 2004) or D-STAG (Danlos, 2009) as well as those that predict interactions between referential and discourse structure (Grosz & Sidner 1986; Cristea et al., 1998; Webber, 1991; Chiarcos & Krasavina, 2005, inter alia).

5. References

- Al-Saif, A., Markert, K. (2010): Annotating discourse connectives for Arabic. In Proc. LREC 2010.
- Asher (1993): Reference to Abstract Objects in Discourse. Kluwer, Dordrecht.
- Asher, N., Lascarides, A. (2003): Logics of Conversation. Cambridge University Press, Cambridge.
- Asher, N., Vieu, L. (2005): Subordinating and coordinating discourse relations. *Lingua* 115, 591-610.
- Bras, M., Le Draoulec, A., Asher, N. (2006): Evidence for a Scalar Analysis of Result in SDRT from a Study of the French Temporal Connective 'alors'. In: SPRIK Conference "Explicit and Implicit Information in Text - Information Structure across Languages".
- Burchardt, A., Erk, K., Frank, A., Kowalski, A., Padó, S., Pinkal, M. (2006): The SALSA Corpus: a German Corpus Resource for Lexical Semantics. In Proceedings of LREC 2006.
- Büring, D. (2003): On D-Trees, Beans, and B-Accents. *Linguistics and Philosophy* 26(5), pp. 511-545.
- Carlson, L., Marcu, D. (2001): Discourse Tagging Manual. ISI Tech Report ISI-TR-545.
- Carlson, L., Marcu, D., Okurowski, M. E. (2003): Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In: *Current Directions in Discourse and Dialogue*, Kluwer.
- Chiarcos, C., Krasavina, O. (2005): Rhetorical Distance

- Revisited: A Parametrized Approach. In *Workshop on Constraints in Discourse (CID 2005)*.
- Cristea, D., Ide, N., Romary, L. (1998): *Veins Theory: A Model of Global Discourse Cohesion and Coherence*. In *Proc. CoLing 1998*.
- Danlos L. (2009): *D-STAG : Un formalisme d'analyse automatique de discours basé sur les TAG synchrones*. *Revue TAL* 50 (1), pp. 111-143.
- Grosz, B., Sidner, C. (1986): *Attention, Intentions, and the structure of discourse*. *Computational Linguistics* 12(3), pp. 175-204.
- Hearst, M. (1997): *TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages*, *Computational Linguistics*, 23 (1), pp. 33-64.
- Hobbs, J. (1985): *On the Coherence and Structure of Discourse*, Report No. CSLI-85-37, Center for the Study of Language and Information, Stanford University.
- Hunter, J., Baldridge, J., N. Asher (2007): *Annotation for and Robust Parsing of Discourse Structure on Unrestricted Texts*. *Zeitschrift für Sprachwissenschaft* 26, pp. 213-239.
- Knott, A., Oberlander, J., O'Donnell, M., Mellish, C. (2001): *Beyond Elaboration: The interaction of relations and focus in coherent text*. In: Sanders, Schilperoord, Spooren (eds.), *Text representation: linguistic and psycholinguistic aspects*. John Benjamins.
- Mann, W. C., Thompson, S. A. (1998): *Rhetorical Structure Theory: Toward a functional theory of text organization*. *Text* 8, pp. 243-281.
- Marcu, D., Amorrortu, E., Romera, M. (1999): *Experiments in Constructing a Corpus of Discourse Trees*. *ACL Workshop on Standards and Tools for Discourse Tagging*.
- Polanyi, L., Scha. R. (1983): *On the Recursive Structure of Discourse*. In K. Ehlich & H. Van Riemsdijk (Eds.), *Connectedness in sentence, discourse and text*, pp. 141-178. Tilburg: Tilburg University
- Prasad, R., Miltsakaki, M., Dinesh, N., Lee, A., Joshi, A., Robaldo, L., Webber, B. (2007): *The Penn Discourse Treebank 2.0 Annotation Manual*. Technical Report, University of Pennsylvania.
- Reese, B., Denis, P., Asher, N., Baldridge, J., Hunter, J. (2007): *Reference Manual for the Analysis and Annotation of Rhetorical Structure*. Technical Report, University of Texas at Austin.
- Roberts, C. (1996): *Information Structure in Discourse: Towards an Integrated Formal Theory of Pragmatics*. In Yoon, Kathol (eds.), *OSU Workin Papers in Linguistics* 49: *Papers in Semantics*, pp. 91-136.
- Sanders, T. J. M., Spooren, W. P. M., Noordman, L. G. M. (1992): *Toward a Taxonomy of Coherence Relations*. *Discourse Processes* 15, pp. 1-35.
- Sanders, T. J. M., Spooren, W. P. M. (1999): *Communicative intentions and coherence relations*. In Bublitz, Lenk, Ventola (eds.) *Coherence in Text and Discourse*, pp. 235-250. John Benjamins, Amsterdam.
- Schilder, F. (2002): *Robust discourse parsing via discourse markers, topicality and position*. *Natural Language Engineering* 8(2), pp. 235-255.
- Somasundaran, S., Namata, G., Wiebe, J., Getoor, L. (2009): *Supervised and Unsupervised Methods in Employing Discourse Relations for Improving Opinion Polarity Classification*. In *Proc. EMNLP 2009*.
- Stede, M. (2004): *The Potsdam Commentary Corpus*. In *Proc. ACL Workshop on Discourse Annotation*.
- Telljohann, H., Hinrichs, E. W., Kübler, S., Zinsmeister, H., Beck, K. (2009): *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*. Technical Report, Seminar für Sprachwissenschaft, Universität Tübingen.
- Txurruka, I. G. (2003): *The Natural Language Conjunction*. *And. Linguistics and Philosophy* 26(3), pp. 255-285.
- van der Vlieth, N., Berzlanovich, I., Bouma G., Egg, M., Redeker, G. (2011): *Building a Discourse-Annotated Dutch Text Corpus*. In *Proceedings of the DGfS Workshop "Beyond Semantics"*, Bochumer Linguistische Arbeitsberichte 3.
- van Kuppevelt, J. (1995): *Discourse Structure, Topicality and Questioning*. *Linguistics* 31, pp. 109-147.
- Webber, B. (1991): *Structure and Ostension in the Interpretation of Discourse Deixis*. *Natural Language and Cognitive Processes* 6(2), pp. 107-135.
- Webber, B. (2004): *DLTAG: Extending Lexicalized TAG to Discourse*. *Cognitive Science* 28, pp. 751-779.

Devil's Advocate on Metadata in Science

Christina Hoppermann, Thorsten Trippel, Claus Zinn

General and Computational Linguistics, University of Tübingen

Wilhelmstraße 19, D-72074 Tübingen

E-mail: christina.hoppermann@uni-tuebingen.de, thorsten.trippel@uni-tuebingen.de, claus.zinn@uni-tuebingen.de

Abstract

This paper uses a devil's advocate position to highlight the benefits of metadata creation for linguistic resources. It provides an overview of the required metadata infrastructure and shows that this infrastructure is in the meantime developed by various projects and hence can be deployed by those working with linguistic resources and archiving. Possible caveats of metadata creation are mentioned starting with user requirements and backgrounds, contribution to academic merits of researchers and standardisation. These are answered with existing technologies and procedures, referring to the Component Metadata Infrastructure (CMDI). CMDI provides an infrastructure and methods for adapting metadata to the requirements of specific classes of resources, using central registries for data categories, and metadata schemas. These registries allow for the definition of metadata schemas per resource type while reusing groups of data categories also used by other schemas. In summary, rules of best practice for the creation of metadata are given.

Keywords: metadata, Component Metadata Infrastructure (CMDI), infrastructure, sustainable archives

1. Introduction

The creation of primary research data and its analysis is a large share of a researcher's workload. In linguistics, research data comprises many different types: there are resources such as corpora, lexicons, and grammars; there are various kinds of experimental data resulting, for example, from perception and production studies with sensor data originating from eye-tracking and MRI (magnetic resonance imaging) devices. There is data in the form of speech recordings, written text, videotaped gestures, which, in part, is annotated or transcribed along many different layers; there is audio and video data of other forms of human-human communication such as cultural or religious songs or dances; and there is also a large variety of software tools for the manipulation, analysis and interpretation of all these types of data sources.

Once a study of research data yields statistically and scientifically significant results, it is documented and published, usually complementing a description of research methodology, interpretations of results, etc., with a depiction of the underlying research data. Reputable journals are archived so that its articles are deemed accessible for a long time. Access to articles is usually facilitated via Dublin Core (DC) metadata

categories such as "author", "title", "journal", "publisher" or "publication year". In general, however, there is no infrastructure in place to access the research data underlying a reported study, although some researchers make such data available via their webpage or institution, and some conferences or journals ask authors to supplement their article with primary data, which is then also made public.¹ So far, it is not the general rule to describe research data with metadata for indexing or cataloguing by themselves or others. In part, this is due to caveats for the provision of metadata held by large parts of the scientific community. In this paper, the Devil's Advocate (DA) will articulate some of these caveats. We will aim at rebutting each of them, given the recent advances for metadata management, in particular, in the area of linguistics.

2. Playing Devil's Advocate

DA: There is little if any scientific merit to be gained from resource and metadata provision.

This is a view mentioned in a recent statement by the Wissenschaftsrat² which says that infrastructure does

¹ For example, Interspeech 2011 invited authors to submit supporting data files to be included on the Proceedings CD-ROM in case of paper acceptance.

² The German Wissenschaftsrat is a joined council of German

hardly provide for an increased scholarly reputation (Wissenschaftsrat, 2011:23). Though this might be true for a restricted notion of scientific merit, that is the merit being defined by the number of published journal articles and books, it is not true in a less restricted sense. Furthermore, the Wissenschaftsrat (Wissenschaftsrat, 2011:23) points out that infrastructural projects offer the opportunity for methodical innovations, generate new research questions, and help attracting new researchers. If new researchers, methods and research questions are part of the scientific merit, the claim that there is no scientific merit in metadata provision is thus not true. There are even more reasons for arguing that additional scientific merits are gained, at least in three overlapping areas: (1) by providing a complete overview over the field, (2) by fostering interoperability and providing reproducible, non-arbitrary results, and (3) by increasing the pace of gaining research results.

First of all, in an ideal case, a metadata-driven resource inventory gives an accurate picture of a scientific landscape by containing all resource types such as corpora, lexical databases, or experiments. By having access to all these resources, in principle, nothing is gained because it is too time-consuming to analyse and reproduce research questions from the data. But as soon as resources are described by metadata, it is possible to classify, sort and provide an overview over them using the descriptions as such. Though descriptions contain generalisations, they are still sufficient to provide an outline of resources. This also serves the purpose of providing essential background for steering research activities and funding projects as well as to discover trends and gaps, all allowing to increase the researcher's reputation and merit.

Second, the metadata-based publicity fosters communication between researchers, for example, because contact information are required to gain access to resources, comparable data structures are needed to be reusable by other methods, or because selections of resources (e.g. subcorpora) have to be created. Resources can be merged and cross-evaluated to discover which results are reproducible. This helps to avoid fraud and plagiarism. At the same time, the investigation of research questions different from the original ones can be

Research Foundation officials and researchers appointed by the government for consulting it on research related issues.

applied to existing resources. In all cases, good scientific practice will credit the resource creator, and thus add to his or her reputation when a publication makes reference to its underlying research data, which is possible on the basis of appropriate metadata. The references pointing to the resources can be indexed by others and are consequently added to the scientific map.

Third, more and faster results can be created. By providing metadata, researchers new to a discipline gain a faster overview over the research questions and activities of a discipline as well as easier access to existing linguistic resources and tools. Moreover, accurate metadata descriptions can help avoiding the duplication of research work by providing insights and access to existing work. Hence, researchers who are applying new methods do not always have to recreate resources but can rely on existing ones, providing a jumpstart. At the same time, the resources as such are providing added benefit by being more widely used, thereby also increasing the reputation of the creator.

DA: Expert knowledge on metadata is required to properly describe research data. Thus metadata experts rather than researchers are called for duty.

The library sciences, with their long tradition and expertise in metadata, have many different classification systems in place to organise collections. But is it realistic to ask researchers, such as linguists, to properly describe language resources and tools with metadata, given their lack of knowledge in metadata provision, the variety and complexity of research data, and the missing dominant metadata schemes in the field? On the other hand, it seems clear that metadata provision cannot be done properly without the researchers' involvement. It is unrealistic to assume that some research data can be just given to a librarian with expertise in linguistics (or a linguist with expertise in archiving methodology) with the task to assign proper metadata to it. There needs to be considerable involvement of the resource creator in describing the resource in formal (where possible) and informal terms (possibly by filling out a questionnaire). The "librarian" can then enter the provided information into a formal schema, ensuring that, at least, obligatory descriptors are properly provided. In sum, to put a proper metadata-based infrastructure in place, some minimal researcher training in metadata provision is needed. This

needs to be complemented with infrastructure personnel, or, if possible, with user-friendly metadata editors that trained researchers can learn to use.

DA: There is a little if any consensus on the set of metadata descriptors or metadata schemes to be used in describing language resources and tools.

It is clear that a common vocabulary for metadata provision is required. Otherwise it will be hard to offer effective metadata-based search and retrieval services. It is also evident that established metadata standards such as Dublin Core are insufficient, as they do not include every data category (DatCat) needed for describing specific types of resources. However, given the complexity of the research field in linguistics with its many different resource types, it is naïve to assume that established metadata schemas can be reused without losing descriptive power. For example, resource types need to be indicated and for different resource types additional descriptive categories need to be defined. For lexical resources it is common to describe the lexical structures, for annotations the annotation tag sets, for experiments the size of the samples and the free and bound variables. Each of these data categories is only relevant for the individual type of a resource, but for these they can be more essential than categories such as “title” and “author”. As this list of data categories may require additions, since new resource types become available, it needs to be treated as an open list.

In recent times, some consensus on the procedure of creating elementary field names for the description of linguistic research data has been achieved in order to allow for a standardisation of data categories. It is formally captured by the ISOcat data category registry for the description of language resources and tools (ISO 12620; International Organization of Standardization, 2009; <http://www.isocat.org>). ISOcat (Figure 1) is an open web-based registry of data categories into which everybody can insert his own data categories with (human-readable) definitions of their intended use. This is done in a private space with limited access that can be used by researchers to include new data categories not yet intended or not ready for standardisation. For private use, these data categories can already be referenced via persistent identifiers (PIDs) but they can also be stored in a public space with unrestricted access and be proposed

as standard data categories. If the data categories are submitted for standardisation, a standardisation process involving domain experts is being initiated with community consensus building, quality assurance, voting and maintenance cycles.

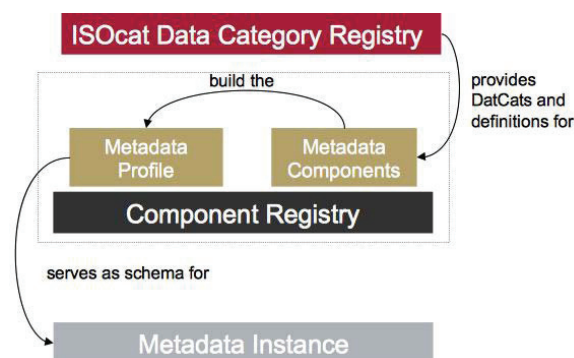


Figure 1: Relation between ISOcat, Component Registry and metadata instances

The registry provides a solid base to start from, but the sheer size of available DatCats may overwhelm untrained users. Additional structures are needed to minimise cases where different users may apply different descriptors to provide similar resources with metadata. For this purpose, the Component Registry for metadata (Figure 1; Broeder et al., 2010; <http://catalog.clarin.eu/ds/ComponentRegistry/#>) contains a rich set of prefabricated metadata building blocks that aggregate elementary blocks of data categories into larger compounds. Researchers can select and combine existing building blocks – or define new ones – in a schema, which can then be instantiated to describe a given resource with the help of a so-called metadata instance (Figure 1). The concept of reusing building blocks is part of the Component Metadata Infrastructure (CMDI, <http://www.clarin.eu/cmdi>). For many resource types the registry already contains prefabricated schemas that can be re-used by researchers. Moreover, there exists at least one fully functional metadata editor (<http://www.lat-mpi.eu/tools/arbil/>) with interfaces to both ISOcat and the Component Registry. It is freely available and support is provided by the programmers to facilitate the use of the editor for non-expert users who otherwise might be overwhelmed by the total range of functions the editor offers. There are also other XML editors supporting the schemas. Once a schema is defined with these tools, these off-the-shelf

XML editors are available to describe resources with metadata according to the metadata schema. These schemas can then be used to validate the metadata instances with the help of syntactical parsers to ensure the adherence to syntactic structures and controlled vocabulary.

DA: There is rarely a right time to make a resource public (via metadata description).

Research rarely follows a fully planned path. A resource such as a corpus or a lexicon is adjusted, additional layers of annotation or transcription are added, data may get re-annotated with different coders, lexical entries may get revised or extended to reflect new insights, etc. Nevertheless, the moment publications are created and project reports are written, it shall be good scientific practice to archive the underlying research data and to assign and publish metadata about the resource. Here, the current status of the resource can be marked with metadata about, for instance, the resource's life cycle or versioning information.

There is also a policy change in the funding agencies. The German Research Association (DFG), for instance, sets the terms that resources ought to be maintained by the originating institution; researchers are responsible for the proper documentation of resources, and procedures need to be defined for the case when they leave an institution (Deutsche Forschungsgemeinschaft, 1998:13). A proper documentation of resources has to include their description in terms of metadata to facilitate their archival and future retrieval.

Therefore, at the latest, metadata shall be provided (or revised) at the end of a research project, at best by the researchers who have created the resource. Ideally, the life cycle stage at archiving time is already defined in the project work plan. Even if the desired final state was not accomplished, the primary data needs to be archived by the end of the project with proper metadata assigned to it.

DA: Without a central metadata agency, all the added values advertised will not materialise.

Added values such as searchability and citation of resources require some point of access to the metadata. It is correct that there is not a single central metadata agency but there are various interconnected agencies providing services to the community in terms of metadata.

For instance, the German NaLiDa project (<http://www.sfs.uni-tuebingen.de/nalida/>) serves as a metadata centre for resources and tools created in Germany. The project as such does not claim exclusive representation, but aims at cooperating with other archives in providing a service to the community for accessing metadata in the form of catalogues and allowing easy access to resources. It harvests metadata from participating institutions and also provides metadata management support for German research institutions (Barkey et al., 2011). Within the project, a faceted search interface was developed with complementation of a full-text search engine (<http://www.sfs.uni-tuebingen.de/nalida/katalog>), with currently access to more than 10,000 metadata records of language resources and tools. Though the NaLiDa project could be seen as a central metadata agency, its implementation has a rather decentralised flavour. Metadata is harvested from various sources and then aggregated and indexed into a single database. To kick-start or increase the inflow of data, participating institutions receive help both in terms of setting-up an OAI-PMH³-based data provision service and in other aspects of metadata creation and maintenance. Once the local metadata providers – the primary research data remains with the institutions – are set up, other parties than NaLiDa are free to crawl their data sets and to provide services in terms of all data.

At the European level, the CLARIN project (<http://www.clarin.eu>) has also devised such a crawler, and is likewise offering a faceted search interface for language resources and tools (CLARIN Virtual Language Observatory, <http://www.clarin.eu/vlo/>). Since both (and other) parties work towards the realisation of a common infrastructure, with different foci but similar goals, there is much to be gained from a healthy competition and exchange of ideas for the scientific community to profit from.

3. Summary

Given the recent advances in linguistics with regard to metadata provision for linguistic resources and tools, there is little left to offer excuses for not using the existing infrastructure. In general, this results in the

³ Open Archives Initiative Protocol for Metadata Harvesting

following rules of best practice for the documentation of resources:

- 1) One of the best strategies for preserving research data is by publishing it into repositories and networks. This way, multiple archives serve as backup. Additionally, it allows for an easier sharing and spreading of resources, contributing to the academic merits of resource providers.
- 2) Archived data is easier accessible if the data is sufficiently described. As flexible metadata schemas can adapt for various types of resources, it is possible to create such descriptions as required by the type of a resource. Metadata can then be used to make resources public, in order for others to use (harvest) them.
- 3) Data categories are best defined in central (standardised) registries, such as ISOcat, that allow for references via persistent identifiers. No data categories should be used that are not centrally defined to avoid fragmentation of the resource community.
- 4) For interoperability purposes, components as collections of data categories should be reused where adequate or defined as new entries in the Component Registry for reuse by others.
- 5) The flexibility of the framework helps to avoid tag abuse if data providers adhere to data category definitions or, if not available, define their own modified categories. This will contribute to the consistency and reusability of data.
- 6) Syntactic evaluation of metadata should always be performed to ensure harvesting, usability of applications and consistency. By checking for content models, tag abuse can be avoided further.
- 7) When using research data, it should be referred to them as stated in the data's metadata.
- 8) Resource creators might need some training and assistance, which is provided by various projects. Some time for this work should be included.

4. Acknowledgements

Work on this paper was conducted within the *Centre for Sustainability of Linguistic Data (Zentrum für Nachhaltigkeit Linguistischer Daten, NaLiDa)*, which is funded by the German Research Foundation (DFG) in the Scientific Library Services and Information Systems

(LIS) framework, and within the infrastructure project *Heterogeneous Primary Research Data: Representation and Processing* of the Collaborative Research Centre *The Construction of Meaning: the Dynamics and Adaptivity of Linguistic Structures* (SFB 833), which is also funded by the DFG.

5. References

- Barkey, R., Hinrichs, E., Hoppermann, C. Trippel, T., Zinn, C. (2011): Komponenten-basierte Metadatenschemata und Facetten-basierte Suche - Ein flexibler und universeller Ansatz. In J. Griesbaum, T. Mandl & C. Womser-Hacker (eds.), *Information und Wissen: global, sozial und frei? Internationales Symposium der Informationswissenschaft (Hildesheim)*. Boizenburg: Verlag Werner Hülsbusch (vwh), pp. 62-73.
- Broeder, D., Kemps-Snijders, M., Van Uytvanck, D., Windhouwer, M., Withers, P., Wittenburg, P., Zinn, C. (2010): A Data Category Registry- and Component-based Metadata Framework. In *Proceedings of the 7th Conference on International Language Resources and Evaluation, 19-21 May 2010*, European Language Resources Association.
- Deutsche Forschungsgemeinschaft (1998): *Vorschläge zur Sicherung guter wissenschaftlicher Praxis: Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft“*, Denkschrift. Weinheim: Wiley-VCH. See http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf (retrieved March 31, 2011).
- International Organization of Standardization (2009): *Terminology and other language and content resources - Specification of data categories and management of a Data Category Registry for language resources (ISO-12620-2009)*, Geneva. Go to www.isocat.org to access the registry.
- Wissenschaftsrat (2011): *Empfehlung zu Forschungsinfrastrukturen in den Geistes- und Sozialwissenschaften*. Berlin: 28/01/2011. See <http://www.wissenschaftsrat.de/download/archiv/10465-11.pdf> (retrieved March 31, 2011).

Improving an Existing RBMT System by Stochastic Analysis

Christian Federmann, Sabine Hunsicker

DFKI – Language Technology Lab

Stuhlsatzenhausweg 3, D-66123 Saarbrücken, GERMANY

E-mail: {cfedermann,sabine.hunsicker}@dfki.de

Abstract

In this paper we describe how an existing, rule-based machine translation (RBMT) system that follows a transfer-based translation approach can be improved by integrating stochastic knowledge into its analysis phase. First, we investigate how often the rule-based system selects the wrong analysis tree to determine the potential benefit from an improved selection method. Afterwards we describe an extended architecture that allows integrating an external stochastic parser into the analysis phase of the RBMT system. We report on the results of both automatic metrics and human evaluation and also give some examples that show the improvements that can be obtained by such a hybrid machine translation setup. While the work reported on in this paper is a dedicated extension of a specific rule-based machine translation system, the overall approach can be used with any transfer-based RBMT system. The addition of stochastic knowledge to an existing rule-based machine translation system represents an example of a successful, hybrid combination of different MT paradigms into a joint system.

Keywords: Machine Translation, Hybrid Machine Translation, Stochastic Parsing, System Combination

1. Introduction

Rule-based machine translation (RBMT) systems that employ a transfer-based translation approach, highly depend on the quality of their analysis phase as it provides the basis for its later processing phases, namely transfer and generation. Any parse failures encountered in the initial analysis phase will proliferate and cause further errors in the following phases. Very often, bad translation results can be traced back to incorrect analysis trees that have been computed for the respective input sentences. Consequently, any improvements that can be achieved for the analysis phase of some RBMT system lead to improved translation output, which makes this an interesting topic in the context of hybrid machine translation.

In this paper we describe how a stochastic parser can supplement the rule-based analysis phase of a commercial RBMT system. The system in question is the rule-based engine Lucy LT. This engine uses a sophisticated RBMT transfer approach with a long research history, as explained in detail in (Wolf et al., 2010). The output of its analysis phase is a forest containing a small number of tree structures. For this study we investigated if the existing rule base of the Lucy LT system chooses the best tree from the analysis forest

and how the selection of this best tree out of the set of candidates can be improved by adding stochastic knowledge to the RBMT system.

The paper is structured in the following way: in Section 2 we describe the Lucy RBMT system and its transfer-based architecture. Afterwards, in Section 3, we provide details on the integration of a stochastic parser into the Lucy analysis phase of this rule-based system. Section 4 describes the experiments we performed and reports the results of both automated metrics and human evaluation efforts before Section 5 discusses some examples that show how the proposed approach has improved or degraded machine translation quality. Finally, in Section 6, we conclude and provide an outlook on future work in this area.

2. Lucy System Architecture

The Lucy LT engine is a renowned RBMT system that follows a classical, transfer-based translation approach. The system first analyses the given source sentence resulting in a forest of several analysis trees. One of these trees is then selected (as “best” analysis) and transformed in the transfer phase into a tree structure from which the target text can be generated.

It is clear that any errors that occur during the initial

analysis phase proliferate and cause negative side effects on the quality of the resulting translation. As the analysis phase is of special importance, we describe it in more detail. The Lucy LT analysis consists of several phases:

- 1) The input is tokenised with regards to the source language lexicon.
- 2) The resulting tokens then undergo a morphological analysis, which identifies possible combinations of allomorphs for a token.
- 3) This leads to a chart which forms the basis for the actual parsing, using a head-driven strategy. Special treatment is performed for the analysis of multi-word expressions and also for verbal framing.

At the end of the analysis, there is an extra phase named phrasal analysis that is called whenever the grammar was not able to construct a legal constituent from all the elements of the input. This happens in several different scenarios:

- The input is ungrammatical according to the LT analysis grammar.
- The category of the derived constituent is not one of the allowed categories.
- A grammatical phenomenon in the source sentence is not covered.
- There are missing lexical entries for the input sentence.

During the phrasal analysis, the LT engine collects all partial trees and greedily constructs an overall interpretation of the chart. Based on our findings from experiments with the Lucy LT engine, phrasal analyses are performed for more than 40% of the sentences from our test sets and very often result in bad translations.

Each resulting analysis tree, independent of whether it is a grammatical or phrasal analysis, is also assigned an integer score by the grammar. The tree with the highest score is then handed over to the transfer phase, thus pre-defining the final translation output.

3. Adding Stochastic Analysis

An initial, manual evaluation of the translation quality based on the tree selection of the analysis phase showed that there is potential for improvement. For this, we changed the RBMT system to produce translations for all its analysis trees and ranked them according to their quality. In many cases, one of the alternative trees would have lead to a better translation.

Next to the assigned score, we examined the significance of two other features:

- 1) The size of the analysis trees themselves, and
- 2) The tree edit distance of each analysis candidate to a stochastic parse tree.

An advantage of stochastic parsing lies in the fact that parsers from this class can deal very well even with ungrammatical or unknown output, which we have seen is problematic for a rule-based parser. We decided to make use of the Stanford Parser as described in (Klein & Manning, 2003), which uses an unlexicalised, probabilistic context-free grammar trained on the Penn Treebank. We parse the original source sentence with this PCFG grammar to get a stochastic parse tree that can be compared to the trees from the Lucy analysis forest.

In our experiments, we compare the stochastic parse tree with the alternatives given by Lucy LT. Tree comparison is implemented based on the Tree Edit Distance, as originally defined in (Zhang & Shasha, 1989}. In analogy to the Word Edit or Levenshtein Distance, the distance between two trees is the number of editing actions that are required to transform the first tree into the second tree. The Tree Edit Distance knows three actions:

- Insertion
- Deletion
- Renaming (substitution in Levenshtein Distance)

We use a normalised version of the Tree Edit Distance to estimate the quality of the trees from the Lucy analysis forest. The integration of the stochastic selection has been possible by using an adapted version of the rule-based system, which allowed performing the selection of the analysis tree from an external process.

4. Experiments

Two test sets were used in our experiments. The first test set was taken from the WMT shared task 2008, consisting of a section of data from Europarl (Koehn, 2005). The second test set, which was taken from the WMT shared task 2010 contained news text. Phrasal analyses caused by unknown lexical items occurred more often in the news text, as that text sort tends to more often use colloquial expressions. In our experiments, we translated from English→German; evaluation was performed using both automated metrics and human evaluation using an annotation tool similar to e.g. Appraise (Federmann, 2010).

First, only the Tree Edit Distance and internal score from the Lucy analysis phase were used and we select the tree with the lowest edit distance. If the lowest distance holds for two or more trees, the tree with the highest LT internal score is chosen. Later we added the size of the candidate trees as an additional feature, with a bias to prefer larger trees as they proved to create better translations in our experiments. Results from automatic scoring using BLEU (Papineni et al., 2001) and the derived NIST score are reported in Table 1 and Table 2 for test set #1 and test set #2, respectively. The BLEU scores for the new translations are a little bit worse, but still comparable to the quality of the original translations. The difference is not statistically significant.

Test set #1	BLEU	NIST
Baseline	0.1100	4.4059
Stochastic Selection	0.1096	4.3946

Table 1: Automatic scores for test set #1.

Test set #2	BLEU	NIST
Baseline	0.1529	5.5725
Stochastic Selection	0.1514	5.5469
Selection+Size	0.1511	5.5341

Table 2: Automatic scores for test set #2.

We also manually evaluated a sample of 100 sentences. For this, we created all possible translations for each phrasal analysis and had human annotators judge on their quality. Then, we checked whether our stochastic selection mechanism returned a tree that led to the best translation. In case it did not, we investigated the reasons for this. Sentences for which all trees created the same translation were skipped.

Table 3 shows the error rate of our stochastic analysis component that chose the optimal tree for 56% of the sentences, while Table 4 shows the selection reasons that resulted in the selection of a non-optimal tree. We also see that the minimal tree edit distance seems to be a good feature to use for comparisons, as it holds for 71% of the trees, including those examples where the best tree was not scored highest by the LT engine. This also means that additional features for choosing the tree out of the group of trees with the minimal edit distance are required.

Best translation?	Yes (56%)	No (44%)
Minimal distance?	Yes (71%)	No (29%)

Table 3: Error rate of the stochastic analysis.

More than 50 tokens in source	36.4%
Time-out before best tree is reached	29.5%
Chosen tree had minimal distance	34.1%

Table 4: Reasons for erroneous tree selection.

Even for the 29% of sentences, in which the optimal tree was not chosen, little quality was lost: in 75.86% of those cases, the translations didn't change at all (obviously the trees resulted in equal translation output). In the remaining cases the translations were divided evenly between slight degradations and equal quality.

In cases when the best tree was not chosen, the first tree (which is the default tree) was selected in 70.45%. This is due to a combination of robustness factors that are implemented in the RBMT system and have been beyond our control in the experiments. The LT engine has several different indicators that may each throw a time-out exception, if, for example, the analysis phase takes too long to produce a result. To avoid getting time-out errors, only sentences with up to 50 tokens are treated by our stochastic selection mechanism. Additionally, the component itself checks the processing time and returns intermediate results, if this limit is reached. We are currently working on eliminating this time-out issue as it prevents us from driving our approach to its full potential. As with the internal score, we see that the Tree Edit Distance on its own is a good indicator of the quality of the analysis, but that additional features are required to prevent suboptimal decisions to be taken. As such, we included the size of the trees. Here the bigger trees are preferred to smaller ones as experimental results have confirmed that these are more likely to produce better translations.

The manual evaluation shows results that are similar to the automated metrics. We are currently investigating in more detail what happened in case of the degradations to improve that misbehaviour. It seems as if additional features might be needed to more broadly improve the rule-based machine translation engine using our stochastic selection mechanism.

5. Examples

We now provide some examples from our experiments that illustrate how the stochastic selection mechanism changed the translation output of the rule-based system. For example, the analysis of the following sentence is now correct:

Source: “They were also protesting against bad pay conditions and alleged persecution.”

Translation A: “Sie protestierten auch gegen schlechte Soldbedingungen und behaupteten Verfolgung.”

Translation B: “Sie protestierten auch gegen schlechte Soldbedingungen und angebliche Verfolgung.”

Translation A is the default translation. The analysis tree associated with this translation contains a node for the adjective “alleged” which is wrongly parsed as a verb.

The next example shows how an incorrect word order problem is fixed:

Source: “If the finance minister can't find the money elsewhere, the project will have to be aborted and sanctions will be imposed, warns Janota.”

Translation A: “Wenn der Finanzminister das Geld nicht anderswo finden kann, das Projekt abgebrochen werden müssen wird und Sanktionen auferlegt werden werden, warnt Janota.”

Translation B: “Wenn der Finanzminister das Geld nicht anderswo finden kann, wird das Projekt abgebrochen werden müssen und Sanktionen werden auferlegt werden, warnt Janota.”

Lexical items are associated with a domain area in the lexicon of the rule-based system. Items that are contained within a different domain area than the input text are still accessible, but items in the same domain are preferred. In the following example, this leads to an incorrect disambiguation of multi-word expressions:

Source: “Apparently the engine blew up in the rocket's third phase.”

Translation A: “Offenbar blies der Motor hinauf die dritte Phase der Rakete in.”

Translation B: “Offenbar flog der Motor in der dritten Phase der Rakete in die Luft.”

Again, the stochastic selection allows choosing a better tree, which leads to the correct idiomatic translation. Something similar happens in the following case:

Source: “As of January, they should be paid for by the insurance companies and not compulsory.”

Translation A: “Ab Januar sollten sie für von den

Versicherungsgesellschaften und nicht obligatorisch bezahlt werden.”

Translation B: “Ab Januar sollten sie von den Versicherungsgesellschaften und nicht obligatorisch gezahlt werden.”

These changes remain at a rather local scope, but we also have observed instances where the sentence improves globally:

Source: “In his new book, ‘After the Ice’, Alun Anderson, a former editor of New Scientist, offers a clear and chilling account of the science of the Arctic and a gripping glimpse of how the future may turn out there.”

Translation A: “In seinem neuen Buch bietet Alun Anderson, ein früherer Redakteur von Neuem Wissenschaftler, ‘Nach dem Eis’ einen klaren und kalten Bericht über die Wissenschaft der Arktis und einen spannenden Blick davon an, wie die Zukunft sich hinaus dort drehen kann.”

Translation B: “In seinem neuen Buch, ‘Nach dem Eis’, bietet Alun Anderson, ein früherer Redakteur von Neuem Wissenschaftler, einen klaren und kalten Bericht über die Wissenschaft der Arktis und einen spannenden Blick davon an, wie die Zukunft sich hinaus dort drehen kann.”

In translation A, the name of the book, “After the Ice”, has been moved to an entirely different place in the sentence, removing it from its original context.

6. Conclusion and Outlook

The analysis phase proves to be crucial for the quality of the translation in rule-based machine translation systems. In this paper, we have shown that it is possible to improve the analysis results of such a rule-based engine by introducing a better selection method for the trees created by the grammar. Our experiments show that the selection itself is not a trivial task and requires fine-grained selection criteria.

While the work reported on in this paper is a dedicated extension of a specific rule-based machine translation system, the overall approach can be used with any transfer-based RBMT system. Future work will concentrate on the circumvention of e.g. the time-out errors that prevented a better performance of the stochastic selection mechanism. Also, we will more closely investigate the issue of decreased translation

quality and experiment with additional decision factors that may help to alleviate the negative effects.

The addition of stochastic knowledge to an existing rule-based machine translation system represents an example of a successful, hybrid combination of different MT paradigms into a joint system.

7. Acknowledgements

This work was also supported by the EuroMatrixPlus project (IST-231720) that is funded by the European Community under the Seventh Framework Programme for Research and Technological Development.

8. References

- Federmann, C. (2010). Appraise: An open-source toolkit for manual phrase-based evaluation of translations. In Proceedings of the Seventh conference on International Language Resources and Evaluation. European Language Resources Association (ELRA).
- Klein, D., Manning, C. D. (2003). Accurate unlexicalized parsing. In Proceedings of the 41st Annual Meeting of the ACL, pp. 423–430.
- Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In Proceedings of the MT Summit 2005.
- Papineni, K., Roukos, S., Ward, T., Zhu, W.-J. (2001). Bleu: a method for automatic evaluation of machine translation. IBM Research Report RC22176 (W0109-022), IBM.
- Wolf, P., Alonso, J., Bernardi, U., Llorens, A. (2010). EuroMatrixPlus WP2.2: Study of Example- Based Modules for LT Transfer.
- Zhang, K., Shasha, D. (1989). Simple fast algorithms for the editing distance between trees and related problems. SIAM J. Comput., 18, pp. 1245–1262.

Terminology extraction and term variation patterns: a study of French and German data

Marion Weller^a, Helena Blancafort^b, Anita Gojun^a, Ulrich Heid^a

^aInstitut für maschinelle Sprachverarbeitung, Universität Stuttgart

^bSyllabs, Paris

E-mail: {wellermn|gojunaa|heid}@ims.uni-stuttgart.de, blancafort@syllabs.com

Abstract

The terminology of many technical domains, especially new and evolving ones, is not fully fixed and shows considerable variation. The purpose of the work described in this paper is to capture term variation. For term extraction, we apply hand-crafted POS patterns on tagged corpora, and we use rules to relate morphological and syntactic variants. We discuss some French and German variation patterns, and we present first experimental results from our tools. It is not always easy to distinguish (near) synonyms from variants that have a slightly different meaning from the original term; we discuss ways of operating such a distinction. Our tools are based on POS tagging and an approximation of derivation and compounding; however, we also propose a non-symbolic, statistics-based line of development. We discuss general issues of evaluating variant detection and present a small-scale precision evaluation.

Keywords: terminology, term variation, comparable corpora, pattern-based term extraction, compound nouns

1. Introduction

The objective of the EU-funded project TTC¹ (*Terminology Extraction, Translation Tools and Comparable Corpora*) is the extraction of terminology from comparable corpora. The tools under development within the project address the issues of compiling corpus collections, monolingual term extraction and the alignment of terms into pairs of multilingual equivalence candidates, as well as the management and the export of the resulting terminological data towards CAT and MT tools.

Since parallel corpora of specialized domains are scarce and not necessarily available for a broad range of languages (TTC deals with English (EN), Spanish (ES), German (DE), French (FR), Latvian (LV), Russian (RU), Chinese (ZH)), comparable corpora are used instead: textual material from specialized domains is accessible for many languages, either on the Internet or in publications of companies.

In technical domains which are rapidly evolving, documents published on the Internet are often the most recent sources of data. In such domains, terminology typically has not yet been standardized, and thus numerous variants co-exist in published documents. Tools which support the extraction, identification and interrelating of term variants are thus necessary to capture the full range of expressions used in the respective domain. End users may then decide (e.g. on the basis of variant frequency and sources of variants) which expression to prefer.

A second, more technical motivation for term variant extraction is provided by the procedures for term alignment (either lexical or statistical strategies), for which data sparseness is a problem. In order to reduce the complexity of term alignment, TTC intends to gather monolingual variants into sets of related terms. Particularly for this application, we do not only allow for (quasi) synonyms, but also for variants with a slight difference in meaning as shown in 1.

1) *production d'électricité* ↔ *électricité produite*
(*production of electricity* ↔ *produced electricity*)

Terms may be of different forms (single-word vs. multi-word terms) in different languages: this is a challenge

¹<http://www.ttc-project.eu>

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement n. 248005.

for term alignment. For example, compound nouns play an important role in German terminology, but have no equivalents of the same morpho-syntactic structure in many other languages. Grouping equivalent terms of different syntactic structures can help to deal with such cases, as illustrated in 2:

- 2) *Energieproduktion* ↔ *Produktion von Energie* ↔ *production d'électricité*
(*energy production* ↔ *production of energy*)

2. Methodology

The steps required for term extraction and for variant identification follow a simple pipeline architecture: first, a corpus collection is compiled, which then undergoes linguistic pre-processing. Following these steps, monolingual term candidates are extracted. As not all extracted items are domain relevant, we apply statistical filtering. Since we intend to detect term variation on a morpho-syntactic level, this last step requires morphological processing in order to model derivational relationships between word classes.

2.1. Compiling a corpus and pre-processing

To collect corpus data, we use the focused Web crawler *Babouk* (de Groc, 2011) which has been developed within the TTC project. *Babouk* starts with a set of seed terms or URLs given by the user which are combined into queries and submitted to a search engine. *Babouk* scores the relevance of the retrieved web pages using a weighted-lexicon-based thematic filter. Based on the content of relevant retrieved pages, the lexicon is extended and new search queries are combined.

One objective of the TTC project is to rely on flat linguistic analysis that is available for all languages. One strand of research thus goes towards the development of knowledge-poor strategies, such as using a pseudo part-of-speech tagger (Clark, 2003) as a basis for probabilistic NP-extraction (Guégan & Loupy, 2011). A knowledge-rich approach is term extraction based on hand-crafted part-of-speech (POS) patterns, which is the method we chose for the present work.

Pre-processing of our data collection consists of tokenizing, POS-tagging and lemmatization using *TreeTagger* (Schmid, 1994). For efficiency reasons, with German and French being morphologically rich

languages, we work with lemmas rather than inflected forms.

2.2. Term candidate extraction and filtering

Our main focus is on the extraction of nominal phrases such as [NN NN] or [NN PRP NN] constructions (cf. tables 2-5), but [V NN] collocations are also of interest². For each language, we identify term candidates by using hand-crafted POS patterns. In contrast to nominal phrases, which are relatively easy to capture by POS patterns, the identification of [V NN] collocations is more challenging, as verbs and their object nouns do not necessarily occur in adjacent positions, depending on the general structure of the sentence. This applies particularly to German where constituent order is rather flexible and allows for long distances between verbs and their objects.

In order to reduce the extracted term candidates to a set of domain-relevant items, we estimate their domain specificity by comparing them with terms extracted from general language corpora (Ahmad et al, 1992). The underlying idea of this procedure is the assumption that terms which occur in both domain-specific and general language corpora are not domain-relevant, whereas terms occurring only or predominantly in the domain-specific data can be considered as specialized terms. We use the quotient q of a term's relative frequency in the specialized data and in the general language corpus as an indicator for its domain relevance (see table 1).

term candidate	f domain	f general	q
Gleichstrom (<i>direct current</i>)	128	4	22362,7
Jahr (<i>year</i>)	2157	221.213	1,2

Table 1: Domain-specific vs. general language

2.3. Term variation

In TTC we define a *term variant* as “an utterance which is semantically and conceptually related to an original term” (Daille, 2005). Thus, term variants are bound to texts (“utterance”) and require the presence of an “original term” identified e.g. by means of a morpho-syntactic term pattern.

²NN:noun, PRP: preposition, V: verb, VPART: participle

The relationship between term variant and original term is supposed to mainly be one of (quasi-) synonymy or of controlled modification (e.g. by attributive adjectives, NPs or PPs). We formalize this by explicitly classifying relationships between patterns.

We distinguish the following types of variants:

- **graphical** *air flow* ↔ *airflow*
- **morphological** (derivation, compounding)
Energieproduktion ↔ *Produktion von Energie*
(*production of energy*)
solare Energie ↔ *Solarenergie* (*solar energy*)
- **paradigmatic** e.g. omissions
les énergies renouvelables ↔ *les renouvelables*
(*the renewable energies* ↔ *the renewables*)
- **abbreviations, acronyms**
Windenergieanlage ↔ *WEA* (*wind energy plant*)
- **syntactic variants**³ *consommation d'énergie* ↔ *consommation annuelle d'énergie*
(*energy consumption* ↔ *yearly energy consumption*)

Assuming that German technical texts contain many domain-specific compounds, we focus in this work on compound nouns and their variant [NN PRP NN] as illustrated above (morphological variants).

For French, we choose a similar pattern [NN de NN] ↔ [NN VPART]. In our current work, we restrict this pattern to nouns ending in *-tion*. The addition of French morphology tools is planned to widen the scope of these patterns.

2.4. Morphological processing

In order to identify morphological variants of German compounds, we need to split compounds into their components: in the present work, we opt for a statistical compound splitter; the implementation is based on (Koehn & Knight, 2003).

Searching for the most probable split of a given word, the basic idea is that the components of a compound also appear as single words and consequently should occur in corpus data. A word frequency list serves as training data, supplemented with a hand-crafted set of rules to model transitional elements, such as the *s* in *Produktions|kosten* (*production costs*).

³This last type of variants is not necessarily synonymous with the original term.

For French, we created a set of rules to model the relationship between nouns ending in *-tion* and the respective verbs:

- *production* → *produire* (*production* → *produce*)
- *évolution* → *évoluer* (*evolution* → *evolve*)
- *condition* → *conditionner* (*condition* → *condition*)
- *protection* → *protéger* (*protection* → *protect*)

Similar rules can be formulated, e.g. for nouns ending in *-ment* or *-eur*, e.g. *chargement* (nominalized action) → *charger* (verb), as well as *convertisseur* (nominalized tool name) → *convertir* (verb). Similarly, terms containing adjectives ending in *-able*, such as *utilisable* → *utiliser* (cf. table 5) or relational adjectives (*prototypique* → *prototype*) are under study. A further type of pattern that could be added are rules to handle prefixation (e.g. *anti-corrosion* → *corrosion*).

2.5. Processing formally related items

A very common form of graphic variation is hyphenation, e.g. *Luftwärmepumpe* vs. *Luft-Wärmepumpe* (*air-source heat pump*). This type of variation is dealt with by the splitting program, which uses hyphens as splitting points. Hyphenated and non-hyphenated forms are treated as one term.

To a certain extent, our variant detection tools also deal with alternating transitional elements (*Kraftwerkbetrieb* vs. *Kraftwerksbetrieb*). This is modeled by hand-crafted rules which allow for several realizations. Additionally, there are relatively regular forms of spelling variation, e.g. the *new/old orthography* in German, resulting in e.g. *ph/f* variation. This can be dealt with either by rules or using a method based on string-distance.

3. Experiments and examples of results

Our experiments are based on comparable corpora crawled from the Web. While they are generally easy to obtain with a focused crawler, such corpora might be inhomogeneous with respect to domain coverage or types of sources. When working with several languages, the degree of comparability may also vary.

We use a collection of 1000 documents each for French and German, with a total size of 1.55 M tokens (FR) and 1.29 M tokens (DE) of the domain of *wind energy*.

When looking at the extracted German data, we find that

Abgabe von Wärme	1	Wärmeabgabe	18	<i>release of warmth</i>
Beleuchtung von Straße	1	Straßenbeleuchtung	49	<i>street lighting</i>
Erzeugung von Strom	32	Stromerzeugung	569	<i>power generation</i>
Produktion von Strom	4	Stromproduktion	72	<i>power production</i>
Speicherung von Energie	7	Energiespeicherung	37	<i>energy storage</i>
Verbrauch an Primärenergie	1	Primärenergieverbrauch	114	<i>primary energy consumption</i>
Versorgung mit Fernwärme	2	Fernwärmeversorgung	13	<i>district heating</i>
Nutzung von Biomasse	8	Biomassenutzung	7	<i>biomass utilization</i>

Table 2: Prepositional phrases vs. compound nouns

consommation d'électricité	<i>electricity consumption</i>	28	électricité consommée	<i>consumed electricity</i>	15
consommation d'énergie	<i>energy consumption</i>	66	énergie consommée	<i>consumed energy</i>	22
importation de pétrole	<i>import of petroleum</i>	9	pétrole importé	<i>imported petroleum</i>	1
production d'électricité	<i>electricity production</i>	225	électricité produite	<i>produced electricity</i>	95
production de chaleur	<i>heat production</i>	26	chaleur produite	<i>produced heat</i>	21
installation d'éolienne	<i>wind turbine installation</i>	5	éolienne installée	<i>installed wind turbine</i>	16
installation de puissance	<i>installation of power</i>	1	puissance installée	<i>installed power</i>	69
utilisation d'énergie	<i>use of energy</i>	5	énergie utilisée	<i>used energy</i>	19

Table 3: Related French terms: prepositional phrases vs. noun-participle constructions.

Nutzenergie	<i>useful energy</i>	89
nutzbar Energie	<i>usable energy</i>	24
genutzt Energie	<i>used energy</i>	5
nutzbar Energieform	<i>usable energy form</i>	9
genutzt Energieform	<i>used energy form</i>	4
nutzbar Energiegehalt	<i>usable energy content</i>	3
Nutzenergie-Anteil	<i>proportion of useful energy</i>	1
nutzbar Energiemenge	<i>usable amount of energy</i>	1

Table 4: Variants of the compound *Nutzenergie*.

the realization of a term as a compound is often more frequent than the alternative structures [NN PRP NN] or [NN ART_{gen} NN_{gen}], as illustrated in table 2. This does not only apply to common words like *Stromerzeugung* (*power generation*), but also to comparatively long and more complex words like *Fernwärmeversorgung* (lit. *long-distance heat supply: district heating*). We consider this as evidence that the respective compound nouns are established as terms in the domain or even in general language. The degree of preference varies, up to the point of there not being an alternative realization, as is the case with *Windgeschwindigkeit* (*wind speed*, freq=149), for which one could imagine a construction like **Geschwindigkeit des Windes* (*speed of the wind*), which does not occur in our corpus.

In contrast to the German structures, the French terms

énergie utilisée	<i>used energy</i>	19
énergie utile	<i>useful energy</i>	14
énergie utilisable	<i>usable energy</i>	14
forme d'énergie utile	<i>useful energy form</i>	2
form d'énergie utilisable	<i>form of useable energy</i>	2
source d'énergie utilisable	<i>source of usable energy</i>	1

Table 5: Different combinations of the components *energie* and *utile*.

of the pattern pair⁴ [NN de NN] ↔ [NN VPART] in table 3 are not (near) synonyms, but could rather be considered as related. While some terms seem to prefer one of the two patterns, the overall tendency for preference is less clear than for the German examples. The difference in meaning (i.e. action vs. situation) does not allow for full interchangeability of related terms, and the use of the different forms of realization is context dependent. Some terms from the pairs contained in table 3 have different meanings, as is the case with *puissance installée* vs. *installations de puissance élevée* in example (3).

- 3) Par contre, le coût et la complexité des installations les réservent le plus souvent à des installations de puissance élevée pour

⁴Note that the extracted lemma of the participle is its infinitive; we show the inflected form for better readability, i.e. *consommée* instead of *consommer*.

bénéficier d'économies d'échelle.

However, due to the cost and complexity of the installations, they are mostly restricted to installations of high power in order to benefit from the scaling effects.

In other cases, grammatical and/or stylistic constraints may lead authors to use one variant rather than another. For example, compounds in enumerations are rather split in order to facilitate the combination with other nouns, e.g. *Meeresboden* vs. *Boden von Meeren* in example (4).

- 4) Methanhydrat bildet sich am Boden von Meeren bzw. tiefen Seen
Methane hydrate develops at the ground of the sea or deep lakes

In table 4, we show examples of variants in a wider sense: starting with the compound *Nutzenergie* (*useful energy*), we find the synonym *nutzbare Energie* (*usable energy*) and the related form *genutzte Energie* (*used energy*). In the entries in the lower part of the table (grey background), the component *Energie* is part of a compound noun while still preserving the (basic) meaning of the term *Nutzenergie* (*useful energy*).

The French examples in table 5 correspond to the German ones (table 4), with related terms consisting of the basic components in the upper part of the table, and terms expanded by an additional component in the lower part of the table (gray background). The forms *nutzbar* and *utilisable* (*usable*) in table 4 and 5 illustrate one of the above mentioned variation pattern for adjectives.

4. Evaluation and discussion

4.1. Issues in measuring precision and recall

While it is relatively easy to measure the precision of identified (near) synonyms (such as the compound ↔ [NN PRP NN] pairs), it is comparatively difficult to determine the precision of related terms like the ones in tables 4 and 5, as it is often difficult to decide on the degree of relatedness.

Even more difficult is the evaluation of recall, which largely depends on the set of term variation patterns, but also on the patterns used for term candidate extraction.

In order to avoid noise, term candidate extraction is restricted to productive patterns; this implies that not all term variants might be extracted and consequently, that some may not be available for variant grouping. The

same applies to the set of rules used to group variants.

For example, the French pattern [NN PRP NN] is restricted to the prepositions *de* and *à*. While there might be valid terms containing other prepositions, they are excluded from being extracted. Similarly, the large number of potential paraphrases of German compounds cannot be captured.

The examples in tables 4 and 5 illustrate the wide range of possible types of variation and thus the difficulty to capture and relate the different types of variation. In addition to the problem of pattern coverage, another factor is the quality of the morphological tools used to model the relationship between word classes.

4.2. Evaluation of precision

In a small experiment, we measured the precision of the 100 most-frequent German compound nouns and their proposed variants: 74 of the variants are valid. Most of the 26 invalid variants are due to bad PP-attachment, as illustrated by the following example:

- 5) *Stromkunde* (*energy customer*) → **Kunde mit Strom* (*customer with energy*)

which is part of the verbal phrase *Kunden mit Strom versorgen* (*supply costumers with energy*). This kind of error can rather be considered a problem of the extraction step than of the variant detection.

However, in the examined set of 100 items, there was one term-variant pair whose derivation is technically correct, but the meaning is not related:

- 6) *Grundwasser* (*ground water*) → *Wasser am Grund eines Sees* (*water on the ground of a lake*)

4.3. Symbolic vs. non-symbolic approach

By relying on a fixed set of rules for extraction, we clearly favour precision at the cost of recall.

In order to extract terms without a set of patterns, we present a knowledge-poor approach for term extraction using a probabilistic NP extractor and string-level term variation detection. First, we apply a probabilistic NP extractor trained on a small corpus annotated manually with NPs (300 to 600 sentences): this tool has been described in Guégan & Loupy (2011) for the extraction of NP chunks and uses a pseudo part-of-speech tagger (Clark, 2003).

A further non-symbolic procedure consists in relating

extracted terms without relying on a predefined set of variation patterns. We experimented with comparing NPs on a string level (using Levenshtein distance ratio) and grouping terms by similarity. The resulting term groups also provide a basis for the automatic derivation of term variation patterns, which can be used as an input to the symbolic method.

4.4. Relatedness of term candidates

Using a predefined set of term variation patterns facilitates the decision whether terms are (near) synonyms or related. As synonyms, we consider for example the type [compound noun] ↔ [NN PRP NN]. Structures involving relational adjectives ([ADJ NN] (DE), [NN ADJ] (FR)), can be expressed by prepositional phrases, e.g. *production énergétique* ↔ *production d'énergie* (*energy production* ↔ *production of energy*).

Similarly, patterns can also help to specify the degree of relatedness: by explicitly formulating term variation rules we can differentiate between merely related terms (e.g. *consumption* vs. *annual consumption*) and term variants where we assume quasi synonymy (cf. compound nouns in table 2).

A difficult task is the identification of (neoclassical) synonyms: without additional information (e.g. a dictionary), it is impossible to relate terms like *Sonnenenergie* ↔ *Solarenergie* (*solar energy*), as the relation between *Sonne* and *solar* is not known to the system and cannot be derived by morphological means. While the terms in the example above are synonyms, there can be some slight difference in meaning between neoclassical compounds and their native form: the term *hydroélectricité* (*hydroelectricity*) is more precise than *énergie de l'eau* (*water energy*), and not necessarily a synonym.

5. Conclusion and next steps

We presented a method for terminology extraction and for the identification of a certain type of term variation. Preliminary results show that there are preferences for a certain type of realization, especially when considering German compound nouns.

Since our current work only deals with a small part of variation possibilities, we intend to enlarge our

inventory by exploring more variation patterns. We particularly plan to include high-quality morphological tools, e.g. SMOR (Schmid et al., 2004) for German, and DériF (Namer, 2009) for French. SMOR has proven to outperform our statistical splitter.

Another strand of research is the exploration of term variation across languages, e.g. relations between term variants that are similar within different language pairs.

References

- Ahmad, K., Davies, A., Fulford, H., Rogers, M. (1992): What is a Term? The semi-automatic extraction of terms from text. In Translation Studies - an Interdiscipline. John Benjamins Publishing Company.
- Clark, A. (2003): Combining distributional and morphological information for part of speech induction. In Proceedings of the 10th conference of the European chapter of the Association for Computational Linguistics. Budapest, Hungary.
- Daille, B. (2005): Variants and application-oriented terminology engineering. In Terminology, volume. 1.
- Guégan, M., de Loupy, C. (2011): Knowledge-Poor Approach to Shallow Parsing: Contribution of Unsupervised Part-of-Speech Induction. RANLP 2011 - Recent Advances in Natural Language Processing.
- de Groc, C. (2011): Babouk: Focused web crawling for corpus compilation and automatic terminology extraction. In Proceedings of the IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. Lyon, France.
- Koehn, P., Knight, K. (2003): Empirical Methods for Compound Splitting. In Proceedings of the 10th conference of the European chapter of the Association for Computational Linguistics. Budapest, Hungary.
- Namer, F. (2009): Morphologie, Lexique et Traitement Automatique des Langues - Le système DériF. Hermès – Lavoisier Publishers.
- Schmid, H. (1994): Probabilistic part-of-speech tagging using decision trees. In Proceedings of the international conference on new methods in language processing. Manchester, UK.
- Schmid, H., Fitschen, A., Heid, U. (2004): SMOR: A German computational morphology covering derivation, composition and inflection. In Proceedings of LREC '04. Lisbon, Portugal.

Ansätze zur Verbesserung der Retrieval-Leistung kommerzieller Translation-Memory-Systeme

Dino Azzano^a, Uwe Reinke^b, Melanie Sauer^b

^aitl AG, ^bFachhochschule Köln

^aElsenheimerstr. 65, 80687 München

^bGustav-Heinemann-Ufer 54, 50968 Köln

E-mail: dino.azzano@gmail.com, uwe.reinke@fh-koeln.de, melanie.sauer@fh-koeln.de

Abstract

Translation-Memory-Systeme (TM-Systeme) zählen zweifelsohne zu den wichtigsten und am weitesten verbreiteten Werkzeugen der computergestützten Übersetzung. Kommerzielle Systeme sind inzwischen seit über zwei Jahrzehnten auf dem Markt. Im Hinblick auf die Erkennung semantischer Ähnlichkeiten wurde ihre Retrieval-Leistung bislang jedoch nicht entscheidend verbessert. Demgegenüber stellt die Computerlinguistik seit langem robuste Verfahren bereit, die zu diesem Zweck sinnvoll eingesetzt werden könnten. Ausgehend von den derzeitigen Grenzen der Retrieval-Leistung kommerzieller TM-Systeme zeigt der vorliegende Beitrag mögliche Ansätze zur Retrieval-Optimierung auf. Dabei wird zwischen Ansätzen mit und ohne Nutzung von linguistischem Wissen unterschieden. So genannte platzierbare und lokalisierbare Elemente können ohne linguistisches Wissen effizient behandelt werden. Im Gegensatz zu normalem Fließtext sind diese Elemente grundsätzlich eindeutig erkennbar und bleiben in der Übersetzung unverändert oder sie werden gemäß vorgegebenen Regeln angepasst. Die Erkennung mancher Elemente kann durch reguläre Ausdrücke erzielt werden und – basierend auf der Erkennung – verbessert eine optimierte Ähnlichkeitsberechnung das Retrieval der Segmente, in denen die Elemente vorkommen. Für die Optimierung des Retrievals von Paraphrasen und Subsegmenten (Phrasen, Teilsätzen) sowie für eine Verbesserung der Terminologieerkennung sind demgegenüber linguistische Verfahren erforderlich. Im Rahmen eines Forschungsprojekts wird an der Fachhochschule Köln derzeit versucht, vorhandene computerlinguistische Verfahren in kommerzielle TM-Systeme zu integrieren.

Keywords: computergestützte Übersetzung, Translation-Memory-Systeme, Retrieval-Optimierung, Fuzzy-Matching, platzierbare und lokalisierbare Elemente

1. Translation-Memory-Systeme

TM-Systeme sind Software-Applikationen, die den Übersetzungsprozess unterstützen und seit Jahren für alle am Übersetzungsprozess Beteiligten ein wichtiges computergestütztes Werkzeug darstellen. Ihr Hauptzweck ist die Wiederverwendung bereits übersetzten Textmaterials (Trujillo, 1999; Reinke, 2005). Unter den professionellen Übersetzern arbeitet die Mehrheit regelmäßig mit einem oder mehreren TM-Systemen (Massion, 2005; Lagoudaki, 2006). Zu den bekanntesten kommerziellen Produkten zählen Across, Déjà Vu, memoQ, MultiTrans, SDL Trados, Similis, Transit und Wordfast. Als nicht kommerzielles Produkt sei Omega-T erwähnt. Kernstück eines TM-Systems ist das Translation-Memory (TM), eine Datenbank oder eine Kollektion

von Dateien, welche Einzelsegmente – die in der Regel einem Satz entsprechen – in der Ausgangssprache sowie in mindestens einer Zielsprache enthält. Zwischen den ausgangssprachlichen und den zielsprachlichen Einträgen besteht eine feste Zuordnung. TMs stellen daher alignierte parallele Textkorpora dar, die Metainformationen (wie Anlagedatum, Erzeuger usw.) enthalten, aber nicht linguistisch annotiert sind (Kenning, 2010; Zinsmeister, 2010). Weitere Komponenten von TM-Systemen wie Terminologiedatenbank, Editor, Filter zur Konvertierung von Dateiformaten sowie Projektmanagementwerkzeuge seien hier nur aus Gründen der Vollständigkeit erwähnt.

TM-Systeme generieren keine eigenen Texte. Sie sind daher klar von Systemen zur maschinellen Übersetzung (MÜ) zu unterscheiden, wobei hybride Lösungen existieren.

tieren, die TM und MÜ integrieren. Kernaufgabe eines TM-Systems ist das Nachschlagen und Auffinden von Treffern im TM (Reinke, 2004; Jekat & Volk, 2010). Ein TM-System ist somit in erster Linie ein (monolinguales) Information-Retrieval-System. Die Suche erfolgt zunächst auf Segmentebene. Während der Übersetzung wird der zu bearbeitende Ausgangssprachliche Text segmentweise mit dem TM verglichen. Wird ein Ausgangssprachlicher Treffer gefunden, kann die zugeordnete Zielsprachliche Entsprechung zur Weiterverarbeitung verwendet werden. Dabei ist die Suche unscharf, so dass auch ähnliche Segmente (Fuzzy-Treffer) gefunden werden können (Sikes, 2007). Die Ähnlichkeit zwischen Suchanfrage und Treffer wird durch einen Prozentwert quantifiziert. Beim Übersetzen werden dem TM die neu erstellten Segmentpaare hinzugefügt, so dass dessen Umfang kontinuierlich zunimmt.

Moderne TM-Systeme bieten darüber hinaus Funktionen, um die Suche ggf. auf die Subsegmentebene auszudehnen. Hierbei werden Ausgangs- und Zielsprachliche Subsegmente einander mit Hilfe statistischer Verfahren zugeordnet (Macken, 2009; Chama, 2010) und während der Übersetzung vorgeschlagen (z.B. mit Hilfe einer Autovervollständigenden-Funktion).

Obwohl TM-Systeme inzwischen seit über zwei Jahrzehnten auf dem Markt sind, wurde ihre Retrieval-Leistung auf Segmentebene bislang qualitativ und quantitativ nicht entscheidend verbessert. Selbst Ansätze, die ohne linguistisches Wissen auskommen und somit auf recht einfache Weise Verbesserungen erzielen könnten, haben in kommerziellen TM-Systemen bislang wenig Beachtung gefunden. Diese werden zunächst im 2. Abschnitt des Beitrags behandelt. Der 3. Abschnitt geht dann auf Möglichkeiten zur linguistischen Optimierung der Retrieval-Leistung ein.

2. Retrieval-Optimierung ohne linguistisches Wissen

Bei den bisherigen Evaluierungen der Retrieval-Leistung von TM-Systemen wurde der Schwerpunkt auf den Fließtext gelegt (Reinke, 2004; Sikes, 2007; Baldwin, 2010). Das ist berechtigt, birgt jedoch die Gefahr, andere Textelemente, so genannte platzierbare und lokalisierbare Elemente, außer Acht zu lassen, die im Übersetzungsprozess eine beachtliche Rolle spielen.

2.1. Platzierbare und lokalisierbare Elemente

Platzierbare Elemente wie Tags, Inline-Grafiken und Felder bestehen nicht oder nur teilweise aus reinem Text und können häufig unverändert in den Zieltext übernommen werden. Tags sind Auszeichnungselemente in HTML- und XML-Dateien. XML-Formate haben in den letzten Jahren im Bereich der technischen Dokumentation – auch als Austauschformat – stark an Bedeutung gewonnen (Reinke, 2008; Anastasiou, 2010; Pelster, 2011) und spielen daher auch im Übersetzungsprozess eine wichtige Rolle. Inline-Grafiken und Felder sind typische Elemente in Desktop-Publishing-Formaten sowie in Formaten aus MS Word¹, die im Alltag der meisten Übersetzer von zentraler Bedeutung sind (Lagoudaki, 2006:12).

Lokalisierbare Elemente wie Zahlen, Datumsangaben, Eigennamen mit eindeutiger Oberflächenstruktur, URLs und E-Mail-Adressen sind hingegen Elemente aus reinem Text, die meist ohne linguistisches Wissen erkennbar sind und deren Lokalisierung – im Unterschied zum normalen Fließtext – vorgegebenen Regeln obliegt und häufig keine Auswirkung auf den restlichen Text hat.

2.2. Untersuchung

Im Rahmen einer Promotionsarbeit (Azzano, 2011) wurde untersucht, welchen Einfluss platzierbare und lokalisierbare Elemente auf das Retrieval kommerzieller TM-Systeme ausüben. Zu diesem Zweck wurden acht kommerzielle TM-Systeme verglichen: Across, Déjà Vu, Heartsome, memoQ, MultiTrans, SDL Trados, Transit und Wordfast. Aus unterschiedlichen Korpora wurden Segmente extrahiert, in denen platzierbare und lokalisierbare Elemente vorkamen, wobei möglichst viele Variationsmuster berücksichtigt wurden. Diese Segmente wurden anschließend mit den TM-Systemen bearbeitet, um die Erkennung der Elemente sowie die vorgeschlagenen Ähnlichkeitswerte zu prüfen.² Die Hauptergebnisse der vergleichenden Analyse werden im Folgenden zusammenfasst.

¹ Mit DOCX verwendet MS Word zwar ein XML-basiertes Format, seine Betrachtung und Bearbeitung im Übersetzungsprozess unterscheidet sich jedoch von üblichen XML-Dokumenten.

² Eine nähere Erläuterung der Testmethoden und Testdaten ist in diesem Beitrag aus Platzgründen nicht möglich; siehe Azzano (2011) für weitere Einzelheiten.

2.2.1. Recall

Prinzipiell sollten in einem TM gespeicherte Ausgangssprachliche Segmente auch dann gefunden werden, wenn sie sich vom aktuell zu übersetzenden Ausgangssprachlichen Segment nur durch die oben genannten Elemente unterscheiden.³ Die Tests zeigten jedoch, dass das Retrieval in solchen Fällen fehlschlagen kann oder dass es, wie im folgenden Beispiel, trotz minimalen Unterschieds zu sehr hohen Abzügen kommt:

- Armstrong stepped off Eagle's footpad [...]
- Armstrong stepped off *Eagle's* footpad [...]

Die meisten TM-Systeme bieten Ähnlichkeitswerte zwischen 91% und 99%; eines bietet aber 85% und eines nur 46%.

Auf Grund der kommerziellen Natur der untersuchten TM-Systeme und der dadurch bedingten Black-Box-Evaluierung lassen sich die Ursachen für diese Fehler nicht eindeutig identifizieren. Dennoch sind einige Rückschlüsse aus den Testergebnissen möglich.

Bei platzierbaren Elementen liegt die Ursache für hohe Abzüge manchmal darin, dass diese Elemente bei der Ermittlung der Segmentlänge wie übliche Wörter aus dem Fließtext gewichtet und damit überbewertet werden.⁴

Hingegen stellt ein fester Abzug für Unterschiede bei platzierbaren Elementen eine gute Lösung dar und wird von vier der getesteten TM-Systeme i.d.R. auch angewendet. Im Unterschied zu Fließtext kann die Art der Änderung (Hinzufügung, Löschung, Ersetzung oder Umstellung) bei der Gewichtung von Abzügen ignoriert werden. Ein fester Abzug sowie dessen Unabhängigkeit von der Art der Änderung dürften keine allzu großen Anpassungen der Algorithmen zur Ermittlung des Ähnlichkeitswertes in den kommerziellen TM-Systemen darstellen.

Bei lokalisierbaren Elementen kann die Retrieval-Leistung ebenfalls verbessert werden, wenn diese als „Sonderelemente“ und nicht als normaler Fließtext erkannt werden. Es können prinzipiell dieselben Strategien wie für platzierbare Elemente angewendet werden (z.B. fester Abzug). Zur Erkennung solcher Elemente, die bes-

³ Eine mögliche Ausnahme bilden hier solche platzierbare Elemente, die als Attribut- oder Feldwerte längere zu übersetzende Fragmente enthalten, wie z.B. **.

⁴ Die Segmentlänge, d.h. die Anzahl der Token (Wortanzahl) im Segment, wird als Normalisierungsfaktor zur korrekten Ermittlung des Änderungsumfanges im Verhältnis zum Segmentumfang verwendet (Trujillo, 1999; Manning & Raghavan & Schütze, 2008).

timten Mustern folgen, bewähren sich reguläre Ausdrücke. Aktuell zeigen kommerzielle TM-Systeme jedoch noch Schwächen. Zum einen sind Mechanismen zur Erkennung zwar prinzipiell vorhanden, aber sie schlagen oft fehl, beispielsweise bei Zahlen. Tabelle 1 führt die Erkennungsrate der untersuchten TM-Systeme bei Zahlentoken auf.⁵

TM-System	Erkennungsrate
Wordfast	0,99
memoQ	0,99
Across	0,96
Transit	0,90
Déjà Vu	0,89
SDL Trados	0,71

Tabelle 1: Erkennungsrate von Zahlen

Zum anderen werden einige lokalisierbare Elemente völlig ignoriert. Beispielsweise sind zuverlässige reguläre Ausdrücke zur Erkennung von URLs im reinen Text bereits ohne Weiteres verfügbar (Goyvaerts & Levithan, 2009), aber nur ein TM-System implementiert sie. Daher wurden reguläre Ausdrücke zur Erkennung der jeweiligen lokalisierbaren Elemente präsentiert bzw. entwickelt. Die Vorteile einer geeigneten Behandlung platzierbarer und lokalisierbarer Elemente gehen über das reine Retrieval hinaus. Solche Elemente können häufig automatisch ersetzt oder gelöscht werden, wobei der Rest des Fließtextes gleich bleibt. Diese – zum Teil in den TM-Systemen bereits angewendeten – automatischen Anpassungen können zum einen Zeit bei der Übersetzung sparen, zum anderen erhöhen sie den Ähnlichkeitswert.

2.2.2. Precision

Unter 2.2.1 wurden Beispiele präsentiert, bei denen die Ähnlichkeitswerte zu niedrig ausfallen. Allerdings tritt auch der umgekehrte Fall ein.

Nicht erkannt werden häufig Unterschiede zwischen den zu übersetzenden und den im TM gefundenen Segmenten, wenn sich die Position bzw. die Reihenfolge der platzierbaren Elemente unterscheidet. Im folgenden Beispiel bieten drei TM-Systeme für das zweite Segment einen 100%-Treffer, obwohl sich die Position der Tags geändert

⁵ Insgesamt wurden 79 Zahlentoken getestet, wobei jedes Token ein einmaliges Muster aufweist. Für weitere Informationen zu den Einzeltests und den getesteten Versionen siehe Azzano (2011).

hat und folglich auch in der Fremdsprache angepasst werden müsste.

- This statement is true *only* when [...]
- This statement is *true* only when [...]

Der Fehler liegt vermutlich darin, dass diese Elemente lediglich in ungeordneter Reihenfolge berücksichtigt bzw. vor der Auswertung durch inhaltsleere nicht-positionale Platzhalter ersetzt werden. Darüber hinaus wird die Anzahl der Änderungen teilweise nicht berücksichtigt so dass die Ähnlichkeitswerte zu positiv ausfallen. Im folgenden Beispiel bieten vier TM-Systeme für beide Variationen des ersten Segments den gleichen Ähnlichkeitswert, obwohl im dritten das Tag *
* zweimal hinzugefügt worden ist.

- Last transmission February 6, 1966, 22:55 UTC.
- Last transmission *
*February 6, 1966, 22:55 UTC.
- Last transmission *
*February 6, 1966, *
* 22:55 UTC.

All diese Unzulänglichkeiten dürften mit wenigen Eingriffen in die Retrieval-Algorithmen der TM-Systeme beseitigt werden können.

3. Retrieval Optimierung mit linguistischem Wissen

3.1. Aktuelle Ansätze

Grundsätzlich lassen sich bei der Optimierung der Retrieval-Ergebnisse von TM-Systemen zwei Zielsetzungen unterscheiden:

- 1) Die Verbesserung von Recall und Precision des (monolingualen) Retrievals (Optimierung der Treffermenge und des Rankings der Treffer)
 - a. auf Segmentebene
 - b. auf Subsegmentebene (Retrieval von ‚Chunks‘, (komplexen) Phrasen, Teilsätzen)
- 2) Die Anpassung der gefundenen Treffer zur Optimierung ihrer Wiederverwendbarkeit.

In der Forschung sind derzeit in erster Linie Bemühungen festzustellen, die Wiederverwendbarkeit von Fuzzy-Treffern durch Verfahren der statistischen maschinellen Übersetzung zu erhöhen (Biçici & Dymetman, 2008; Zhechev & van Genabith, 2010; Koehn & Senellart, 2010). Dabei werden solche Fragmente, die den Unterschied zwischen einem zu übersetzenden Segment und einem in der TM-Datenbank gefundenen Fuzzy-Treffer ausmachen,

mit Hilfe statistischer Übersetzungsverfahren so bearbeitet, dass die Anpassung der im Translation-Memory gefundenen Übersetzung an den aktuellen Kontext für den Übersetzer idealerweise keinen zusätzlichen Posteditionsaufwand bedeutet. Welche Auswirkung eine solche ‚Verschmelzung‘ von Humanübersetzung und maschineller Übersetzung auf Segmentebene tatsächlich auf die Postedition von Fuzzy-Treffern und somit auf Produktivität und Textqualität hat, müsste aber in jedem Fall empirisch untersucht werden.

Unter dem Aspekt einer effizienten Einbindung vorhandener linguistischer Verfahren in kommerzielle TM-Systeme scheint es zunächst durchaus lohnenswert, eine Optimierung von Recall und Precision zu verfolgen. Eines der wenigen kommerziellen TM-Systeme, das zur Optimierung der Retrieval-Leistung nicht nur zeichenkettenbasierte, sondern (einfache) computerlinguistische Verfahren anwendet, ist das Programm Similis der französischen Firma Lingua et Machina, das morphosyntaktische Analysen und flache Parsing-Verfahren einsetzt, um Fragmente unterhalb der Segmentebene zu identifizieren (Planas, 2005).

Neben der Identifikation von Subsegmenten ist vor allem auch eine Verbesserung des Retrievals solcher Ausgangssprachlicher Segmente erforderlich, die lediglich Paraphrasen bereits übersetzter Sätze darstellen und somit auf Zielsprachlicher Seite häufig keinerlei Veränderung erfordern. Durch morphosyntaktische, lexikalische und/oder syntaktische Variation gekennzeichnete Paraphrasen machen einen nicht zu unterschätzenden Anteil in solchen Fachtexten aus, die ständig aktualisiert, modifiziert und wiederverwendet werden.

3.2. Zielsetzungen und Ansätze im Rahmen des Projekts iMEM

Möglichkeiten, vorhandene computerlinguistische Verfahren in kommerzielle TM-Systeme zu integrieren, werden derzeit an der Fachhochschule Köln im Rahmen des Forschungsprojekts „Intelligente Translation Memories durch computerlinguistische Optimierung (iMEM)“ untersucht. iMEM zielt auf eine Optimierung der Retrieval-Leistung von TM-Systemen sowohl im Hinblick auf die bessere Erkennung von Fragmenten unterhalb der Segmentebene als auch hinsichtlich einer Optimierung der Verfahren zur Terminologieerkennung und -prüfung. Dabei sollen robuste Verfahren zur morphosyntaktischen

Analyse sowie zur regelbasierten Satzsegmentierung zum Einsatz kommen. Ziel ist die Entwicklung von Schnittstellenmodellen und prototypischen Schnittstellen zwischen kommerziellen TM-Systemen und „Lingware“. Hierbei werden exemplarisch das TM-System SDL Trados Studio 2009 und das morphosyntaktische Analysewerkzeug MPRO (Maas & Rösener & Theofilidis, 2009) eingesetzt. Ausgehend von den Sprachen Deutsch und Englisch sollen Erfahrungen für die Entwicklung weiterer Sprachmodule sowie für die Übertragung der Ergebnisse auf andere TM-Systeme gewonnen werden.

Für die Einbindung morphosyntaktischer Informationen in das TM-System wurde eine eigenständige SQL-Datenbank konzipiert, die aus den Daten der TM-Datenbank als paralleles „linguistisches TM“ aufgebaut wird und über entsprechende IDs mit der TM-Datenbank verknüpft ist. Das „linguistische TM“ enthält neben den Token der Textoberfläche derzeit im Wesentlichen Ergebnisse der mit MPRO durchgeführten Kompositaanalyse, wobei die Daten zur Beschleunigung des Retrievals in Form von Suffix Arrays (Aluru, 2004) vorgehalten werden.

In der Retrieval-Phase wird das aktuell zu übersetzende Segment zunächst analog zum „linguistischen TM“ linguistisch analysiert und annotiert, so dass ein Abgleich stattfinden kann. Die Abfrage des TM erfolgt in zwei unabhängigen Teilprozessen, bei denen einerseits die Tokenketten und andererseits die Ergebnisse der Kompositazerlegungen des zu übersetzenden Segments mit den entsprechenden Daten der im „linguistischen TM“ gespeicherten ausgangssprachlichen Segmente verglichen werden. Dabei werden für alle Ergebnisse der beiden Abfragen unter Verwendung von Generalized Suffix Arrays (GSA) (Rieck & Laskov & Sonnenburg, 2007) die längsten gemeinsamen Zeichenketten (Longest Common Substring, LCS) von zu übersetzendem Segment und im „linguistischen TM“ gefundenem Segment ermittelt. Für das Ranking der gefundenen Treffer ist noch eine Formel zu entwickeln, die die Ergebnisse beider Teilsuchen kombiniert und gewichtet, wobei jeweils u.a. Anzahl und Länge der LCS sowie deren Position in den zu vergleichenden Segmenten berücksichtigt werden sollte (vgl. auch Hawkins & Giraud-Carrier, 2009).

Im weiteren Verlauf des Projekts soll untersucht werden, inwieweit das bisherige Verfahren durch satzsyntaktische Analysen so erweitert werden kann, dass unterhalb der Segmentebene nicht nur vergleichsweise einfache Phra-

sen, sondern vor allem auch Übersetzungseinheiten wie Teilsätze und komplexe Phrasen gefunden werden, die für das computergestützte Humanübersetzen mit TM-Systemen relevant sind.

4. Fazit und Ausblick

Zusammenfassend kann festgestellt werden, dass für die Lösung der Retrieval-Probleme kommerzieller TM-Systeme aus Sicht der Computerlinguistik bewährte Verfahren zur Verfügung stehen, die aber bisher kaum oder nur vereinzelt Eingang in die TM-Systeme gefunden haben. Ein sprachunabhängiger, rein zeichenkettenbasierter Ansatz ohne Nutzung linguistischen Wissens, wie er derzeit bei fast allen kommerziellen TM-Systemen verfolgt wird, liefert ungeachtet seiner offensichtlichen Vorteile hinsichtlich der Sprachabdeckung keine optimalen Precision- und Recall-Werte. Es liegt daher nahe, einen differenzierteren Ansatz zu verfolgen und für die vom Übersetzungsvolumen her ‘großen’ Sprachen damit zu beginnen, vorhandene robuste Verfahren der linguistischen Datenverarbeitung in kommerzielle TM-Systeme zu integrieren. Bei Sprachen, für die entsprechende Verfahren nicht zur Verfügung stehen oder nicht ausreichend robust sind, kann zunächst weiter mit den herkömmlichen Retrieval-Mechanismen gearbeitet werden, wobei Verbesserungen in der Handhabung von platzierbaren und lokalisierbaren Elementen möglich sind.

5. Danksagung

Das Projekt „iMEM – Intelligente Translation Memories durch computerlinguistische Optimierung“ wird vom Bundesministerium für Bildung und Forschung im Rahmen des Programms „Forschung an Fachhochschulen“ gefördert.

6. Literatur

- Aluru, S. (2004): „Suffix Trees and Suffix Arrays“. In Mehta, D. P. und Sahni, S. (Eds.), *Handbook of Data Structures and Applications*. Boca Rayton: Chapman & Hall/CRC.
- Azzano, D. (2011): *Placeable and localizable elements in translation memory systems*. Dissertation. Ludwig-Maximilians-Universität München.
- Anastasiou, D. (2010): *Survey on the Use of XLIFF in Localisation Industry and Academia*. In *Proceedings of the 7th International Conference on Language Re-*

- sources and Evaluation.
- Baldwin, T. (2010): The hare and the tortoise: speed and accuracy in translation retrieval. *Machine Translation*, 23(4), pp. 195-240.
- Bıçıcı, E. und Dymetman, M. (2008): Dynamic Translation Memory: Using Statistical Machine Translation to improve Translation Memory Fuzzy Matches. In Gelbukh, A. F. (Ed.), *Computational Linguistics and Intelligent Text Processing*, 9th International Conference, Proceedings. Lecture Notes in Computer Science 4919. Berlin, Heidelberg: Springer, pp. 454-465.
- Chama, Z. (2010): Vom Segment zum Kontext. *technische kommunikation*, 32(2), pp. 21-25.
- Goyvaerts, J. und Levithan, S. (2009): *Regular expressions cookbook*. Sebastopol, O'Reilly.
- Hawkins, B. und Giraud-Carrier, C. (2009): „Ranking search results for translated content“. In Zhang, K. und Alhajj, R. (Eds.), *IRI'09 - Proceedings of the 10th IEEE international conference on Information Reuse & Integration*. Piscataway, NJ: IEEE Press, pp. 242-245.
- Jekat, S. und Volk, M. (2010): Maschinelle und computer-gestützte Übersetzung. In Carstensen, K.-U. et al. (Eds.), *Computerlinguistik und Sprachtechnologie: eine Einführung*. Heidelberg: Spektrum, pp. 642-658.
- Kenning, M.-M. (2010): What are parallel and comparable corpora and how can we use them? In O'Keefe, A. und McCarthy, M. (Eds.), *The Routledge Handbook of Corpus Linguistics*. New York: Routledge, pp. 487-500.
- Koehn, Ph. und Senellart, J. (2010): Convergence of Translation Memory and Statistical Machine Translation. In Zhechev, V. (Ed.), *Proceedings of the Second Joint EM+/CNGL Workshop "Bringing MT to the User: Research on Integrating MT in the Translation Industry"*.
- Lagoudaki, E. (2006): Translation Memory systems: Enlightening users' perspective. <http://www3.imperial.ac.uk/portal/pls/portallive/docs/1/7307707.PDF> (26.07.2011).
- Maas, H. D., Rösener, Ch. und Theofilidis, A. (2009): „Morphosyntactic and semantic analysis of text: The MPRO tagging procedure“. In Mahlow, C. und Piotrowski, M. (Eds.), *State of the Art in Computational Morphology: Workshop on Systems and Frameworks for Computational Morphology*. Proceedings. Berlin et al.: Springer, pp. 76-87.
- Macken, L. (2009): In search of the recurrent units of translation. In Daelemans, W. und Hoste, V. (Eds.), *Evaluation of Translation Technology*. Brussels: Academic and Scientific Publishers, pp. 195-212.
- Manning, C., Raghavan, P., Schütze, H. (2008): *Introduction to Information Retrieval*. Cambridge et al.: Cambridge University Press.
- Massion, F. (2005): *Translation-Memory-Systeme im Vergleich*. Reutlingen: Doculine.
- Pelster, U. (2011): XML für den passenden Zweck. *technische kommunikation*, 33(1), pp. 54-57.
- Planas, E. (2005): SIMILIS: Second-generation translation memory software. In *Translating and the Computer 27: Proceedings of the Twenty-seventh International Conference on Translating and the Computer*. London: Aslib.
- Reinke, U. (2004): *Translation Memories: Systeme – Konzepte – linguistische Optimierung*. Frankfurt am Main: Lang.
- Reinke, U. (2006): *Translation Memories*. In Brown, K. (Ed.), *Encyclopedia of Language and Linguistics*. Oxford: Elsevier, pp. 61-65.
- Reinke, U. (2008): XML-Unterstützung in Translation-Memory-Systemen. In *tekom Jahrestagung 2008, Zusammenfassung der Referate*. Stuttgart: Gesellschaft für technische Kommunikation e.V.
- Rieck, K., Laskov, P. und Sonnenburg, S. (2007): Computation of Similarity Measures for Sequential Data using Generalized Suffix Trees. In Schölkopf, B., Platt, J. und T. Hoffman (Eds.), *Advances in Neural Information Processing Systems 19*. Cambridge, MA: MIT Press, pp. 1177-1184.
- Sikes, R. (2007): Fuzzy matching in theory and practice. *MultiLingual*, 18(6), pp. 39-43.
- Trujillo, A. (1999): *Translation Engines: Techniques for Machine Translation*. London: Springer.
- Zinsmeister, H. (2010): Korpora. In Carstensen, K.-U. et al. (Eds.), *Computerlinguistik und Sprachtechnologie: eine Einführung*. Heidelberg: Spektrum, pp. 482-491.
- Zhechev, V. und van Genabith, J. (2010): Maximising TM Performance through Sub-Tree Alignment and SMT. In *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas*.

WikiWarsDE: A German Corpus of Narratives Annotated with Temporal Expressions

Jannik Strötgen, Michael Gertz

Institute of Computer Science, Heidelberg University

Im Neuenheimer Feld 348, 69120 Heidelberg, Germany

E-mail: stroetgen@uni-hd.de, gertz@uni-hd.de

Abstract

Temporal information plays an important role in many natural language processing and understanding tasks. Therefore, the extraction and normalization of temporal expressions from documents are crucial preprocessing steps in these research areas, and several temporal taggers have been developed in the past. The quality of such temporal taggers is usually evaluated using annotated corpora as gold standards. However, existing annotated corpora only contain documents of the news domain, i.e., short documents with only few temporal expressions. A remarkable exception is the recently published corpus WikiWars, which is the first temporal annotated English corpus containing long narratives that are rich in temporal expressions. Following this example, in this paper, we describe the development and the characteristics of WikiWarsDE, a new temporal annotated corpus for German. Additionally, we present evaluation results of our temporal tagger HeidelbergTime on WikiWarsDE and compare them with results achieved on other corpora. Both, WikiWarsDE as well as our temporal tagger HeidelbergTime are publicly available.

Keywords: temporal expression, TIMEX2, corpus annotation, temporal information extraction

1. Introduction and Related Work

In the last decades, the extraction and normalization of temporal expressions have become hot topics in computational linguistics. In many research areas, temporal information plays an important role, e.g., in information extraction, document summarization, and question answering (Mani et al., 2005). In addition, temporal information is valuable in information retrieval and can be used to improve search and exploration tasks (Alonso et al., 2011). However, the tasks of extracting and normalizing temporal expressions are challenging due to the fact that there are many different ways to express temporal information in documents and that temporal expressions may be ambiguous.

Besides explicit expressions (e.g., “April 10, 2005”) that can directly be normalized to some standard format, relative and underspecified expressions are very common in many types of documents. To determine the semantics of such expressions, context information is required. For example, to normalize the expression “Monday” in phrases like “on Monday”, a reference

time and the relation to the reference time have to be identified. Depending on the domain of the documents that are to be processed, this reference time can either be the document creation time or another temporal expression in the document. While the document creation time plays an important role in news documents, it is almost irrelevant in narrative style documents, e.g., documents about history or biographies. Despite these challenges, all applications using temporal information mentioned in documents rely on high quality temporal taggers, which correctly extract and normalize temporal expressions from documents.

Due to the importance of temporal tagging, there have been significant efforts in the area of temporal annotation of text documents. Annotation standards such as TIDES TIMEX2 (Ferro et al., 2005) and TimeML (Pustejovsky et al., 2003b; Pustejovsky et al., 2005) were defined and temporal annotated corpora like TimeBank (Pustejovsky et al., 2003a) were developed – although most of the corpora contain English documents

only. Furthermore, research challenges were organized where temporal taggers were evaluated. The ACE (Automatic Content Extraction) time expression and normalization (TERN) challenges were organized in 2004, 2005, and 2007.¹ In 2010, temporal tagging was one task in the TempEval-2 challenge (Verhagen et al., 2010). However, so far, research was limited to the news domain, i.e., the documents of the annotated corpora are short with only a few temporal expressions. The temporal discourse structure is thus usually easy to follow. Only recently, a first corpus containing narratives was developed (Mazur & Dale, 2010). This corpus, called WikiWars, consists of Wikipedia articles about famous wars in history. The documents are much longer than news documents and contain many temporal expressions. As the developers point out, normalizing the temporal expressions in such documents is more challenging due to the rich temporal discourse structure of the documents.

Motivated by this observation and by the fact that no temporal annotated corpus for German was publicly available so far, we created the WikiWarsDE corpus², which we present in this paper. WikiWarsDE contains the corresponding German articles of the documents of the English WikiWars corpus. For the annotation process, we followed the suggestions of the WikiWars developers, i.e., annotated the temporal expressions according to the TIDES TIMEX2 annotation standard using the annotation tool Callisto³. To be able to use publicly available evaluation scripts, the format of the ACE TERN corpus was selected. Thus, evaluating a temporal tagger on the WikiWarsDE corpus is straightforward and evaluation results of different taggers can be compared easily.

The remainder of the paper is structured as follows. In Section 2, we describe the annotation schema and the corpus creation process. Then, in Section 3, we present detailed information about the corpus such as statistics on the length of the documents and the number of temporal expressions. In addition, evaluation results of

our own temporal tagger on the WikiWarsDE corpus are presented. Finally, we conclude our paper in Section 4.

Temporal Expression	Value of the VAL attribute
November 12, 2001	2001-11-12
9:30 p.m.	2001-11-12T21:30 ⁴
24 months	P20M
daily	XXXX-XX-XX

Table 1: Normalization examples (VAL) of temporal expressions of the types date, time, duration, and set.

2. Annotation Schema and Corpus Creation

In Section 2.1, we describe the annotation schema, which we used for the annotation of temporal expressions in our newly created corpus. Furthermore, we explain the task of normalizing temporal expressions using some examples. Then, in Section 2.2, we detail the corpus creation process and explain the format, in which WikiWarsDE is publicly available.

2.1. Annotation Schema

Following the approach of Mazur and Dale (2010), we use TIDES TIMEX2 as annotation schema to annotate the temporal expressions in our corpus. The TIDES TIMEX2 annotation guidelines (Ferro et al., 2005) describe how to determine the extents of temporal expressions and their normalizations. In addition to date and time expressions, such as “November 12, 2001” and “9:30 p.m.”, temporal expressions describing durations and sets are to be annotated as well. Examples for expressions of the types duration and set are “24 months” and “daily”, respectively.

The normalization of temporal expressions is based on the ISO 8601 standard for temporal information with some extensions. The following five features can be used to normalize a temporal expression:

- VAL (value)
- MOD (modifier)
- ANCHOR_VAL (anchor value)
- ANCHOR_DIR (anchor direction)
- SET

The most important feature of a TIMEX2 annotation is the “VAL” (value) feature. For the four examples above,

¹ The 2004 and 2005 training sets and the 2004 evaluation set are released by the LDC as is the TimeBank corpus; see <http://www ldc.upenn.edu/>

² WikiWarsDE is publicly available on http://dbs.ifi.uni-heidelberg.de/temporal_tagging/

³ <http://callisto.mitre.org/>

⁴ Assuming that “9:30 p.m.” refers to 9:30 p.m. on November 12, 2001.

the values of VAL are given in Table 1. Furthermore, “MOD” (modifier) is used, for instance, for expressions such as “the end of November 2001”, where MOD is set to “END”, i.e., to capture additional specifications not captured by VAL. ANCHOR_VAL and ANCHOR_DIR are used to anchor a duration to a specific date, using the value information of the date and specifying whether the duration starts or ends on this date. Finally, SET is used to identify set expressions.

Often, for example in the TempEval-2 challenge, the normalization quality of temporal taggers is evaluated based on the VAL (value) feature, only. This fact points out the importance of this feature and was the motivation to evaluate the normalization quality of our temporal tagger based on this feature as described in Section 3.

2.2. Corpus Creation

For the creation of the corpus, we followed Mazur and Dale (2010), the developers of the English WikiWars corpus. We selected the 22 corresponding German Wikipedia articles and manually copied sections describing the course of the wars.⁵ All pictures, cross-page references, and citations were removed. All text files were then converted into SGML files, the format of the ACE TERN corpora containing “DOC”, “DOCID”, “DOCTYPE”, “DATETIME”, and “TEXT” tags. The document creation time was set to the time of downloading the articles from Wikipedia. The “TEXT” tag surrounds the text that is to be annotated.

Similar to Mazur and Dale (2010), we used our own temporal tagger, which is described in Section 3.2, containing a rule set for German as a first-pass annotation tool. The output of the tagger can then be imported to the annotation tool Callisto for manual correction of the annotations. Although this fact has to be taken into account when comparing the evaluation results on WikiWarsDE of our temporal tagger with other taggers, this procedure is motivated by the fact that “annotator blindness” is reduced to a minimum, i.e., that annotators miss temporal expressions. Furthermore, the annotation effort is reduced significantly since one does

not have to create a TIMEX2 tag for the expressions already identified by the tagger.

At the second annotation stage, the documents were examined for temporal expressions missed by the temporal tagger and annotations created by the tagger were manually corrected. This task was performed by two annotators – although Annotator 2 only annotated the extents of temporal expressions. The more difficult task of normalizing the temporal expressions was performed by Annotator 1 only, since a lot of experience in temporal annotation is required for this task. At the third annotation stage, the results of both annotators were merged and in cases of disagreement the extents and normalizations were rechecked and corrected by Annotator 1.

To compare our inter-annotator agreement for the determination of the extents of temporal expressions to others, we calculated the same measures as the developers of the TimeBank-1.2 corpus. They calculated the average of precision and recall with one annotator's data as the key and the other's as the response. Using a subset of ten documents, they report inter-annotator agreement of 96% and 83% for partial match (lenient) and exact match (strict), respectively.⁶ Our scores for lenient and exact match on the whole corpus are 96.7% and 81.3%, respectively.

Finally, the annotated files, which contain inline annotations, were transformed into the ACE APF XML format, a stand-off markup format used by the ACE evaluations. Thus, the WikiWarsDE corpus is available in the same two formats as the WikiWars corpus, and the evaluation tools of the ACE TERN evaluations can be used with this German corpus as well.

3. Corpus Statistics and Evaluation Results

In this section, we first present some statistical information about the WikiWarsDE corpus, such as the length of the documents and the number of temporal expressions in the documents (Section 3.1). Then, in Section 3.2, we shortly introduce our own temporal tagger HeidelbergTime, present its evaluation results on WikiWarsDE, and compare them with results achieved on other corpora.

⁵ Due to the shortness of the Wikipedia article about the Punic Wars in general, we used sections of three separate articles about the 1st, 2nd, and 3rd Punic Wars.

⁶ For more information on TimeBank, see <http://timeml.org/site/timebank/documentation-1.2.html>.

Corpus	Docs	Token	Timex	Token / Timex	Timex / Document
ACE 04 en train	863	306.463	8.938	34,3	10,4
TimeBank 1.2	183	78.444	1.414	55,5	7,7
TempEval2 en train	162	53.450	1.052	50,8	6,5
TempEval2 en eval	9	4.849	81	59,9	9,0
WikiWars	22	119.468	2.671	44,7	121,4
WikiWarsDE	22	95.604	2.240	42,7	101,8

Table 2: Statistics of the WikiWarsDE corpus and other publicly available or released corpora.

3.1. Corpus Statistics

The WikiWarsDE corpus contains 22 documents with a total of more than 95,000 tokens and 2,240 temporal expressions. Note that the fact that the WikiWars corpus contains almost 25,000 tokens more than WikiWarsDE can be partly explained by the differences between the two languages. In German compounds are very frequent, e.g., the 3 English tokens "course of war" is just 1 token in German ("Kriegsverlauf").

In Table 2, we present some statistics of the corpus in comparison to other publicly available corpora. On the one hand, the density of temporal expressions (Token/Timex) is similar among the documents of all the corpora. In WikiWarsDE, one temporal expression occurs every 42.7 tokens on average.

On the other hand, one can easily see that the documents of the WikiWarsDE and the WikiWars corpora are much longer and contain many more temporal expressions than the documents of the news corpora. While WikiWars and WikiWarsDE contain 121.4 and 101.8 temporal expressions per document on average, the number of temporal expressions on the news corpora ranges between 6.5 and 10.4 temporal expressions only. Thus, the temporal discourse structure is much more complex for the narrative-style documents in WikiWars and WikiWarsDE. Further statistics on the single documents of WikiWarsDE are published with the corpus.

3.2. Evaluation Results

After the development of the corpus, we evaluated our temporal tagger HeidelbergTime on the corpus. HeidelbergTime

is a multilingual, rule-based temporal tagger. Currently, two languages are supported (English and German), but due to the strict separation between the source code and the resources (rules, extraction patterns, normalization information), HeidelbergTime can be easily adapted to further languages. In the TempEval-2 challenge, HeidelbergTime achieved the best results for the extraction and normalization of temporal expressions from English documents (Strötgen & Gertz, 2010; Verhagen et al., 2010). Since HeidelbergTime uses different normalization strategies depending on the type of the documents that are to be processed (news- or narrative-style documents), we were able to show that HeidelbergTime achieves high quality results on both kinds of documents for English.⁷

With the development of WikiWarsDE, we are now able to evaluate HeidelbergTime on a German corpus as well. For this, we use the well-known evaluation measures of precision, recall, and f-score. In addition, we distinguish between lenient (overlapping match) and strict (exact match) measures. For the normalization, one can calculate the measures for all expressions that were correctly extracted by the system (value). This approach is used by the ACE TERN evaluations. However, similar to Ahn et al. (2005) and Mazur and Dale (2010), we argue that it is more meaningful to combine the extraction with the normalization tasks, i.e., to calculate the measures for all expressions in the corpus (lenient+value and strict+value).

⁷ More information on HeidelbergTime, its evaluation results on several corpora, as well as download links and an online demo can be found at <http://dbs.ifi.uni-heidelberg.de/heideltime/>.

Corpus	lenient			strict			value			lenient + value			strict + value		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
TimeBank 1.2	90.5	91.4	90.9	83.5	84.3	83.9	86.2	86.2	86.2	78.0	78.8	78.4	73.2	74.0	73.6
WikiWars	93.9	82.4	87.8	86.0	75.4	80.4	89.5	90.1	89.8	84.1	73.8	78.6	79.6	69.8	74.4
WikiWarsDE	98.5	85.0	91.3	92.6	79.9	85.8	87.0	87.0	87.0	85.7	74.0	79.4	82.5	71.2	76.5

Table 3: Evaluation results of our temporal tagger on an English news corpus (TimeBank 1.2), an English narratives corpus (WikiWars) and our newly created German narratives corpus WikiWarsDE.

On WikiWarsDE, HeidelTime achieves f-scores of 91.3 and 85.8 for the extraction (lenient and strict, respectively) and 79.4 and 76.5 for the normalization (lenient + value and strict + value, respectively).

For comparison, we present the results of HeidelTime on some English corpora. As shown in Table 3, our temporal tagger achieves equally good results on both, the narratives corpora (WikiWars and WikiWarsDE) and the news corpus (TimeBank). Note that our temporal tagger uses different normalization strategies depending on the type of the corpus that is to be processed. This might be the main reason why HeidelTime clearly outperforms the temporal tagger of the WikiWars developers. For the WikiWars corpus, Mazur and Dale (2010) report f-scores for the normalization of only 59,0 and 58,0 (lenient + value and strict + value, respectively). Compared to these values, HeidelTime achieves much higher f-scores, namely 78.6 and 74.4, respectively.

4. Conclusions

In this paper, we described WikiWarsDE, a temporal annotated corpus containing German narrative-style documents. After presenting the creation process and statistics of WikiWarsDE, we used the corpus to evaluate our temporal tagger HeidelTime. While Mazur and Dale (2010) report lower evaluation results of their temporal tagger on narratives than on news documents, HeidelTime achieves similar results on both types of corpora. Nevertheless, we share their opinion that the normalization of temporal expressions on narratives is challenging. However, using different normalization strategy for different types of documents (news-style and narrative-style documents), this problem can be tackled.

By making available WikiWarsDE and HeidelTime, we provide useful contributions to the community in support of developing and evaluating temporal taggers and of improving temporal information extraction.

5. Acknowledgements

We thank the anonymous reviewers for their valuable suggestions to improve the paper.

6. References

- Ahn, D., Adafre, S.F., de Rijke, M. (2005): Towards Task-Based Temporal Extraction and Recognition. In G. Katz, J. Pustejovsky, F. Schilder (Eds.), *Extracting and Reasoning about Time and Events*. Dagstuhl, Germany: Dagstuhl Seminar Proceedings.
- Alonso, O., Strötgen, J., Baeza-Yates, R., Gertz, M. (2011): Temporal Information Retrieval: Challenges and Opportunities. In *Proceedings of the 1st International Temporal Web Analytics Workshop (TAW)*, pp. 1–8.
- Ferro, L., Gerber, L., Mani, I., Sundheim, B., Wilson, G. (2005): TIDES 2005 Standard for the Annotation of Temporal Expressions. Technical report, The MITRE Corporation.
- Mani, I., Pustejovsky, J., Gaizauskas, R.J. (2005): *The Language of Time: A Reader*. Oxford University Press.
- Mazur, P., Dale, R. (2010): WikiWars: A New Corpus for Research on Temporal Expressions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 913-922.
- Pustejovsky, J., Hanks, P., Sauri, R., See, A., Gaizauskas, R.J., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., Lazo, M. (2003a): The

- TIMEBANK Corpus. In Proceedings of Corpus Linguistics 2003, pp. 647–656.
- Pustejovsky, J., Castaño, J.M., Ingria, R., Sauri, R., Gaizauskas, R.J., Setzer, A., Katz, G. (2003b): TimeML: Robust Specification of Event and Temporal Expressions in Text. In New Directions in Question Answering, pp 28–34.
- Pustejovsky, J., Knippen, R., Littman, J., Sauri, R. (2005): Temporal and Event Information in Natural Language Text. Language Resources and Evaluation, 39(2-3):123–164.
- Strötgen, J., Gertz, M. (2010): HeidelTime: High Quality Rule-based Extraction and Normalization of Temporal Expressions. In Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval), pp. 321–324.
- Verhagen, M., Sauri, R., Caselli, T., Pustejovsky, J. (2010): SemEval-2010 Task 13: TempEval-2. In Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval), pp. 57–62.

Translation and Language Change with Reference to Popular Science Articles: The Interplay of Diachronic and Synchronic Corpus-Based Studies

Sofia Malamatidou

University of Manchester

Oxford Road, M13 9PL

E-mail: sofia.malamatidou@manchester.ac.uk

Abstract

Although a number of scholars have adopted a corpus-based approach to the investigation of translation as a form of language contact and its impact on the target language (Steiner, 2008; House, 2004; 2008; Baumgarten et al. 2004), no sustained corpus-based study of translation involving Modern Greek has so far been attempted and very few diachronic corpus-based studies (Amouzadeh & House, 2010) have been undertaken in the field of translation. This study aims to combine synchronic and diachronic corpus-based approaches, as well as parallel and comparable corpora for the analysis of the linguistic features of translated texts and their impact on non-translated ones. The corpus created captures a twenty-year period (1990-2010) and is divided into three sections, including non-translated and translated Modern Greek popular science articles published in different years, as well as the source texts of the translations. Unlike most studies employing comparable corpora, which focus on revealing recurrent features of translated language independently of the source and target language, this study approaches texts with the intention of revealing features that are dependent on the specific language pair involved in the translation process.

Keywords: corpus-based translation studies, language change, diachronic corpora, Modern Greek, passive voice

1. Introduction

Translation as a language contact phenomenon is a phenomenon that neither linguistics nor translation studies has addressed in depth. However, in the era of the information society, the translation of popular science texts tends to be very much a unidirectional process from the dominant lingua franca, which is English, into less widely spoken languages such as Modern Greek. This process is likely to encourage changes in the communicative conventions of the target language. Given the fact that the genre of popular science was developed in Greece mainly through translations from Anglophone sources in the last two decades, it is interesting to examine whether and how the translations from English encouraged the dissemination of particular linguistic features in the target language in the discourse of this particular genre. A number of scholars, mostly within the English-German context, have taken interest in investigating translation as a form of language contact and its effects on the target language. Steiner (2008) has investigated grammatical and syntactic features of

explicitness as a result of the contact between English and German, which however did not involve diachronic analyses of corpora. Most importantly, House and a group of scholars have investigated how translation from English affects German, but also Spanish and French (House, 2004; 2008; Baumgarten et al. 2004; Becher et al. 2009). However, these studies mainly involved manual analyses of texts, that is, they were not corpus-based studies as they are understood by Baker (1995), i.e. they did not involve the automatic or semi-automatic analysis of machine-readable texts. Diachronic corpus-based approaches to translation are limited (Amouzadeh & House, 2010) and in terms of Modern Greek, no similar study has ever been conducted.

This study aims to examine whether and how translation can encourage linguistic changes in the target language by investigating a diachronic corpus of non-translated and translated Modern Greek popular science articles, along with their source texts, in order to examine how translation can be understood as a language contact phenomenon. The linguistic change that is examined is

the frequency of the passive voice, since it has been claimed to be found more frequently in translated Modern Greek texts (Apostolou-Panara, 1991), especially those translated from English.

This paper first presents the theoretical model that informs the study, namely the Code-Copying Framework (Johanson, 1993; 1999; 2002). Then the research methodology is presented in detail and data analysis techniques are analysed. Finally, some preliminary findings are discussed. It must be mentioned, that this is still an ongoing project and for that reason the results are limited to a number of small sample studies.

2. The Code-Copying Framework

The Code-Copying Framework is a widely applicable linguistic model that is suitable for the description of phenomena that have consistently been neglected, such as translation as a form of language contact and a propagator of change. Some of its concepts have recently been used by translation scholars to describe similar phenomena (Steiner, 2008), suggesting that it is a conceptual model suitable for analysing diverse cases of language contact, in particular cases where translation plays a central role in the dissemination of linguistic features.

The Code-Copying Framework was developed by Johanson (1993; 1999; 2002) who is critical of the terminology, especially that of borrowing, used in the field of language change studies and it is this critique that serves as a point of departure towards developing a new explanatory framework of language contact, where 'copying' replaces traditional terms and provides a different vintage point from which to analyse the phenomenon. Johanson (1999:39) argues that in any situation of code-interaction, that is, in a situation where two or more codes interact, two linguistic systems, i.e. two codes are employed. The Model Code is the source code, whereas the Basic Code is the recipient code which also provides the necessary morphosyntactic and other information for inserting and adapting the copied material (Johanson, 2008:62). Although, there are different directions of copying, this study focuses on the case of 'adoption' which involves elements being inserted from the Model Code into the Basic Code and

views translation as a language contact situation where translators are likely to copy elements from the source language, i.e. the Model Code, when translating into the target language, which is the Basic Code.

Two types of copying are possible within this model: global and selective copying. The linguistic properties that can be copied are material (i.e. phonic), semantic, combinational (i.e. collocations and syntax) and frequential properties, namely the frequency of particular linguistic units. In the case of global copying, a linguistic item is copied along with all its aforementioned properties. In the case of selective copying, one or more properties are copied resulting in distinct types of copying. Thus, there is material (M), semantic (S), combinational (C) and frequential (F) copying.

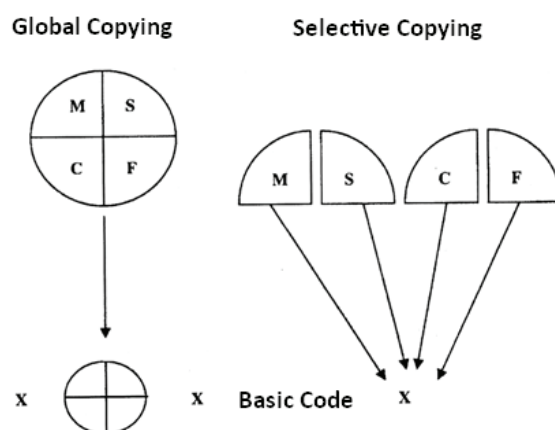


Figure 1: The Code-Copying Framework (Johanson, 2006:5)

During the process of translation, selective copying is more probable than global copying (Verschik, 2008:133). For that reason, the type of copying that is dealt with in this study is selective copying and in particular frequential copying, which results in a change in the frequency patterns of an existing linguistic unit. Apostolou-Panara (1991) notes that the passive constructions are used more frequently in Modern Greek than they once were. Traditionally, it has been argued that the passive voice structures are used in Modern Greek though not as often as in English (Warburton, 1975:576), where the passive voice is quite frequent especially in terms of informative texts such as popular science articles. As far as translation is concerned, different frequencies

and proportionalities of native patterns often result in texts having a 'non-native' feeling (Steiner, 2008:322). The frequent translation of source text patterns with grammatical, yet marginal, target language linguistic patterns may ultimately override prevailing patterns and result in new communicative preferences in the target language (Baumgarten & Özçetin 2008:294).

Copies usually begin as momentary code-copies, that is, occasional instances of copying. When copies start being frequently and regularly used by a group of individuals or by a particular speech community, they become habitualised code-copies. Copies may also become conventionalised and become integrated and widely accepted by a speech community. The final stage is for copies to become monolingual, i.e. when copies are used by monolinguals and do not presuppose any bilingual ability (Johanson, 1999:48). Since momentary copies are difficult to trace (Csató, 2002:326), emphasis in this study is placed on habitualised code-copies. Translators are considered as part of a particular speech community and copies are regarded as habitualised when they are frequently and regularly used by translators. Conventionalised copies are not examined in this study, since they presuppose measuring social evaluation that is outside the scope of this research. However, it is safe to assume that if a copy is monolingualised, that is, it is used in non-translated texts; it is also in general terms socially approved.

Translation in this study is understood as a social circumstance facilitating copying. It is not considered as a cause of change, but rather as an instance of contact during which copying may occur and change may proliferate through language, since translated texts, especially newspaper and magazine articles, are widely circulating texts that are likely to exert a powerful linguistic impact on a large audience. The main factors of copying are considered to be extra-linguistic, especially the cultural dominance of English in relation to Modern Greek, as far as the production of scientific texts is concerned, and the prestige that English enjoys as a prominent language and culture, both in the general sense of a lingua franca and in terms of scientific research.

3. Data and Methodology

3.1. Corpus design

Based on the availability of data and the research aims of this thesis, an approximately 500,000 corpus of Modern Greek non-translated and translated popular science articles, along with their source texts was created. The corpus is named TROY (TRAnslation Over the Years) and covers a 20-year period (1990-2010), which is considered to be an adequate time span for language change to occur and is amenable to being systematically observed.

Newspapers and magazines dedicated to scientific issues provide are the two main sources of popular science articles. The corpus is specialised in terms of both genre and domain, i.e. it involves popular science articles from the domain of technology and life sciences. These domains were chosen due to the fact that the majority of articles, especially translations, seem to belong to either one of the two domains. This in turn indicates that interest is expressed for these domains from the general public, which consequently suggests that a high number of people will read articles belonging to the domains of technology and life sciences, a fact that is likely to result in a powerful linguistic impact on a large audience.

The TROY corpus is divided into three subcorpora. The first subcorpus consists of non-translated Modern Greek popular science articles published in 1990-1991. The second subcorpus consists of non-translated and translated Modern Greek popular science articles published in 2003-2004, as well as the source texts of the translations. The years 2003-2004 were selected because translations of popular science texts started circulating more widely in Greece during that period than in previous years. The third subcorpus includes non-translated as well as translated texts and their source texts, all published in 2009-2010. The subcorpora are evenly balanced, both in terms of their overall size and between the two domains.

3.2. Corpus Methodology

The corpus methodology employed in this study has three aims. Firstly, it aims to investigate whether certain features have changed over time in Modern Greek. Secondly, it aims to examine whether this change is

related or mirrored in the process of translation. Finally, it aims to investigate whether influence can be traced back to the English source texts. Ultimately, this methodology aims at combining most corpus-based methodologies under one research aim. Thus, synchronic and diachronic corpus-based approaches, as well as parallel and comparable corpora are employed in order to illustrate the way in which combined methodologies can assist in the analysis of the linguistic features of translated texts and their impact on non-translated ones.

Firstly, the corpus methodology aims at examining language change in Modern Greek and in particular to investigate whether the frequency of the passive voice has changed over time. This involves a longitudinal corpus-based study, during which a comparable corpus is analysed diachronically. For the purposes of this study, the non-translated articles published in 1990-1991 will be compared to the non-translated articles published in 2009-2010.

The second aim of this corpus-based methodology is to examine the role of translation in this language change phenomenon. This involves a comparable corpus-based analysis where translated and non-translated Modern Greek popular science articles are analysed synchronically. First, the non-translated articles published in 2003-2004 will be compared to the translated articles published during the same years. Then, the same type of analysis will be conducted for articles published in 2009-2010. Two separate phases of analysis are included in order to investigate the extent to which the linguistic features in the translated texts differ from those of the non-translated ones at different time periods. More particularly, the first phase of analysis focuses on a period of time when the influence from English translations of popular science articles was at its initial stage. The second phase of analysis focuses on a later stage of the contact between English and Modern Greek through translation, as far as the particular genre or popular science is concerned.

Finally, this corpus-based methodology aims to investigate the role of the source texts in this language contact situation. This involves the synchronic analysis of a parallel corpus of translated articles and their

originals, which consists of two phases of analysis, i.e. the translated popular science articles that were published in 2003-2004 will be compared to their source texts and the same analysis will be conducted for the articles published in 2009-2010.

The analyses will be conducted with the help of the Concordance tool of WordSmith Tools 5.0 and will be based on semi-automatic methods, since at points where a closer examination of the texts is required, they will be analysed manually. The verb form is considered to be the unit of analysis and auxiliary verbs are excluded from the counts, since they do not provide any lexical information. For the sample studies discussed below, a part-of-speech (POS) tagger is not being used due to the fact that available Modern Greek POS taggers score relatively low on accuracy and Modern Greek verbs can be quite accurately identified from their suffixes with the use of wildcards.

4. Preliminary Results

Although this is still an ongoing project, a number of sample studies indicate that a corpus-based methodology that combines synchronic and diachronic corpus-based approaches, as well as parallel and comparable corpora can considerably assist in the analysis of the linguistic features of translated texts and their impact on non-translated ones. Articles for the sample studies are taken from the newspaper *Βήμα* (The Tribune), which includes a section dedicated to scientific issues.

4.1. Language Change in Modern Greek

In terms of the first aim of this corpus-based methodology, that is, the examination of language change in Modern Greek, a sample study of popular science articles published in 1991 and 2010 involving 4,000 words was conducted in order to examine changes in the frequency of the passive voice. Although this is a very small sample study, it was found that the passive voice has become more frequent in Modern Greek in the last 20 years, at least in terms of the specific genre of popular science articles. In particular, in the articles published in 1991, 273 verb forms were found, 42 of which involved passive verb forms. In the articles published in 2010, 217 instances of verb forms were identified, 42 of which were passive. This means that there is an approximately 5%

increase in the frequency of the distribution of passive voice constructions in Modern Greek. However, this 5% increase may be attributed to a number of factors that are irrespective of contact-induced language change, i.e. it may be a result of internal language changes. An analysis of translated texts is necessary in order to establish the extent to which contact through translation has encouraged a frequential copying of passive voice structures from English.

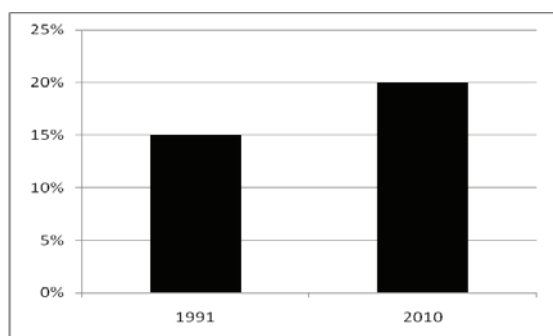


Figure 2: Change in the frequency of the passive voice in Modern Greek (1991-2010)

4.2. The Role of the Translations

A second sample study was conducted in order to examine the role of the translation in this language change situation. In particular, a small corpus of 20,000 words taken from translated and non-translated Modern Greek popular science articles published in 2010 was analysed. The analysis revealed that the frequency of the passive voice in the translated and non-translated articles is very similar, i.e. approximately 20%. In the non-translated articles, 1,081 verb instances were identified, 215 of which were passive, whereas the translated articles included 1,234 verb forms, out of which 243 involved passive voice occurrences.

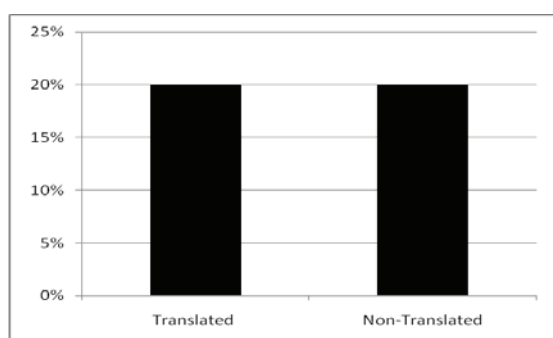


Figure 3: Frequency of the passive voice in translated and non-translated articles published in 2010

This similarity in terms of the proportions of the passive voice suggests that the translated texts at least mirror the changes in the frequency of the passive voice that is attested in Modern Greek. This sample study focuses on a later stage of contact between English and Modern Greek in terms of popular science publications and it is assumed that this later stage indicates more established instances of copying, if we accept that some kind of copying has taken place. Although a comparable analysis of articles published in 2003-2004, when the influence from Anglophone source texts was at its initial stage, has not yet been attempted, such an analysis is likely to reveal a different patterning than the one discussed above, i.e. that the frequency of the passive voice is higher in translated texts than in non-translated ones. This will indicate that the frequential copying of passive voice gradually habitualised in the context of translation.

4.3. The Role of the Source Texts

Finally, in terms of the last aim of this corpus-based methodology, namely the investigation of the role of the English source texts in this language change phenomenon, it should be mentioned that although a sample study is not available at the moment for this type of analysis, it can be predicted based on the previous sample study that translated texts are likely to follow the patterns of the source texts. Corpus studies (Biber et al. 1999:476) suggest that the English passives account for approximately 25% of all finite verbs in academic prose and for 15% in news. Popular science articles are considered to be somewhere in between these two genres, since they present scientific issues using a journalistic language. Thus, the frequency of the passive voice in English popular science articles can be expected to be somewhere between these two percentages, i.e. 20%. The distribution of the frequency of the passive voice in the previous sample study represents exactly this proportion. If this prediction is confirmed, it will suggest that the translation of popular science articles from Anglophone sources tends to encourage the frequential copying of the passive voice in Modern Greek. In that case, Modern Greek being the Basic Code copied the frequency of the passive voice patterns from the Model Code, which is English. The copies first habitualised in the discourse of the translation and then spread into the general linguistic community and became monolingual copies.

5. Conclusion

Although the results are only preliminary, the importance of this corpus-based study lies in a number of factors. Firstly, it is one of the first diachronic corpus-based studies ever to be attempted within the field of translation studies and it raises collective awareness of how translation can encourage the dissemination of particular source language linguistic features. If this scholarly strand is to be consolidated, more research across a wider range of language pairs and linguistic features has to be conducted. Secondly, it is one of the first sustained corpus-based studies ever to be conducted in the Modern Greek context within the field of translation studies, which aims at analysing systematically and in depth the Modern Greek linguistic features of translated texts. Finally, this study combines all corpus-based methodologies, i.e. diachronic, synchronic, comparable and parallel, under one research aim: the investigation of translation as a language contact phenomenon. This is probably the most important aspect of this study since it stresses the numerous advantages of collaborative techniques and engages them in a mutually profitable dialogue.

6. References

- Amouzadeh, M., House, J. (2010): Translation and Language Contact: The case of English and Persian. *Languages in Contrast*, 10(1), pp. 54-75.
- Apostolou-Panara, A. (1991): English Loanwords in Modern Greek: An overview. *Terminologie et Traduction*, 1(1), pp. 45-60.
- Baker, M. (1995): Corpora in Translation Studies: An overview and some suggestions for future research. *Target*, 7(2), pp. 223-243.
- Baumgarten, N., House, J., Probst, J. (2004): English as a Lingua Franca in Covert Translation Processes. *The Translator*, 10(1), pp. 83-108.
- Baumgarten, N., Özçetin, D. (2008): Linguistic Variation through Language Contact in Translation. In E. Siemund & N. Kintana (Eds.), *Language Contact and Contact Languages*. Amsterdam: John Benjamins, pp. 293-316.
- Becher, V., House, J., Kranich, S. (2001): Convergence and Divergence of Communicative Norms through Language Contact in Translation. In K. Braunmüller & J. House (Eds.), *Convergence and Divergence in Language Contact Situations*. Amsterdam: John Benjamins, pp. 125-152.
- Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. (1999): *Longman Grammar of Spoken and Written English*. Harlow: Longman.
- Csató, É.Á. (2002): Karaim: A high-copying language. In M.C. Jones & E. Esch (Eds.), *Language Change: The interplay of internal, external and extra-linguistic factors*. Berlin: Mouton de Gruyter, pp. 315-327.
- Johanson, L. (1993): Code-Copying in Immigrant Turkish. In G. Extra & L. Verhoeven (Eds.), *Immigrant Languages in Europe*. Clevedon, Philadelphia and Adelaide: Multilingual Matters, pp. 197-221.
- Johanson, L. (1999): The Dynamics of Code-Copying in Language Encounters. In B. Brendemoen, E. Lanza & E. Ryen (Eds.), *Language Encounters across Time and Space*. Oslo: Novus Press, pp. 37-62.
- Johanson, L. (2002): *Structural Factors in Turkic language Contacts*. London: Curzon.
- Johanson, L. (2008): Remodelling Grammar: Copying, conventionalisation, grammaticalisation. In E. Siemund & N. Kintana (Eds.), *Language Contact and Contact Languages*. Amsterdam: John Benjamins, pp. 61-79.
- House, J. (2004): English as Lingua Franca and its Influence on Other European Languages. In J.M. Bravo (Ed.), *A New Spectrum of Translation Studies*. Valladolid: Universidad de Valladolid, pp. 49-62.
- House, J. (2008): Global English and the Destruction of Identity?. In P. Nikolaou & M.V. Kyritsi (Eds.), *Translating Selves: Experience and identity between languages and literatures*. London and New York: Continuum, pp. 87-107.
- Steiner, E. (2008): Empirical Studies of Translations as a Mode of Language Contact: 'Explicitness' of lexicogrammatical encoding as a relevant dimension. In E. Siemund & N. Kintana (Eds.), *Language Contact and Contact Languages*. Amsterdam: John Benjamins, pp. 317-346.
- Verschik, A. (2008): *Emerging Bilingual Speech: From Monolingualism to Code-Copying*. London: Continuum.
- Warburton, I. (1975): The Passive in English and Greek. *Foundations of Language*, 13(4), pp. 563-578.

A Comparable Wikipedia Corpus: From Wiki Syntax to POS Tagged XML

Noah Bubenhofer, Stefanie Haupt, Horst Schwinn

Institut für Deutsche Sprache IDS

Mannheim

E-mail: bubenhofer@ids-mannheim.de, st.haupt@gmail.com, schwinn@ids-mannheim.de

Abstract

To build a comparable Wikipedia corpus of German, French, Italian, Norwegian, Polish and Hungarian for contrastive grammar research, we used a set of XSLT stylesheets to transform the mediawiki annotations to XML. Furthermore, the data has been annotated with word class information using different taggers. The outcome is a corpus with rich meta data and linguistic annotation that can be used for multilingual research in various linguistic topics.

Keywords: Wikipedia, Comparable Corpus, Multilingual Corpus, POS-Tagging, XSLT

1. Background

The project EuroGr@mm¹ aims at describing German grammar from a multi-lingual perspective. Therefore, an international research team consisting of members from Germany, France, Italy, Norway, Poland and Hungary, collaborates in bringing in their respective language knowledge to a contrastive description of German. The grammatical topics that have been tackled so far are morphology, word classes, tense, word order and phrases. A corpus-based approach is used to compare the grammatical means of the languages in focus. But so far, no comparable corpus of the chosen languages was at the project's disposal. Of course, for all the languages big corpora are available, but they consist of different text types and are in different states of preparation regarding linguistic markup.

Hence we wanted to build our own corpus of comparable data in the different languages. The Wikipedia is a suitable source for building such a corpus. The disadvantage of the Wikipedia is its limitations regarding text types: The articles are (or are at least intended to be) very uniform in their linguistic structure. To overcome this problem we decided to include also the discussions of the articles in our corpus, which can broaden at least slightly the text type diversity.

In this paper we describe, how the Wikipedia was converted to an XML format and part-of-speech-tagged.

2. Wikipedia conversion to XCES

To be able to integrate the linguistic annotated version of the Wikipedia into our existing corpus repository, the data has to be in the XML format XCES². There are already some attempts to convert the Wikipedia to a corpus linguistic usable data source (Fuchs, 2010:136). But they offer either only the data of a specific language version of the Wikipedia in an XML format (Wikipedia XML Corpus, Denoyer & Gallinari, 2006; SW1, Atserias et al., 2008), the format isn't suitable for our needs (WikiPrep, Gabrilovich & Markovitch, 2006; WikIDF, Krizhanovsky, 2008; Java Wikipedia Library, Zesch et al. 2008) or the conversion tool does not work anymore with the current mediawiki engine (WikiXML Collection; Wiki2TEI, Desgraupes & Loiseau 2007). To have a lasting solution, the conversion routines need to be useable also in the future which would allow us to get from time to time a new version of the Wikipedia. Therefore we developed our own solution of XSLT transformations to get an XCES version of the data.

All Wikipedia articles and their discussions are available as mediawiki database dumps in XML (Extensible Markup Language, Bray et al., 1998). These database dumps contain different annotations. Metadata of articles display in XML while the articles display in mediawiki language. We convert these documents into XCES format using XSLT 2.0 transformations to ease research.

¹ See

<http://www.ids-mannheim.de/gra/eurogr@mm.html>.

² <http://www.xces.org/>

This process is divided into 2 sections:

- 1) The conversion from mediawiki language to XML
- 2) The conversion from the generated XML to XCES format

The mediawiki language consists of a variety of special signs for special annotations. E.g. to describe a level 2 header the line displays as text wrapped into two equal signs on each side, like this:

```
== head ==
```

Likewise lists display as a chain of hash or asterisk signs, according to the level, e.g. a level 3 list entry:

```
### list entry
```

During the first conversion we process the paragraphs according to their type and detect headers, lists, tables and usual paragraphs. We convert these signs into clean XML, so

```
== head ==
```

turns to

```
<head2>text</head2>
```

and

```
### list entry
```

turns to

```
<item level=3>list entry</item>.
```

Of course inside the paragraphs there may be text-highlighting markup. We access the paragraphs and convert these wikimedia annotations to XML, too. Here we follow a certain pattern to detect text-highlighting signs.

Still the document's hierarchy is flat. In the next step we add structure to the lists. We group the list items according to their level to highlight the structure. In a later step we group all articles into sections depending on the occurrence of head elements. Whenever we add structure we need to take care of possible errors in the mediawiki syntax.

Now the articles need to be transformed into the XCES structure. Here we sort the articles into alphanumerical sections. We transform the corpus and enrich every article with meta data. We provide a unique id for every article and discussion so that they can easily be referenced. Also the actual article text can be distinguished from the discussion part of the article, which is important because they are different text types. These conversion routines should work for all the language versions of the Wikipedia, but have so far only

been tested with the languages necessary for the project: German, French, Italian, Norwegian (Bokmål), Polish and Hungarian.

3. POS-Tagging

To enable searching for word class information in the corpus, the data needs being part-of-speech tagged. This task has not been finished yet, but preliminary tests have been done already. Not having any additional resources, we have to rely on ready to use taggers and cannot do any improvements or adjustments of the taggers.³ We are using the following taggers:

German TreeTagger (Schmid, 1994) with the available training library for German (STTS-Tagset, Schiller et al., 1995)

French TreeTagger with the available training library for French

Italian TreeTagger with the available training library for Italian

Polish TaKIPI (Piasecki, 2007), based on Morfeusz SLaT (Saloni et al., 2010)

Hungarian System developed by the Hungarian National Corpus team (Váradi, 2002), based on TnT (Brants, 2000)

Norwegian (Bokmål) Oslo-Bergen Tagger (Hagen et al., 2000)⁴

The input for the taggers are raw text files without any XML mark-up and containing only those parts of the Wikipedia, which need to be tagged. So all meta information is being ignored.

A Perl script is used to send the input data in manageable chunks to the tagger. The script also transfers the output of the tagger to a XML file that contains to each token the character position reference to the original data file. Because of the size of the Wikipedia, the tagging process is very time consuming. E.g. the XCES file of the German Wikipedia holds about 15.4 GB of data (785'791'766 tokens). The size of the stand-off file containing the linguistic mark-up produced by the

³ Nevertheless we get support of the developers of the taggers, which we greatly appreciate.

⁴ See <http://tekstlab.uio.no/obt-ny/english/history.html> for the newest developments of the tagger.

TreeTagger (POS information to each token) is about 157.9 GB. It took about 30 hours on a standard double core PC to process this file.

4. Corpus Query System

Our existing corpus management software COSMAS II⁵ is used as corpus query system. COSMAS II is currently used to manage the DeReKo (German Reference Corpus, see Kupietz et al., 2010), which contains about 4 billion tokens. Therefore COSMAS II is also able to cope with the Wikipedia data.

To be able to build from time to time new versions of our corpus based on the latest Wikipedia, we can rely on the same version controlling mechanisms as the DeReKo does.

For technical reasons, COSMAS II cannot handle UTF-8 encoding. Therefore the encoding of the XCES files have to be changed to ISO-8859-1 and characters outside this range converted to numeric character references referring to the Unicode code point.

At the end of this process, the Wikipedias in the XCES and the tagged format will be made publicly available to the scientific community.

5. Conclusion

While the Wikipedia is a often used and attractive source for various NLP and corpus linguistic tasks, it is not easy to get an enduring XML conversion routine which produces proper XML versions of the data. It was our attempt to find such a solution using XSLT stylesheets.

After the part-of-speech tagging of the six language versions of the Wikipedia (German, French, Italian, Polish, Hungarian, Norwegian) we are able to build a multilingual comparable corpus for contrastive grammar research in our project.

For future investigations, the advantage of a XML version of the Wikipedia is clearly visible: The XML structure holds all the meta information available in the mediawiki code and can therefore be used to differentiate findings of grammatical structures: Are there variants of specific constructions in different text types (lexicon entry vs. user discussion)? Or does the usage of the constructions depend on topic domains? And how do

these observations change in the light of inter-lingual comparisons?

6. References

- Atserias, J., Zaragoza, H., Ciaramita, M., Attardi, G. (2008): Semantically Annotated Snapshot of the English Wikipedia. In Proceedings of the Sixth International Language Resources and Evaluation (LREC 08), Marrakech, pp. 2313–2316.
- Brants, T. (2000): TnT – A Statistical Part-of-Speech Tagger. In Proceedings of the Sixth Conference on Applied Natural Language Processing (ANLP), Seattle, WA.
- Bray, T., Paoli, J., Sperberg-McQueen, C. M. (1998): Extensible Markup Language (XML) 1.0. W3C Recommendation
<<http://www.w3.org/TR/1998/REC-xml-19980210>>.
- Denoyer, L., Gallinari, P. (2006): The Wikipedia XML Corpus. In SIGIR Forum.
- Desgraupes, B., Loiseau, S. (2007): Wiki to TEI 1.0 project <<http://wiki2tei.sourceforge.net/>>.
- Fuchs, M. (2010): Aufbau eines linguistischen Korpus aus den Daten der englischen Wikipedia. In Semantic Approaches in Natural Language Processing. Proceedings of the Conference on Natural Language Processing 2010 (KONVENS 10), Saarbrücken: Universitätsverlag des Saarlandes, pp. 135–139.
- Gabrilovich, E., Markovitch, S. (2006): Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. In Proceedings of The 21st National Conference on Artificial Intelligence (AAAI), Boston, pp. 1301–1306.
- Hagen, K., Johannessen, J. B., Nøklestad, A. (2000): A Constraint-based Tagger for Norwegian. In 17th Scandinavian Conference of Linguistics, Lund, Odense: University of Southern Denmark, 19, pp. 31–48 (Odense Working Papers in Language and Communication).
- Krizhanovsky, A. A. (2008): Index wiki database: design and experiments. In CoRR abs/0808.1753.
- Kupietz, M., Belica, C., Keibel, H., Witt, A. (2010): The German Reference Corpus DeReKo: A primordial sample for linguistic research. In Proceedings of the 7th conference on International Language Resources

⁵ See <http://www.ids-mannheim.de/cosmas2/>.

- and Evaluation, Valletta, Malta: European Language Resources Association (ELRA), pp. 1848-1854.
- Piasecki, M. (2007): Polish Tagger TaKIPI: Rule Based Construction and Optimisation. In *Task Quarterly* 11(1-2), pp. 151-167.
- Saloni, Z., Gruszczyński, W., Woliński, M., Wołosz, R. (2010): *Analizator morfologiczny Morfeusz*
<<http://sgjpp.pl/morfeusz/>>.
- Schiller, A., Teufel, S., Thielen, C. (1995): *Guidelines für das Tagging deutscher Textcorpora mit STTS*. Universität Stuttgart, Institut für maschinelle Sprachverarbeitung; Universität Tübingen, Seminar für Sprachwissenschaft, Stuttgart
<<http://www.sfs.uni-tuebingen.de/Elwis/stts/stts.html>>.
- Schmid, H. (1994): *Probabilistic Part-of-Speech Tagging Using Decision Trees*
<<http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf>>.
- Váradi, T. (2002): *The Hungarian National Corpus*. In *Proceedings of the 3rd LREC Conference, Las Palmas, Spanyolország*, pp. 385-389
<<http://corpus.nytud.hu/mnsz>>.
- Zesch, T., Müller, C., Gurevych, I. (2008): *Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary*. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC 08), Marrakech*, pp. 1646-1652
<<http://www.lrec-conf.org/proceedings/lrec2008/>>.

A German Grammar for Generation in OpenCCG

Jean Vancoppenolle* Eric Tabbert* Gerlof Bouma+ Manfred Stede*

* Dept of Linguistics, University of Potsdam, + Dept of Swedish, University of Gothenburg
E-mail: *{vancoppenolle,tabbert,stede}@uni-potsdam.de +gerlof.bouma@gu.se

Abstract

We present a freely available CCG fragment for German that is being developed for natural language generation tasks in the domain of share price statistics. It is implemented in OpenCCG, an open source Java implementation of the computationally attractive CCG formalism. Since generation requires lexical categories to have semantic representations, so that possible realizations can be produced, the underlying grammar needs to define semantics. Hybrid Logic Dependency Semantics, a logic calculus especially suited for encoding linguistic meaning, is used to declare the semantics layer. To our knowledge, related work on German CCG development has not yet focused on the semantics layer. In terms of syntax, we concentrate on aspects of German as a partially free constituent order language. Special attention is paid to scrambling, where we employ CCG's type-changing mechanism in a manner that is somewhat unusual, but that allows us to a) minimize the amount of syntactic categories that are needed to model scrambling, compared to providing categories for all possible argument orders, and b) retain enough control to impose restrictions on scrambling.

Keywords: CCG, Generation, Scrambling, German

Introduction

“Der Kurs der Post ist vom 13. September bis 29. Oktober stetig gefallen und dann bis zum 15. November wieder leicht angestiegen.

Zwischen dem 13. und dem 29. September schwankte der Kurs leicht zwischen 15 und 16 Euro. Anschließend fiel er um mehr als die Hälfte ab und erreichte am 29. Oktober seinen Tiefststand bei 7 Euro. Bis zum 15. November stieg der Kurs nach einigen Schwankungen auf seinen Schlusswert von 10 Euro.”

Consider the graph depicting the development of a share price. Undoubtedly, a human could interpret the mathematical properties of that graph and quite easily describe this information in prose. He would probably produce a text more or less similar to the one presented above. In computational linguistics (or, more general, artificial intelligence), people attempt to go one step further and let the computer do that work for us. Basically, it will have to perform the same steps that a human would need to in order to accomplish this task:

determine the mathematical properties of interest and generate a text that is faithful to the input and easy to read. The present paper addresses the latter sub task – i.e., the text generation.

Our goal is to develop a freely available fragment of a German grammar in OpenCCG that is suitable for natural language generation tasks in the domain of share prices. Related work on German in OpenCCG includes Hockenmaier (2006) and Hockenmaier and Young (2008), who employ grammar induction algorithms to induce CCG grammars automatically from treebanks (e.g. TiGERCorpus). To our knowledge, however, very little resources are actually freely available. In particular, the coverage of a part of the semantic layer is a novel contribution of the grammar that we present here.

1. CCG

CCG (Combinatory Categorical Grammar, Steedman 2000, Steedman & Baldridge 2011) is a lexicalized grammar formalism in which all constituents, lexical

ones included, are assigned a syntactic category that describes its combinatory possibilities. These categories may be atomic or complex. Complex categories are functions from one category into another, with specification of the relative position of the function and its argument. For instance, the notation $s \backslash np$ describes a complex category that can be combined with an np on its left (direction of the slash) to yield an s . Category combination always applies to adjacent constituents and is governed by a set of combinatory rules, of which the simplest is function application. In the example in Fig. 1, we build a sentence (category s) around a transitive verb ($(s \backslash np) / np$). There are two versions of function application used in the derivation: backward ($<$) and forward ($>$), depending on which constituent is the argument and which is the function. An overview of other derivation rules is given in Table 1.

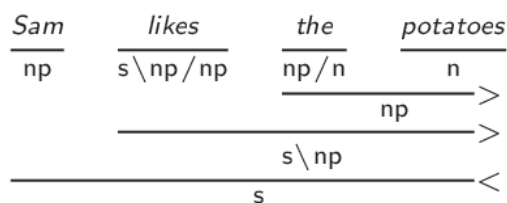


Figure 1: A basic CCG derivation.

The atomic categories in CCG come from a very restricted set. They may be enriched with features to handle case, agreement, clause types, etc. In addition, a grammar writer may choose to handle language-specific phenomena with unary type-changing rules. Finally, the grammar presented uses *multi-modal* CCG (henceforth MMCCG), which gives extended lexical control over derivation possibilities by adding modalities to the slashes in complex categories (see Baldridge 2002; Steedman & Baldridge 2011, for introduction and overview).

In its basic form, CCG has mildly context-sensitive generative power and is thus able to account for non-context-free natural language phenomena (Steedman, 2000). Its attractiveness is due to the linguistic expressiveness on the one hand and the fact that it is efficiently parsable in theory (Shanker & Weir, 1990), as well as in practice (Clark & Curran, 2007).

Function application				
$(>)$	α / β	β	\Rightarrow	α
$(<)$	β	$\alpha \backslash \beta$	\Rightarrow	α
Function composition				
$(> \mathbf{B})$	α / β	β / γ	\Rightarrow	α / γ
$(< \mathbf{B})$	$\beta \backslash \gamma$	$\alpha \backslash \beta$	\Rightarrow	$\alpha \backslash \gamma$
Crossed function composition				
$(> \mathbf{B}_\times)$	α / β	$\beta \backslash \gamma$	\Rightarrow	$\alpha \backslash \gamma$
$(< \mathbf{B}_\times)$	β / γ	$\alpha \backslash \beta$	\Rightarrow	α / γ
Type raising				
$(> \mathbf{T})$	α		\Rightarrow	$\beta / (\beta \backslash \alpha)$
$(< \mathbf{T})$	α		\Rightarrow	$\beta \backslash (\beta / \alpha)$

Table 1: Basic combinatory rules.

1.1. OpenCCG

OpenCCG¹ is an open source Java implementation of an MMCCG parser and generator. On the semantic side, it implements HLDS (Hybrid Logic Dependency Semantics, Blackburn, 2000; Baldridge & Kruijff, 2002), an extended modal logic calculus especially suited for encoding linguistic meaning. Below is an example sentence and its corresponding HLDS formula:

- 1) Der Kurs erreicht seinen Höchststand.
 the share-price reaches its peak

$$\begin{aligned}
 & @_{e:\text{achievement}} (\mathbf{reach} \wedge \\
 & \langle \mathbf{ARG1} \rangle (r:\text{sem-obj} \wedge \mathbf{rate} \wedge \\
 & \quad \langle \mathbf{DEF} \rangle + \wedge \\
 & \quad \langle \mathbf{NUM} \rangle \text{sg} \wedge \\
 & \langle \mathbf{ARG2} \rangle (p:\text{sem-obj} \wedge \mathbf{peak} \wedge \\
 & \quad \langle \mathbf{NUM} \rangle \text{sg} \wedge \\
 & \quad \langle \mathbf{OWNER} \rangle (i:\text{sem-obj} \wedge \text{pro3m})))
 \end{aligned}$$

Figure 2: HLDS formula of 1).

In HLDS, the meaning of (1) is represented by means of dependency relations between discourse referents, such as the OWNER relation between the referents **peaks** and **its** (pro3m) in Fig. 2. Nominals uniquely identify discourse referents and provide a means to reference them at any point in a formula. In Fig. 2, the nominals e , r , p , and i identify the referents for the event of **reaching**, **rate**, **peak** and **its**, respectively. Nominals can be augmented to restrict the semantic type of possible discourse referents with respect to an underlying ontology, for instance to animate or inanimate entities, or, as in Fig. 2, to a more general notion like *semantic*

¹<http://openccg.sourceforge.net/>

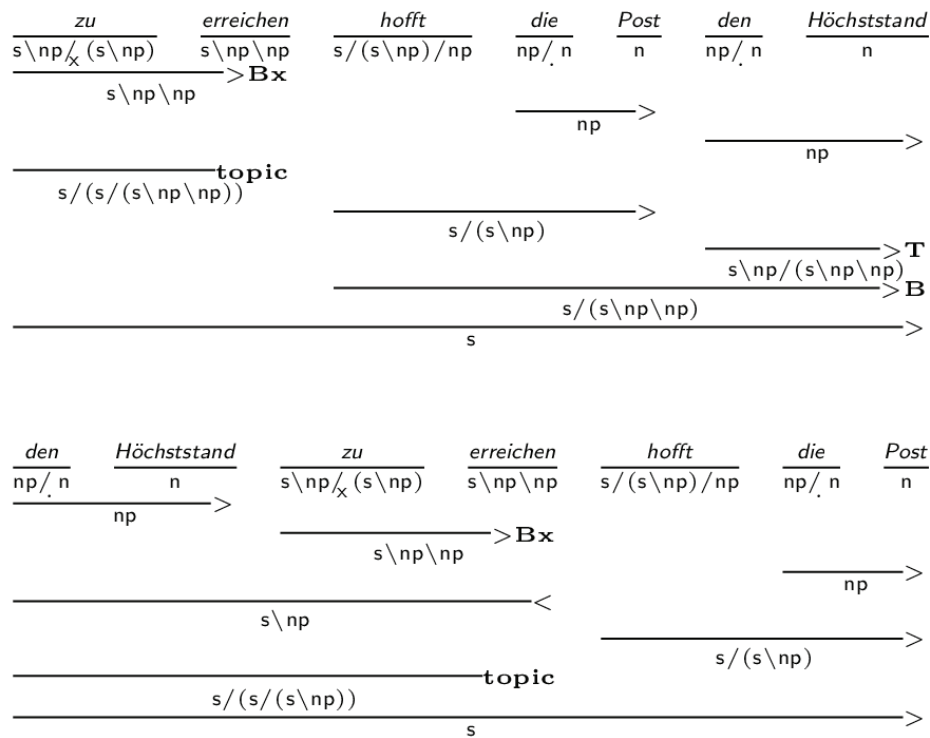


Figure 3: NP fronting

objects. Finally, the satisfaction operator $@$ states that the formula p in $@_i p$ holds at world i .

OpenCCG implements a flexible surface realizer that when given a logical form (LF) like in Fig. 1 returns one or more realizations of it, based on the underlying grammar. Both the number of realizations and their surface forms depend on how much information a LF specifies, thereby allowing to either enhance or restrain non-determinism of the realization process. For example, given that the LF in Fig. 2 does not specify which of the two arguments is fronted, the following two surface forms are possible:

- 2) Der Kurs erreicht seinen Höchststand.
the share-price reaches its peak
- 3) Seinen Höchststand erreicht der Kurs.
Its peak reaches the share-price.

'The share-price reaches its peak'

2. Coverage

Our current work focuses on different aspects of German as a partially free constituent order language, including basic constituent order and scrambling in particular, but also on complex nominal phrases, clausal subordination, and coordination. In the next two

sections, we first give a brief overview of how topicalization is modeled in our grammar, followed by an approach to scrambling that we are currently investigating and that, as far as we know, is new in CCG.

2.1. Topicalization

The finite verb can occupy three different positions that depend on the clause type and determine the sentence mood: matrix clauses are either verb-initial (declarative or yes/no-interrogative), or verb-second (declarative or wh-interrogative), and subordinate clauses are always verb-final (declarative or interrogative).

Following Steedman (2000), Hockenmaier (2006) and Hockenmaier and Young (2008), we implemented a topicalization rule that systematically derives verb-second order from verb-initial order by fronting an argument of the verb, e.g. an NP, a PP, or a clause. This also covers partial fronting (see Fig. 3 for examples):

- 4) $T \Rightarrow s_{v2} / (s_{v1} / T)$, $T = \{ np, pp, s_{lo-inf} \setminus np, \dots \}$
- Sentence modifiers (e.g. *heute* in *heute fällt der Kurs* 'today, the share price is falling') are analyzed as s_{v2} / s_{v1} and can thus form verb-second clauses on their own.

2.2. Scrambling

Much of the constituent order freedom in German is due to the fact that it allows for permutation of verbal arguments within a clause (local scrambling, 5) and 'extraction' of arguments of an arbitrarily deeply embedded infinite clause (long-distance scrambling, 6):

- 5) dass [dem Unternehmen]₂ [das Richtige]₃
 that the enterprise the right-thing
 [der Berater]₁ empfiehlt.
 the counselor advises
 'that the counselor advises the enterprise the right thing'
- 6) dass [dem Unternehmen]₂ [das Richtige]₃
 that the enterprise the right-thing
 [der Berater]₁ [__ zu empfehlen hofft].
 the counselor to advise hopes
 'that the counselor hopes to advise the enterprise the right thing'

Different proposals have been made in MMCCG to account for constituent order freedom in general. To our knowledge, the two most common approaches are to provide separate categories for each possible order (Hockenmaier, 2006; Hockenmaier & Young, 2008) or to allow lexical underspecification of argument order through multi-sets (Hoffman, 1992; Steedman & Baldridge, 2003).

We are investigating an approach to local scrambling that aims at combining the advantages of both methods, namely having fine-grained control over argument permutation on the one hand, and requiring as few categories as possible on the other. It is based on a set of type-changing rules that change categories 'on the fly'. (7) shows a simplified rule that allows to derive plural NPs from plural nouns, reflecting the optionality of determiners in German plural NPs (e.g. *sie isst Kartoffeln* 'she eats potatoes'):

$$7) \quad n_{pl} \Rightarrow np_{pl}$$

Type-changing rules can also be used to swap two consecutive argument NPs, (*i* and *j* denote indexes):

$$8) \quad s/np_{(i)+base}/np_{(j)-pron} \Rightarrow s/np_{(j)}/np_{(i)-base}$$

$$9) \quad s\$ \backslash np_{(i)-pron} \backslash np_{(j)+base, -pron} \Rightarrow s\$ \backslash np_{(j)-base, -pron} \backslash np_{(i)}$$

This essentially emulates the behavior of multi-sets and at the same time reduces the number of categories to a minimum, thereby enhancing the maintainability of the grammar. The advantage over multi-sets is that

restrictions on scrambling can be formulated straightforwardly, such as that full NPs should not scramble over pronouns (i.e. NPs having the *-pron(oun)* feature) (see Uszkoreit (1987) for an overview of scrambling regularities in German).

Rules like (8) and (9) require special caution, though. Type-changing rules are supposed to actually *change* the type of the argument category as they could otherwise apply over and over again, causing an infinite recursion. This is where the $\pm base$ feature comes into play. It indicates whether an NP occupies its base position or has already been scrambled, restricting the application of (8) and (9) to the former case and thereby preventing infinite recursion. The so-called *dollar variable* \$ in (9) ranges over complex categories that have the same functor (here: *s*), such as $s \backslash np$. It is not crucial to our scrambling rules but generalizes (9) to apply to both transitive and ditransitive verbs.

Four more rules are sufficient to capture all possible local permutations and also some of the long-distance permutations, as the one in (6).

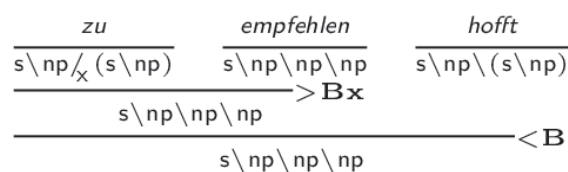


Figure 4: Parse of an infinite clause.

The derivation in Fig. 4 contains the derivation of the complex verb cluster of example (6). The composed category $s \backslash np \backslash np \backslash np$ corresponds to the one of an ordinary ditransitive verb, so although (6) is an instance of long-distance scrambling, it can be derived by means of our local scrambling rules (8) and (9).

3. Lexicon

The grammar is intended for use in the domain of the stock market, thus providing the means to describe the development of share prices. Since the expansion and proper implementation of a lexical database is a full-fledged task of its own and the focus of our current work is to extend the grammar, our current lexicon is still quite limited in its scope.

At a later point one might consider to make use of the

CCGbank lexicon (Hockenmaier, 2006).

3.1. Nouns

Our lexicon currently contains approximately 125 nouns. For the different inflectional paradigms we made use of inflection tables presented on the free online service [canoo.net](http://www.canoo.net).² For each of these paradigms we wrote an 'expansion'. OpenCCG's expansions provide a means to define inflectional paradigms as an applicable rule and link lexical information to them, so that OpenCCG generates the different tokens of a word and its syntactic and semantic properties as interpretable lexical entries. Thus a typical noun entry is a one-liner like this:

```
10) #Höchststand
    noun_infl_1(Höchststand, Höchstständ, masc
    peak, graph_point_definition)
```

The first two arguments contain the singular and plural stem, to which the inflection endings will be attached by the expansion. The following arguments are gender (for agreement), a predicate (as semantic reference) and a semantic type from the ontology. While seemingly plain English, these semantic predicates should be thought of as grossly simplified meta language, which guarantees a unique and unambiguous semantic representation.

3.2. Verbs

For the verbs we followed a similar approach, with three expansions. The first two actually cover the same inflection paradigm, with the difference that for verbs ending in *-ern* like *klettern* (to climb) we duplicated the paradigm and made slight adjustments to circumvent the concatenation of the word stem *kletter* and certain inflectional morphs like *-en* to ungrammatical forms like **(wir) kletteren* (instead of *klettern*). The third expansion covers several modal verbs like *können* (to can) or *müssen* (to have to).

Each of those rules sets the features of the respective inflection (e.g. fin, 1st, sg, pres) and those for past tense. Sample entries:

- 11) regular-vv(schwanken, schwank, schwankte, fluctuate)
- 12) regular-vv-ern(klettern, kletter, kletterte, climb)

²<http://www.canoo.net/>

4. Generation

We would like to conclude with a brief outline of how our grammar fits into the generation scenario presented in the introduction.

The idea is to generate text automatically from share price graphs, i.e., from collections of data points. Graphs are analyzed in terms of different mathematical properties (e.g. extremes and inflection points). These properties, together with user-provided realization parameters that allow fine-grained control over the 'specificity' of LFs (and thus over the number of surface realizations), are input to static LF templates. The filled LF templates are then fed to the OpenCCG realizer where our grammar is used to compute the appropriate surface forms. In the last step, orthographic post-processing, the surface forms are normalized with respect to language-specific orthographic standards (e.g. number or date formats, etc.). The figure below illustrates this procedure:

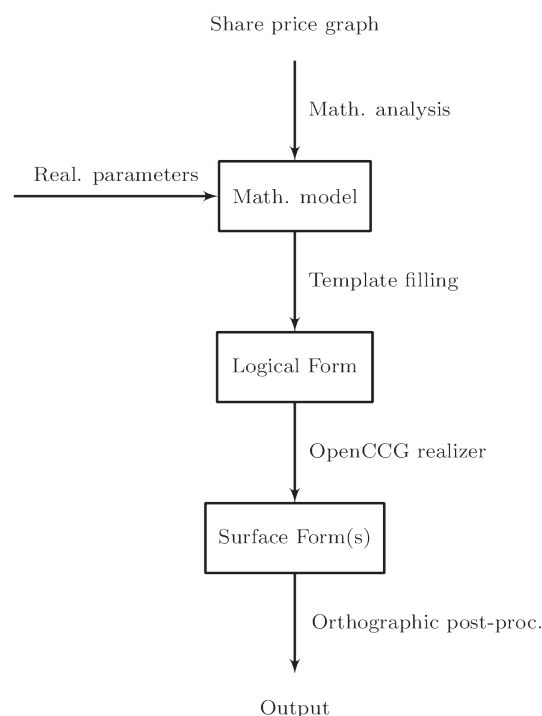


Figure 5: Procedure of the generation process.

5. Summary

We have presented a freely available CCG fragment of a generation grammar for German that is equipped with a semantic layer implemented in Hybrid Logic

Dependency Semantics. In terms of syntax, we have focused on aspects of German as a partially free constituent order language and investigated an approach to scrambling by employing OpenCCG's type-changing rules in a somewhat unconventional manner. In doing so, we aimed at minimizing the amount of categories needed to allow different argument orders while retaining a certain degree of flexibility regarding argument order restrictions. Future work will concentrate more on the lexicon, for instance by refining and extending our expansions for inflectional paradigms of various word classes. We also hope to use OpenCCG's interesting regular expression facilities for derivational morphology.

Our grammar can be downloaded from www.ling.uni-potsdam.de/~stede/AGacl/ressourcen/GerGenGram.

6. Acknowledgements

We would like to thank the other participants of the course *Automatische Textgenerierung* (Winter 2010/11) at the University of Potsdam, and also the GSCL reviewers for their comments.

7. References

- Baldrige, J. (2002): Lexically Specified Derivational Control in Combinatory Categorical Grammar. PhD thesis, School of Informatics, University of Edinburgh.
- Baldrige, J., Kruijff, G.-J.M. (2002): Coupling CCG and Hybrid Logic Dependency Semantics. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002).
- Baldrige, J., Kruijff, G.-J.M. (2003): Multi-Modal Combinatory Categorical Grammar. In Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003).
- Blackburn, P. (1993): Modal Logic and Attribute Value Structures. In M. de Rijke, editors, *Diamonds and Defaults*, Synthese Language Library, pp. 19–65, Kluwer Academic Publishers, Dordrecht, 1993.
- Blackburn, P. (2000): Representation, Reasoning, and Relational Structures: a Hybrid Logic Manifesto. *Logic Journal of the IGPL*, 8(3), pp. 339-625.
- Bozsahin, C., Kruijff, G.-J.M., White, M. (2008): Specifying Grammars for OpenCCG: A Rough Guide. <http://openccg.sourceforge.net/>
- Clark, S., Curran, S. (2007): Wide-coverage Efficient Statistical Parsing with CCG and Log-linear Models. *Computational Linguistics*, 33(4), pp. 493-552.
- Drach, E. (1937): *Grundgedanken der deutschen Satzlehre*. Diesterweg.
- Hockenmaier, J. (2006): Creating a CCGbank and a wide-coverage CCG lexicon for German. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL.
- Hockenmaier, J., Young, P. (2008): Non-local scrambling: the equivalence of TAG and CCG revisited. Proceedings of The Ninth International Workshop on Tree Adjoining Grammars and Related Formalisms, pp. 41–48, Tübingen, Germany.
- Hoffman, B. (1992): A CCG Approach to Free Word Order Languages. Proceedings of the 30th Annual Meeting of ACL, pp. 300-302.
- Müller, S. (2010): *Grammatiktheorie*. Stauffenburg Verlag.
- Steedman, M. (2000): *The Syntactic Process*. MIT Press.
- Steedman, M., Baldrige, J. (2011): Combinatory Categorical Grammar. In Borsley and Börjars (eds), *Non-transformational Syntax: Formal and explicit models of grammar*, Wiley-Blackwell.
- Uszkoreit, H. (1987): *Word Order and Constituent Structure in German*. CSLI.
- Vijay-Shanker, K., Weir, D.J. (1990): Polynomial Time Parsing of Combinatory Categorical Grammars. Proceedings of the 28th Annual Meeting of Computational Linguistics, pp. 1-8, Pittsburgh, PA, June 1990.
- White, M.: *OpenCCG Realizer Manual*. Documentation of the OpenCCG Realizer.
- White, M. (2004): Efficient Realization of Coordinate Structures in Combinatory Categorical Grammar. *Research on Language & Computation*, 4(1), pp. 39-75.
- White, M., Rajkumar R., Martin, S. (2007): Towards Broad Coverage Surface Realization with CCG. In Proceedings of the Workshop on Using Corpora for NLG: Language Generation and Machine Translation (UCNLG+MT).

Multilingualism in Ancient Texts: Language Detection by Example of Old High German and Old Saxon

Zahurul Islam¹, Roland Mittmann², Alexander Mehler¹

¹AG Texttechnology, Institut für Informatik, Goethe-Universität Frankfurt

²Institut für Empirische Sprachwissenschaft, Goethe-Universität Frankfurt

E-mail: zahurul, mittmann, mehler@em.uni-frankfurt.de

Abstract

In this paper, we present an approach to language detection in streams of multilingual ancient texts. We introduce a supervised classifier that detects, amongst others, Old High German (OHG) and Old Saxon (OS). We evaluate our model by means of three experiments that show that language detection is possible even for dead languages. Finally, we present an experiment in unsupervised language detection as a tertium comparationis for our supervised classifier.

Keywords: Language identification, Ancient text, n-gram, classification, clustering

1. Introduction

With the rise of the web, we face more and more on-line resources that mix different languages. This multilingualism of textual resources poses a challenge for many tasks in Natural Language Processing (NLP). As a consequence, Language Identification (LI) is now an indispensable step of preprocessing for many NLP applications. This includes machine translation, automatic speech recognition, text-to-speech systems as well as text classification in multilingual scenarios.

Obviously, LI is a well-established field of application of NLP. However, if one looks at documents that were written in low-density languages or documents that mix several dead languages, adequate models of language detection are rarely found. In any event, ancient languages are becoming more and more central in approach to computational Humanities, historical semantics and studies on language evolution. Thus, we are in need of models of language detection of dead languages.

In this paper, we present such a model. We introduce a supervised classifier that detects amongst others, OHG and OS. To do so, we extend the model of Waltinger and Mehler (2009) so that it also accounts for dead languages. For any segment of the logical document structure of a text, our task is to detect the corresponding language in which it was written. This detection at the segment level rather than at the level of whole texts allows us to make explicit the multilingualism of ancient documents starting from the level of words via the level of sentences up

to the level of texts. As a result, language-specific preprocessing tools can be used in such a way that they focus on those segments that provide relevant input for them. In this way, our approach is a first step towards building a preprocessor of multilingual ancient texts.

The paper is organized as follows: Section 3 describes the corpus of texts that we have used for our experiments. Section 4 briefly introduces our approach to supervised language detection, which is evaluated in Section 5. Section 6 describes unsupervised language classifier. Finally, a conclusion is given in Section 7.

2. Related Work

As we present a model of n-gram-based language detection, we briefly discuss work in this area.

Cavnar and Trenkle (1994) describe a system of *n-gram* based text and language categorization. Basically, they calculate *n-gram* profiles for each target category. Categorization occurs by means of measuring the distances of the profiles of input documents with those of the target categories. Regarding language classification, the accuracy of this system is 99.8%.

The same technique has been applied by Mansur et al. (2006) for text categorization. In this approach, a corpus of newspaper articles has been used as input to categorization. Mansur et al. (2006) show that *n-grams* of length 2 and 3 are most efficiently used as features for text categorization.

Kanaris and Stamatatos (2007) used character level *n-grams* to categorize web genres. Their approach is based on *n-grams* of characters of variable length that were combined with information about most frequently used HTML-tags.

Note that the language detection toolkit of Google translator may also be considered as a related work. However, at present, this system does not recognize sentences in OHG. We have tested 10 example sentences. The toolkit categorized only one of these input sentences as modern German; other sentences were categorized as different languages (e.g., Italian, French, English and Danish).

These approaches basically explore *n-grams* as features of language classification. However, they do that for modern languages. In this paper we present an approach that fills the gap of ancient language detection.

3. The Corpus

The corpus used consists of 160 complete texts in six diachronically and diatopically diverging stages of the German language plus the OS glosses, all collected from the TITUS¹ online database. High German is the language variety spoken historically south of a bundle of isogloss lines stretching from Aachen through Düsseldorf, Siegen, Kassel and Halle to Frankfurt (Oder) and has developed into what today constitutes standard German. Low German was spoken historically north of this line but has undergone a decline in native speakers to the point that it is now considered a regional vernacular of and alongside standard German, despite the fact that Low German and High German were once distinct languages. Table 1 shows the historical and geographical varieties of older German.

New discoveries of texts in the various historical forms and varieties of German are being made continually. Due to the steadily increasing number of transmitted texts from throughout the history of the German language, the focus of the TITUS corpus is on the older stages: it comprises the whole OHG corpus (apart from the glosses) as well as the entire OS corpus, including one mixed OHG and OS text. Of the younger language stages only unrepresentative amounts of texts are contained: several

dozen Middle High German (MHG) texts, some Middle Low German (MLG) texts, a sample of Early New High German (ENHG) texts and one mixed ENHG and Early New Low German (ENLG) text all of them varying considerably in length, from a few words to several tens of thousands per text.

Language Stage	Period of Time
OHG	ca. 750 – 1050 CE
MHG	ca. 1050 – 1350 CE
ENHG	ca. 1350 – 1650 CE
OS	ca. 800 – 1200 CE
MLG	ca. 1200 – 1600 CE
ENLG	ca. 1600 – 1750 CE

Table 1: Historical and geographical varieties

Among the oldest transmissions are interlinear translations of Latin texts, but also free translations and adaptations as well as mixed German-Latin texts. Translations consist mainly of religious literature, prayers, hymns, but also of ancient authors and scientific writings. These are later on complemented by epic and lyrical poetry (minnesongs), prose literature, sermons and other religious works, specialist books, chronicles, legislative texts and philosophical treatises. The latest texts of the corpus cover a biographical and a historical work, a collection of legal texts for a prince, an experimental re-narration of a parodistic novel as well as the German parts of two bilingual texts, a High German-Old Prussian enchiridion and a mixed High and Low German textbook for learning Russian.

Language Stage	#Texts	#Tokens
OHG	101	437,390
MHG	31	1,776,900
ENHG	6	237,432
OS	17	62,706
MLG	4	133,584
ENLG	1	26,679
Total	160	2,674,691

Table 2: Composition of the corpus

The corpus was generated by entering plain text, either completely by hand or by scanning, performing OCR recognition and correcting it manually. The texts were then indexed and provided with information on languages and subdivisions using the

¹Thesaurus of Indo-European Text and Language Materials – see <http://titus.uni-frankfurt.de>

Word-Cruncher² software developed by Brigham Young University in Provo, Utah. They were then converted into HTML format and were simultaneously conveyed into several SQL database files, classified by the words' language family, to enable the set-up of an on-line search.

4. Approach

In this section, we describe our language detection approach. We start with describing how we prepared the corpus from TITUS database to get input for our classifier (Section 4.1), introduce our model (Section 4.2) and describe its system design (Section 4.3).

4.1. Corpus Preparation

The training and test corpora that we used in our experiments were extracted from the database dump of TITUS (see Section 3). Each word in this extraction has been annotated with its corresponding language name (example: German), sub-language name (example: Old High German), document number, division number and its position within the underlying HTML corpus files. TITUS only annotates the boundaries of divisions so that any division may contain one or more sentences. For any sub-language (i.e., OHG, OS, MHG, MLG, ENLG and ENHG), we extracted text as reported in Table 2.

4.2. Language Detection Toolkit

Our approach for language detection is based on Cavnar and Trenkle (1994) and Waltinger and Mehler (2009). As in these studies, for every target category we learn an ordered list of most frequent *n*-grams that occur in descending order. The same is done for any input text so that categorization is done by measuring the distance between *n*-gram profiles of the target categories and the *n*-gram profiles of the test data.

The idea behind this approach is that the more similar two texts are, the more they share features that are equally ordered.

In general, classification is done by using a range of corpus features as are listed in Waltinger and Mehler (2009). Predefined information is extracted from the corpus to build sub-models based on those features. Each sub-model consists of a ranked frequency distribution of subset of corpus features. Corresponding *n*-gram information are extracted for $n = 1$ to 5. Each *n*-gram gets its

own frequency counter. The normalized frequency distribution of relevant features is calculated according to

$$\widehat{f}_{ij} = \frac{f_{ij}}{\max_{a_k \in L(D_j)} f_{kj}} \in (0,1]$$

\widehat{f}_{ij} is the frequency of feature a_i in D_j , divided by the frequency of the most frequent feature a_k in the feature representation $L(D_j)$ of document D_j (see Waltinger & Mehler, 2009). To categorize any document D_m , it is compared to each category C_n using the distance d of the rank r_{mk} of feature a_k in the sub-model of D_m with the corresponding rank of that feature in the representation of C_n :

$$d(D_m, C_n, a_k) = \begin{cases} |r_{mk} - r_{nk}| & a_k \in L(D_m) \wedge a_k \in L(C_n) \\ \max_{a_k \notin L(D_m) \vee a_k \notin L(C_n)} & \end{cases}$$

$d(D_m, C_n, a_k)$ equals *max* if feature a_k does not belong to the representation of D_m or to the one of category C_n . *max* is the maximum that the term $|r_{mk} - r_{nk}|$ can assume.

4.3. System Design

The language detection toolkit (Waltinger & Mehler, 2009) is used to build training models. It creates several *n*-gram models for each language which are used by the same tool for detection. Figure 1 shows the basic system diagram.

To detect the language of a document, the toolkit traverses the document sentence by sentence and detects the language of each sentence. If the document is homogeneous, (i.e., all sentences belong to the same language), then sentence level detection suffices to trigger other tools for further processing (e.g., Parsing, Tagging and Morpho-syntactic analysis) of that document, where language detection is necessary for preprocessing.

In the case that the sentences belong to more than one language (i.e., in the case of a heterogeneous document), the toolkit process the document word by word and detect the language of each token separately. This step is necessary in the case of multilingual documents that contain words from different languages are in single sentences. For example: in a scenario of lemmatization or morphological analysis of a multilingual document, it is necessary to trigger language specific tools to avoid errors. Just one tool needs to be triggered for further processing of a homogeneous document, whereas for a heterogeneous

² <http://wordcruncher.byu.edu>

document the same kind of tool has to be triggered based on the word level.

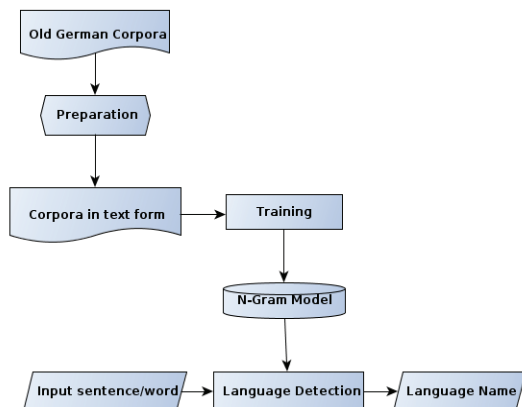


Figure 1: Basic system diagram

Language	Accuracy	F-score
OHG	100%	1
OS	100%	1

Table 3: Sentence level evaluation

5. Evaluation

In order to evaluate the language detection system, we extracted 200 sentences from the OHG corpus and 200 sentences from the OS corpus. These evaluation sets had not been used for training. There are many evaluation metrics used to evaluate NLP tools, we decided to use Accuracy and F-score (Hotho et al., 2005). Table 3 shows the evaluation result of the sentence level language detection, where we obtained 100% accuracy for both test sets. Table 4 shows the evaluation result of the word level language detection. 153 out of 1,259 words in the OHG test set were detected as OS and 33 out of 799 words in the OS test set were classified as OHG. The accuracy of the test set was 79.95% and 91.36% respectively. The evaluation result shows that the OHG test set might contain words from other languages, which is basically true. Petrova et al. (2009) show that the OHG diachronic corpus contains many Latin words. The evaluation becomes more effective when the result is compared with a gold-standard reference set. We came up with a list of 1,548 words (818 types) where each token is manually annotated with the name of the language to which the word belongs. Of 1,548 words, 564 overlapped with training data. Each word in the gold-standard test set

is detected by the toolkit and the result was compared with the reference set. We obtained 91.66% accuracy and an F-score of 95%.

Language	Accuracy	F-score
OHG	79.95%	0.88
OS	91.36%	0.96

Table 4: Word level evaluation

6. Unsupervised Language Classification

In addition to the classifier presented above, we experimented with an unsupervised classifier. The reason was twofold: one the one hand, we wanted to detect the added-value of an unsupervised classifier in comparison to its supervised counterpart. On the other hand, we aimed at extending the number of target languages to be detected. We collected several documents per target language, where each document was represented by a separate feature vector that counts the frequencies of a selected set of lexical features. As target classes we referred to six languages (whose cardinalities are displayed in Table 6): Early New High German (ENHG), Early New Low German (ENLG), Middle High German (MHG), Middle Low German (MLG), Old High German (OHG), and Old Saxon (OS). In order to implement an unsupervised language classifier, we followed the approach described in Mehler (2008). That is, we performed a hierarchical agglomerative clustering together with a subsequent partitioning that is informed about the number of target classes. However, other than in Mehler (2008), we did not perform a genetic search of the best performing subset of features as in the present case their number is too large. Table 5 shows the classification results. Performing a hierarchical-agglomerative clustering based on the cosine measure as the operative measure of object distance, we get an F-score of around 78%. This is a promising result as it is accompanied by a remarkable high accuracy. However, as seen in Table 4, the target classes perform quite differently: while we fail to separate ENHG and ENLG (certainly due to the small number of respective target documents), we separate MHG, MLG, OHG and OS to a reasonable degree. In this sense, the unsupervised classifier makes expectable even higher F-score supposed that we look for better performing features in conjunction with well-trained supervised classifiers. At least, the present study provides a

baseline that can be referred to in future experiments in this area.

Approach	Object Distance	F-Score	Accuracy
hierarchical/complete	cosine	0.78098	0.91134
hierarchical/weighted	cosine	0.69325	0.86934
hierarchical/average	cosine	0.61763	0.8307
hierarchical/single	cosine	0.56675	0.7926

Table 5: *F-scores* and accuracies of classifying historical language data in a semi semi-supervised environment

Language	#Texts	F-score	Recall	Precision
ENHG	6	0	0	0
ENLG	1	0	0	0
MHG	31	0.895	1	0.810
MLG	4	0.8	0.8	0.8
OHG	101	0.762	0.615	1
OS	17	0.889	0.889	0.889

Table 6: F-scores, recalls, and precisions differentiated by the target classes

7. Conclusion

Language detection plays an important role in processing multilingual documents. This is true especially for ancient documents that, due to their genealogy, mix different ancient languages. Here, documents need to be annotated in such a way that preprocessors can activate language specific routines on a segment by segment basis. In this paper, we presented an extended version of the language detection toolkit that allows us decide when to activate language specific analyses. Notwithstanding the low density of training material that is available for these languages, our classification results are very promising. At this point one may object that corpora of ancient texts are essentially so small that language detection can be done by hand. Actually, this objection is wrong if one considers corpora like the *Patrologia Latina* (Jordan, 1995), which mixes classical Latin with medieval Latin as well as with French and other Romance languages that are used in commentaries. From the size of this corpus alone (more than 120 million tokens), it is evident that a reliable means of automatizing segment-based language detection needs to be a viable option. We also described an unsupervised language detector that is evaluated simultaneously by means of OHG, OS, MHG, MLG,

ENLG and ENHG. Although this unsupervised classifier does not outperform its supervised counterpart, it shows that language detection in text streams of ancient languages comes into reach.

8. Acknowledgements

We would like to thank Ulli Waltinger, Armin Hoenen, Andy Lücking and Timothy Price for fruitful suggestions and comments. We also acknowledge funding by the LOEWE Digital-Humanities project in the Goethe-Universität Frankfurt.

9. References

- Cavnar, W. B., Trenkle, J. M. (1994): Ngram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pp. 161–175.
- Hotho, A., Nürnberger, A., Paaß, G. (2005): A Brief Survey of Text Mining. *Journal for Language Technology and Computational Linguistics (JLCL)*, 20(1), pp. 19–62.
- Jordan, M. D., editor (1995): *Patrologia Latina* database. Chadwyck-Healey, Cambridge.
- Kanaris, I., Stamatatos, E. (2007): Webpage genre identification using variable-length character n-grams. In

- Proc. of the 19th IEEE Int. Conf. on Tools with Artificial Intelligence (ICTAI'07), Washington, DC, USA. IEEE Computer Society.
- Mansur, M., UzZaman, N., Khan, M. (2006): Analysis of n-gram based text categorization for Bangla in a newspaper corpus. In Proceedings of the 9th International Conference on Computer and Information Technology (ICCIT 2006).
- Mehler, A. (2008): Structural similarities of complex networks: A computational model by example of wiki graphs. *Applied Artificial Intelligence*, 22(7&8), pp. 619–683.
- Petrova, S., Solf, M., Ritz, J., Chiarcos, C, Zeldes, A. (2009): Building and using a richly annotated interlinear diachronic corpus: The case of old high german tatian. *Journal of Traitement automatique des langues (TAL)*, 50(2), pp. 47–71.
- Waltinger, U., Mehler, A. (2009): The feature difference coefficient: Classification by means of feature distributions. In Proceedings of the Conference on Text Mining Services (TMS 2009), *Leipziger Beiträge zur Informatik: Band XIV*, pp. 159–168. Leipzig University, Leipzig.

Multilinguale Phrasenextraktion mit Hilfe einer lexikonunabhängigen Analysekomponekte am Beispiel von Patentschriften und nutzergenerierten Inhalten

Daniela Becks, Julia Maria Schulz, Christa Womser-Hacker, Thomas Mandl

Universität Hildesheim, Institut für Informationswissenschaft und Sprachtechnologie

Marienburger Platz 22, 31141 Hildesheim

E-mail: {daniela.becks, julia-maria.schulz, womser, mandl}@uni-hildesheim.de

Abstract

Die Extraktion von sinntragenden Phrasen aus Korpora setzt in der Regel eine verhältnismäßig tiefe linguistische Analyse der Texte voraus. Darüber hinaus ist häufig eine Adaptation der verwendeten Wissensbasen sowie der zugrunde liegenden Modelle notwendig, was sich meist als zeit- und arbeitsintensiv erweist. Der vorliegende Artikel beschreibt einen neuen sprach- und domänenübergreifenden Ansatz, der Aspekte von Shallow und Deep Parsing kombiniert. Ein Vorteil des vorgestellten Verfahrens besteht darin, dass es sich mit wenig Aufwand und ohne komplexe Lexika realisieren und auf andere Sprachen und Domänen übertragen lässt. Als Beispiel fungieren englische und deutsche Dokumente aus zwei sehr unterschiedlichen Korpora: Kundenrezensionen (nutzergenerierte Inhalte) und Patentschriften.

Keywords: Shallow Parsing, Multilinguale Phrasenextraktion

1. Einleitung

Vor dem Hintergrund einer globalisierten Welt liegen Informationen häufig in Dokumenten vor, die nicht in der Muttersprache der Benutzer verfasst sind. Um ihnen dennoch die Möglichkeit zu bieten, diese aufzufinden, bedarf es spezieller Methoden. Damit beschäftigt sich das Crosslinguale Information Retrieval (CLIR)¹. Die in diesem Kontext entstehenden Herausforderungen werden unter anderem bei Evaluierungsinitiativen wie CLEF² und NTCIR³ untersucht.

Eine weitere Entwicklung, die sich seit einiger Zeit im Bereich des Information Retrieval abzeichnet, liegt in der zunehmenden Ablösung des klassischen Bag-of-Words-Ansatzes, der bislang sowohl innerhalb des Indexierungsprozesses als auch im Rahmen der Anfrageformulierung Anwendung findet. In der Literatur wird derzeit

vermehrt auf die Vorteile von Phrasen gegenüber einfachen Termen hingewiesen (vgl. z.B. Tseng et al., 2007:1222). Diese zeigen sich auch anhand eines einfachen Recherchebeispiels. Eine Suchanfrage zum Thema *Züge der DB* liefert auch Dokumente zum Thema Datenbanken, da es eine ambige Abkürzung ist, dessen Bedeutung erst im Kontext eindeutig wird. Begreift man die einzelnen Terme jedoch als zusammengehörige Phrase, so wird diese Mehrdeutigkeit aufgelöst und es werden lediglich diejenigen Dokumente ausgewiesen, in denen die Kombination der Terme auftritt.

Das Extrahieren geeigneter Phrasen stellt jedoch vor einem multilingualen Hintergrund eine anspruchsvolle Aufgabe dar, da jedes Korpus unterschiedliche Besonderheiten aufweist, die es zu berücksichtigen gilt. Darüber hinaus spielt die Morphologie der einzelnen Sprachen eine entscheidende Rolle (z.B. abgetrennte Partikel zusammengesetzter Verben im Deutschen).

Innerhalb dieses Artikels wird ein Ansatz vorgestellt, der Shallow und Deep Parsing kombiniert und mit nur geringen Anpassungen sprach- und domänenübergreifend für die Extraktion von sinntragenden Phrasen verwendet werden kann. Als Anwendungsbeispiele fungieren Patentschriften und Kundenrezensionen, die in den Spra-

¹ Im crosslingualen Information Retrieval stimmen die Sprachen der Anfrage- und der Ergebnisdokumente nicht immer überein. Mit Hilfe einer deutschsprachigen Anfrage können beispielsweise auch englischsprachige Dokumente gewonnen werden.

² *Cross Language Evaluation Forum*: <http://clef2011.org>, <http://www.clef-campaign.org>

³ *National Institute of Informatics Test Collection for IR Systems*: <http://research.nii.ac.jp/ntcir/index-en.html>

chen Deutsch und Englisch vorliegen. In Zukunft ist geplant, Dokumente der Sprachen Spanisch und Französisch zu untersuchen.

Im Folgenden werden zunächst die beiden Anwendungsbereiche sowie die zugrunde liegenden Korpora vorgestellt (2.1). Des Weiteren werden einige Verfahren der Phrasenextraktion skizziert (3), an die sich die Beschreibung des sprach- und domänenübergreifenden Ansatzes anschließt (4). Dieser Artikel schließt mit einer Beschreibung des verwendeten Evaluierungsansatzes sowie ersten Ergebnissen ab.

2. Kontext der Forschungen

2.1. Anwendungsbereiche

Als Anwendungsbereiche für die entwickelte Phrasenextraktionskomponente werden in diesem Artikel zwei Projekte vorgestellt. Das erste Projekt findet in Kooperation mit dem FIZ Karlsruhe statt und fokussiert die Patentdomäne. Es zielt darauf ab, den Mehrwert von Phrasen für die Patentrecherche zu evaluieren (vgl. Becks, 2010:423). Das zugrunde liegende Korpus beinhaltet etwa 105.000 Dokumente der CLEF-IP⁴ Testkollektion 2009, welche sich aus ca. 1,6 Millionen Patent- und Anmeldeschriften des Europäischen Patentamtes zusammensetzt. Die Kollektion umfasst sowohl Dokumente in Englisch als auch Patente in Deutsch und Französisch (vgl. Roda et al., 2010:388).


Die Kundenrezensionen, die als Beispiel für nutzergenerierte Inhalte herangezogen werden, stammen aus einem Projekt, das sich mit crosslingualem Opinion Mining befasst, und sich dabei ebenfalls mit der Extraktion von Phrasen beschäftigt. Das Ziel dieses Projektes besteht darin, Phrasen zu extrahieren, die Meinungen bezüglich der rezensierten Produkte und deren Eigenschaften enthalten. Als Grundlage dient in diesem Fall ein Korpus aus Kundenrezensionen (vgl. Hu, Liu, 2004, Ding et al., 2008, Schulz et al., 2010).

Insbesondere im Hinblick auf die Länge der Dokumente unterscheiden sich beide Korpora signifikant, denn im


Falle von Patentschriften handelt es sich um sehr lange und komplexe Dokumente (vgl. u.a. Iwayama et al., 2003). Eine wesentliche Herausforderung besteht somit darin, dass die sprachübergreifende Phrasenextraktion für sehr unterschiedliche Textsorten und Phrasen unterschiedlicher Komplexität gleichermaßen effektiv funktionieren soll.

Eine Phrase wird als eine Kombination von Termen verstanden, die zueinander in einer Head-Modifier-Relation stehen. Diese Beziehung kann in verschiedenen Ausprägungen (z.B. Adjektiv-Nomen-Relation, Nomen-Präpositionalphrasen-Relation) auftreten. Die Phrasen unterscheiden sich jedoch von Chunks, die nach (Abney, 1991) typischerweise aus einem einzelnen *Content Word* bestehen, das von einer Konstellation von Funktionswörtern und Pre-Modifiern umgeben ist, und einem festen Template folgen (vgl. Abney, 1991:257). Betrachtet man das folgende Beispiel, so zeigt sich deutlich, dass eine Phrase über die Grenzen eines Chunks hinausgehen kann. Aufgrund der fokussierten Anwendungsbereiche Information Retrieval und Opinion Mining unterscheidet sich der hier verwendete Phrasenbegriff von der klassischen linguistischen Definition. Er umfasst auch Mehrwertgruppen (z.B. information retrieval system) und Kombinationen aus Subjekt und Prädikat, die im Deutschen auch diskontinuierlich sein können. Eine Liste der erfassten Phrasentypen findet sich in Abschnitt 5.

Beispiel:

[a system] [for information retrieval]

 Chunks

vs.

a [system for information retrieval]

 Phrase

2.2. Problemstellung und Anforderungen an die Phrasenextraktion

Die Entwicklung einer geeigneten Extraktionskomponente wird innerhalb des Projektkontextes durch zwei wesentliche Zielsetzungen bestimmt:

- Die Phrasenextraktion soll mit geringem Anpassungsaufwand für verschiedene europäische Sprachen realisierbar sein (ressourcenarmer Extraktionsansatz).

⁴ *Cross Language Evaluation Forum, Intellectual Property Track*

- Obgleich linguistische Ressourcen noch nicht flächendeckend verfügbar sind, soll die Phrasenextraktion für mehrere Sprachen möglich sein.

Die Phrasenextraktion muss dem „Unknown Words Problem“ entgegenwirken. Infolgedessen soll das System in der Lage sein, Wörter zu bearbeiten, die bislang weder in den vom System benutzten Korpora noch in Wörterbüchern erfasst sind (vgl. Uchimoto et al., 2001:91). Dies spielt insbesondere innerhalb der Patentdomäne eine bedeutende Rolle.

3. Verwandte Ansätze

Zu den traditionellen Methoden der Phrasenextraktion zählen unter anderem regelbasierte Verfahren wie das *Begrenzerverfahren* von Jaene und Seelbach (vgl. Jaene & Seelbach, 1975). Die Autoren haben es sich zur Aufgabe gemacht, für die Inhaltserschließung Phrasen in Form von Mehrwortgruppen, die sie als mehrere eine syntaktisch-semantische Einheit bildende Wörter definieren (vgl. Jaene & Seelbach, 1975:9), aus englischen Fachtexten zu ermitteln. Zu diesem Zweck definieren Jaene und Seelbach sogenannte Begrenzerpaare, die die zu extrahierenden Nominalphrasen einschließen (vgl. Jaene & Seelbach, 1975:7). Ein ähnliches Verfahren für die Extraktion von Nominalphrasen maximaler Länge, die mit dem Ziel der Identifikation von Fachtermini aus französischen Dokumenten dreier Domänen extrahiert werden, beschreiben (Bourigault & Jacquemin, 1999). In diesem Zusammenhang werden die Nominalphrasen in einem zweiten Schritt in ihre Bestandteile (Head und Modifier) zerlegt. Für den Extraktionsprozess werden sowohl Begrenzerpaare als auch die grammatische Struktur der Phrasen herangezogen. Vergleichbare Ansätze beschreiben (Tseng et al., 2007) für die Patentdomäne. Phrasen oder Schlüsselwörter werden hier auf Basis einer Stoppwortliste extrahiert. Als besonders geeignet erweisen sich dabei die längsten sich wiederholenden Phrasen (vgl. Tseng et al., 2007:1223). Auch (Guo et al., 2009) verwenden im Bereich Opinion Mining für die Extraktion von Produkteigenschaften aus Satzsegmenten im semistrukturierten Bereich von Kundenrezensionen Stoppwörter, ergänzt durch meinungstragende Wörter (z.B. Adjektive). Anhand dieses kurzen Überblicks zeigt sich bereits, dass sich die Phrasenextraktion

bislang überwiegend auf die Identifikation einfacher Nominalstrukturen konzentriert. In diesem Zusammenhang kommen neben den regelbasierten Ansätzen auch wörterbuchabhängige Verfahren wie beispielsweise das Dependenzparsing zum Einsatz. Im Information Retrieval kommen Dependenzrelationen häufig in Form von Head/Modifier-Paaren zum Einsatz, welche sich aus einem Head und einem Modifier zusammensetzen, wobei letzterer den Head präzisiert (vgl. Koster, 2004:423).

Head/Modifier-Paare bieten den Vorteil, dass sie neben syntaktischer auch semantische Information beinhalten (vgl. u.a. Ruge, 1989:9). Infolgedessen kommen sie vor allem innerhalb des Indexierungsprozesses zum Einsatz (vgl. Koster, 2004; Ruge, 1995) und erweisen sich in Form von Tripeln (Term-Relation-Term) im Zusammenhang mit Klassifikationsaufgaben als vorteilhaft (vgl. Koster, Beney, 2009).

4. Domänen- und sprachübergreifende Phrasenextraktion

Dieser Artikel beschreibt eine neue Methode für die Extraktion von Phrasen, die die beiden zuvor genannten Kategorien vereinigt. Das Ziel des dargestellten Extraktionsverfahrens besteht im Wesentlichen darin, ein Werkzeug für die Identifikation von Phrasen zur Verfügung zu stellen, das sich mit geringem Aufwand für unterschiedliche Domänen und Sprachen adaptieren lässt (z.B. Anpassung einzelner Begrenzerpaare oder der zulässigen Präpositionen bei Nomen-Genitiv-Phrasen (NG) bzw. Nomen-Präpositionalphrasen (NP)). Dabei wird auf den Einsatz von domänenspezifischen Wissensbasen verzichtet, um die Domänenunabhängigkeit zu gewährleisten. Die Semantik der extrahierten Phrasen darf dabei nicht außer Acht gelassen werden. Infolgedessen handelt es sich um ein Mischverfahren, das die Funktionalität eines Shallow Parsers aufweist, aber eine flache semantische Klassifikation aufgrund linguistischer Regeln gewährleistet (vgl. Becks & Schulz, 2011).

Innerhalb der Phrasenextraktionskomponente findet ein regelbasiertes Verfahren Anwendung, das das Begrenzerverfahren (vgl. Jaene & Seelbach, 1975, Bourigault & Jacquemin, 1999) und die Grundzüge des Dependenzparsings (vgl. z.B. Ruge, 1995) aufgreift. Die Extraktion der Phrasen erfolgt in diesem Fall mit Hilfe verschiede-

ner Regeln, in denen jeweils Paare von Begrenzern, definiert sind. Die Begrenzer sind, anders als in bisherigen Ansätzen, nicht Wörter, sondern morphosyntaktische Wortklassen (Pos-Tags). An dieser Stelle zeigt sich bereits, dass das entwickelte System lediglich auf die Implementierung entsprechender Regeln sowie einen Part-of-Speech-Tagger angewiesen ist. Es handelt sich somit um einen ressourcenarmen Ansatz.

Die implementierten Regeln variieren je nach Phrasentyp. Im Falle einer Adjektiv-Nomen-Relation (AN-R) wird die Phrase häufig von der Klasse *Artikel* und einem Interpunktionszeichen oder einer Präposition eingeschlossen (siehe Abb. 1). Darüber hinaus muss diese mindestens ein Adjektiv und ein Nomen enthalten, damit es sich um eine gültige AN-R handelt. Da die Kategorie *Artikel* sowohl die deutschen Artikel *der, die, das* als auch das englische Pendant *the* umfasst, kann diese Regel auch auf andere Sprachen angewendet werden. Diese abstrahierte Version des Begrenzerverfahrens ist demnach generalisierbar. Eine Einbindung komplexer Wortlisten erübrigt sich.

Wie bereits erwähnt, wurde zudem auf Grundzüge des Dependenzparsings zurückgegriffen. Daher verfügt jede der extrahierten Phrasen sowohl über einen Head als auch einen Modifier, deren Ermittlung ebenfalls regelbasiert erfolgt. Im Falle der in Abbildung 1 dargestellten Beispiele befindet sich der Head am Ende der Phrase („stud“; „front panel button layout“). Der Modifier ist diesem vorangestellt.

Anhand der Beispiele wird deutlich, dass es sich im Falle der extrahierten Phrasen nicht unbedingt um

5. Evaluierung

In der Regel erfolgt die Beurteilung der Qualität des gewonnenen Outputs anhand eines definierten *Goldstandards*. Dieser Ansatz wurde beispielweise von Verbene und Kollegen gewählt (vgl. Verbene et al., 2010). Als Evaluierungsbasis dient eine manuell annotierte Stichprobe bestehend aus 100 Sätzen. Die Berechnung der *Precision* erfolgt auf Basis eines Vergleichs der extrahierten Phrasen mit der annotierten Stichprobe (vgl. Becks & Schulz 2011: 391).

Für die Erstellung des hier verwendeten Goldstandards werden zunächst aus den beiden in Abschnitt 2.1 beschriebenen Korpora für die Sprachen Deutsch und Englisch zufällig Sätze mit einem jeweiligen Gesamtumfang von ca. 2000 Tokens ausgewählt. Basis für die Berechnung der Anzahl der Tokens ist der vom Pos-Tagger generierte Output. Infolgedessen gelten auch Interpunktionszeichen jeweils als ein Token. Diese werden manuell jeweils unabhängig von zwei Annotatoren (der erste und der zweite Autor des Papers) hinsichtlich der folgenden Phrasentypen annotiert:

- Subjekt-Prädikat (z. B. he thinks)
- Prädikat-Objekt (z. B. extract phrases)
- Verb-Adverb (z. B. extract easily)
- Mehrwortgruppen (z. B. information retrieval system)
- Adjektiv-Nomen (z. B. linguistic phrases)
- Nomen-Präpositionalphrase (z.B. system for retrieval)
- Nomen-Genitiv (z. B. rules of extraction)
- Nomen-Relativsatz bzw. Nomen-Partizip (z. B. phrases extracted by the system)

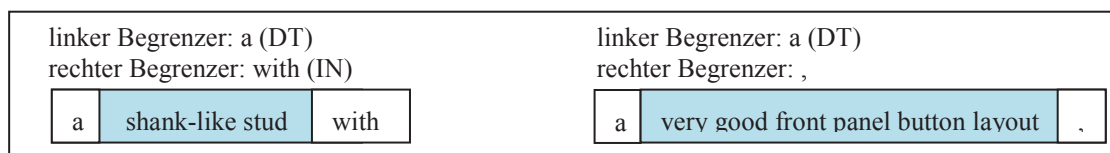


Abbildung 1: Beispiel einer extrahierten Adjektiv-Nomen-Phrase; links: Patentschrift (EP-1120530-B1), rechts: Kundenrezension (Hu & Liu 2004)

Head/Modifier-Paare handeln muss, sondern auch längere Phrasen mit mehreren Head/Modifier-Relationen durch dieses Verfahren abgebildet werden können.

Insgesamt sind für die Auswahl der englischen Sätze aus den Kundenrezensionen 688 und für die deutschen Sätze 639 Phrasen annotiert. Für die Patentdomäne liegen im Englischen 619 und im Deutschen 499 Phrasen vor. Von

den insgesamt 2445 Phrasen im Goldstandard sind ca. 51% unkontrovers, d.h. bei diesen Phrasen stimmen sowohl die von beiden Annotatoren identifizierten Phrasengrenzen als auch die annotierten Relationen überein. Weitere 27% der Phrasen weisen eine identische syntaktische Relation auf, unterscheiden sich jedoch im Hinblick auf die annotierten Phrasengrenzen. Diese Fehlerkategorie umfasst beispielsweise koordinierte Phrasen. Im Falle der nicht bzw. nur teilweise übereinstimmenden Phrasengrenzen wurde mittels Diskussion oder durch Hinzuziehen einer dritten unabhängigen Meinung eine Einigung herbeigeführt. Die zuvor genannten Prozentangaben weisen bereits darauf hin, dass sich die von den Annotatoren identifizierten und klassifizierten Phrasen sehr häufig decken. Die exakte Übereinstimmungsrate lässt sich anhand des berechneten Kappa ablesen. Folgende Formel wurde in diesem Zusammenhang zugrunde gelegt:

$$k = \frac{p_o - p_c}{1 - p_c} \quad (\text{aus Cohen, 1960:40})$$

Gemäß dieser Gleichung ergibt sich ein Kappa von 0.61. Vor dem Hintergrund, dass es sich bei den betrachteten Domänen um sehr divergierende Anwendungsfelder handelt und, dass sehr verschiedenartige, zum Teil diskontinuierliche Phrasen zu annotieren waren, kann dieser Wert als gut erachtet werden.

Für die Evaluierung werden die zusammengestellten Stichproben mit Hilfe der Phrasenextraktionskomponente automatisch annotiert. Der resultierende Output wird anschließend gegen den Goldstandard evaluiert, welcher neben der Phrase die syntaktische Relation und die Angabe der relativen Häufigkeit innerhalb der Stichprobe beinhaltet. Die Evaluierung geht demzufolge über einen Vergleich der Zeichenketten hinaus und erfolgt zusätzlich unter Einbeziehung der folgenden Kriterien:

- Syntaktische Relation
- Häufigkeit

Der Evaluierung liegen somit drei Faktoren zugrunde, welche als gleichgewichtet betrachtet werden. Es werden sowohl Exact als auch Partial Matches mit einer Abweichung von einem Term berücksichtigt. Phrasen, die im Hinblick auf die Phrasengrenze, die identifizierte Relation und die Häufigkeit mit der innerhalb des Gold-

standards hinterlegten Phrase übereinstimmen, gelten im Rahmen der Evaluierung als Exact Matches.

Für das Englische wird domänenübergreifend eine Precision von ca. 52% erzielt. Dabei werden für einige Phrasentypen deutlich bessere Werte erreicht (AN: 86,5%; NG: 76,7%; NN: 76%, VA: 71,8%). Die schlechtere Precision im Falle der übrigen Phrasentypen ist einerseits auf fehlerhafte Pos-Tags (dies gilt besonders für die Patentdomäne) und andererseits auf die Diskontinuität der Phrasen zurückzuführen, welche die Formalisierung deutlich erschwert.

I.d.R. zeigt sich, dass sowohl die Precision- als auch die Recall-Werte im Bereich der nutzergenerierten Inhalte durchschnittlich 13 bzw. 18 Prozentpunkte über denen im Patentbereich liegen. Dies unterstreicht die Schwierigkeit in diesem Kontext und legt die Vermutung nahe, dass es innerhalb der Patentdomäne gewisser Anpassungen bedarf. Um dies zu überprüfen, wurden für die Patentdomäne einige leichte Modifizierungen, z. B. Erweiterung der maximalen Phrasenlänge sowie die Berücksichtigung von Gerundien im Englischen, vorgenommen. Bereits eine geringe Anpassung der Verbalphrasen erhöht die Precision insgesamt auf 60,5% (+8,5%).

Im Deutschen zeigt sich für die bislang realisierten Nominalphrasen ein ähnliches Bild. Hier wird domänenübergreifend eine Precision von 63% erreicht. Auch hier scheiden einzelne Phrasentypen deutlich besser ab (z. B.: AN: 89,4%).

Insgesamt fällt auf, dass der Recall für beide Sprachen (ca. 38%) nicht sehr hoch ist. Dies lässt sich ebenfalls auf den Anwendungskontext zurückführen, denn für die Phrasenextraktion kommt der Precision in diesem Fall eine deutlich größere Bedeutung zu.

6. Schlussbetrachtung

Dieser Artikel bestätigt, dass sich mit einem ressourcenarmen, sprach- und domänenübergreifendem Ansatz z. T. gute Precision-Werte, die für die Phrasenextraktion im Retrieval-Kontext von vorrangiger Bedeutung sind, erzielen lassen. Allerdings weisen die Ergebnisse darauf hin, dass gewisse Modifikationen (z.B. innerhalb der

Patentdomäne) zu einer Steigerung der Ergebnisse führen können.

Zukünftig soll der Ansatz auf weiteren Sprachen (Französisch, Spanisch) getestet und der Einfluss des Pos-Tagging Modells untersucht werden, um die Genauigkeit des Algorithmus weiter zu verbessern.

7. References

- Abney, S. P. (1991): Parsing by Chunks. In: Berwick, R. C.; Abney, S. P.; Tenny, C. (Hrsg.): *Principle-based parsing. Computation and psycholinguistics*. Dordrecht: Kluwer (Studies in linguistics and philosophy, 44), S. 257-278.
- Becks, D. (2010): Begriffliche Optimierung von Patentanfragen. In: *Information - Wissenschaft & Praxis*, Jg. 61, H. 6-7, S. 423.
- Becks, D.; Schulz, J. M. (2011): Domänenübergreifende Phrasenextraktion mithilfe einer lexikonunabhängigen Analysekomponente. In: Griesbaum, J.; Mandl, Th.; Womser-Hacker, Ch. (Hrsg.): *Information und Wissen: global, sozial und frei? Boizenburg: Werner Hülsbusch (Schriften zur Informationswissenschaft Bd. 58)*, S. 388-392.
- Bourigault, D.; Jacquemin, Ch. (1999): Term extraction + term clustering: an integrated platform for computer-aided terminology. In: *Proceedings of the EACL'99* Stroudsburg, PA, USA: Association for Computational Linguistics, S. 15-22.
- Cohen, J. (1960): A Coefficient of Agreement for Nominal Scales. In: *Educational and Psychological Measurement* 20 (1), S. 37-46.
- Ding, X.; Liu, B.; Yu, P. S. (2008): A holistic lexicon-based approach to opinion mining. In: *Proceedings of the WSDM'08*. Palo Alto, California, USA: ACM, S. 231-240.
- Guo, H.; Zhu, H.; Guo, Z.; Zhang, X. X.; Su, Z. (2009): Product feature categorization with multilevel latent semantic association. In: *Proceeding of the CIKM'09*. Hong Kong, China: ACM, S. 1087-1096.
- Hu, M.; Liu, B. (2004): Mining Opinion Features in Customer Reviews. In: McGuinness, D. L.; Ferguson, G. (Hrsg.): *AAAI: AAAI Press/The MIT Press*, S. 755-760.
- Iwayama, M.; Fujii, A.; Kando, N.; Marukawa, Y. (2003): An Empirical Study on Retrieval Models for Different Document Genres: Patents and Newspaper Articles. In: *Proceedings of the SIGIR'03*. New York, NY, USA: ACM, S. 251-258.
- Jaene, H.; Seelbach, D. (1975): *Maschinelle Extraktion von zusammengesetzten Ausdrücken aus englischen Fachtexten*. Berlin, Köln, Frankfurt (Main): Beuth.
- Koster, C. H. A. (2004): Head/Modifier Frames for Information Retrieval. In: *Proceedings of the CICLing'04*. Seoul, Korea: Springer (LNCS 2945), S. 420-432.
- Koster, C. H. A.; Beney, J. G. (2009): Phrase-based document categorization revisited. In: *Proceedings of PaIR'09*. New York, NY, USA: ACM, S. 49-56.
- Roda, G.; Tait, J.; Piroi, F.; Zenz, V. (2010): CLEF-IP 2009: Retrieval Experiments in the Intellectual Property Domain. In: Peters, C.; Di Nunzio, G.; Kurimo, M.; Mandl, Th.; Mostefa, D.; Peñas, A.; Roda, G. (Hrsg.): *Multilingual Information Access Evaluation I. Text Retrieval Experiments*. Proceedings of CLEF '09. Berlin, Heidelberg: Springer (Lecture Notes in Computer Science), Bd. 6241, S. 385-409.
- Ruge, G. (1989): Generierung semantischer Felder auf der Basis von Frei-Texten. In: *LDV Forum* 6, H. 2, S. 3-17.
- Ruge, G. (1995): *Wortbedeutung und Termassoziation. Methoden zur automatischen semantischen Klassifikation*. Hildesheim, Zürich, New York: Olms.
- Schulz, J. M.; Womser-Hacker, Ch.; Mandl, Th. (2010): Multilingual Corpus Development for Opinion Mining. In: Calzolari, N.; Choukri, K.; Maegaard, B.; Mariani, J.; Odijk, J.; Piperidis, S.; Rosner, M.; Tapias, D. (Hrsg.): *Proceedings of the LREC'10*. Valletta, Malta: European Language Resources Association (ELRA), S. 3409-3412.
- Tseng, Y.-H.; Lin, C.-J.; Lin, Y.-I (2007): Text mining techniques for patent analysis. In: *Information Processing and Management*, Jg. 43, H. 5, S. 1216-1247.
- Uchimoto, K.; Sekine, S.; Isahara, H. (2001): The Unknown Word Problem: a Morphological Analysis of Japanese Using Maximum Entropy Aided by a Dictionary. In: Lee, L.; Harman, D. (Hrsg.): *Proceedings of the EMNLP '01: ACL*, S. 91-99.
- Verbene, S.; D'hondt, E.; Oostdijk, N. (2010): Quantifying the Challenges in Parsing Patent Claims. In: *Proceedings of AsPIRe'10*. Milton Keynes, S. 14-21.

Die Digitale Rätoromanische Chrestomathie – Werkzeuge und Verfahren für die Korpuserstellung durch kollaborative Volltexterschließung

Claes Neufeind, Jürgen Rolshoven, Fabian Steeg

Institut für Linguistik, Sprachliche Informationsverarbeitung

Universität zu Köln

Albertus-Magnus-Platz

50923 Köln

E-mail: c.neufeind@uni-koeln.de, rols@spinfo.uni-koeln.de, fabian.steeg@uni-koeln.de

Abstract

Das Paper beschreibt die Entwicklung und den Einsatz von Werkzeugen und Verfahren für die kollaborative Korrektur bei der Erstellung eines rätoromanischen Textkorpus mittels digitaler Tiefenerschließung. Textgrundlage bildet die „Rätoromanische Chrestomathie“ von Caspar Decurtins, die 1891-1919 in der Zeitschrift „Romanische Forschungen“ erschienen ist. Bei dem hier vorgestellten Ansatz werden manuelle und automatische Korrektur unter Einbeziehung von Angehörigen und Interessierten der rätoromanischen Sprachgemeinschaft über eine kollaborative Arbeitsumgebung kombiniert. In dem von uns entwickelten netzbasierten Editor werden die automatisch gelesenen Texte den Digitalisaten gegenübergestellt. Korrekturen, Kommentare und Verweise können nach Wiki-Prinzipien vorgeschlagen und eingebracht werden. Erstmals wird so die Sprachgemeinschaft einer Kleinsprache aktiv in den Prozess der Dokumentation und Bewahrung ihres eigenen sprachlichen und kulturellen Erbes eingebunden. In diesem Paper wird die konkrete Umsetzung der kollaborativen Arbeitsumgebung beschrieben, von der architektonischen Grundlage und aktuellen technologischen Umsetzung bis hin zu Weiterentwicklungen und Potentialen. Die Entwicklung erfolgt von Beginn an quelloffen unter <http://github.com/spinfo/drc>.

Keywords: Volltexterschließung, Korpuserstellung, kollaborative Korrektur

1. Einleitung

Für die Digitalisierung von Texten gibt es seitens der nationalen und internationalen Förderinstitutionen eine Vielzahl von Initiativen, Programmen und Projekten. Über die reine Massendigitalisierung hinaus zielen die Maßnahmen auch auf die digitale Tiefenerschließung von Texten. Diese ermöglicht zum einen den Zugriff über Volltextsuche, zum anderen kann sie zur Erstellung von spezialisierten Korpora genutzt werden, etwa auf Grundlage historischer Textsammlungen.

Ein wesentliches Problem der automatischen Volltexterschließung sind Lesefehler bei der optischen Zeichenerkennung (*Optical Character Recognition, OCR*). Besonders bei älteren Texten machen die unterschiedlichen Verschriftungstraditionen und variierenden Typographien eine fehlerfreie OCR faktisch unmöglich. Im Zuge der hier beschriebenen Digitalisierung der „Rätoromanischen Chrestomathie“ setzen wir deshalb

bei der Korrektur der OCR-Fehler auf die Einbindung von Angehörigen und Interessierten der rätoromanischen Sprachgemeinschaft über eine netzbasierte Arbeitsumgebung, in der die OCR-gelesenen Texte den zugrunde liegenden Digitalisaten gegenübergestellt sind.

2. Ähnliche Arbeiten

Die Idee einer kollaborativen Korrektur von OCR-Ergebnissen findet zunehmend auch im Kontext größerer strategischer Digitalisierungsprogramme Beachtung, so z.B. im IMPACT-Projekt¹ der Europäischen Kommission. Die Einschätzung, dass die Einbindung freiwilliger Korrektoren eine realistische Option ist, wird dabei u. a. durch die positiven Erfahrungen des *Australian Newspapers Digitisation Program (ANDP)*² der *National Library of Australia*

¹*Improving Access To Text*; <http://www.impact-project.eu>.

²Siehe <http://www.nla.gov.au/ndp/>.

gestützt, das im Zuge der Volltexterschließung der zwischen 1803 und 1954 in Australien erschienenen Zeitungen bereits seit 2008 erfolgreich eine Community-orientierte Fehlerkorrektur umsetzt (Holley, 2009). Einen vergleichbaren Ansatz plant auch das Deutsche Textarchiv (DTA)³. In der dort bislang nur intern eingesetzten Korrekturumgebung können Fehler allerdings nicht direkt vom Nutzer bearbeitet, sondern lediglich anhand einer differenzierten Fehlertypologie markiert und an die Mitarbeiter des DTA gemeldet werden, die diese anschließend offline korrigieren. Das Konzept der Verknüpfung von Digitalisat und Text in einem Editor wird zudem in dem im Rahmen des Textgrid-Projekts⁴ entwickelten Text-Bild-Link-Editor aufgenommen, der zwar eine kontrollierte Metadaten-Annotation von Bildelementen durch entsprechend qualifizierte Nutzer ermöglicht, jedoch aufgrund der fehlenden Benutzerverwaltung und Versionierung sowie der ausschließlich manuellen Text-Bild-Verknüpfung nicht für eine netzbasierte, kollaborative Korrektur von OCR-Ergebnissen durch interessierte Laien ausgelegt ist. Da die weiteren Ansätze zu Beginn unserer Arbeiten an der Digitalen Rätoromanischen Chrestomathie zum Teil noch nicht vorlagen (IMPACT, DTA), oder aber starke Differenzen im Ausgangsmaterial und damit im Digitalisierungs-Workflow aufweisen (großformatige Zeitungsseiten im ANDP), haben wir uns für eine Eigenentwicklung entschieden, um dadurch auch auf die speziellen Anforderungen einer mehrsprachigen Textbasis und das Fehlen von Korrekturlexika eingehen zu können. Während im DTA wie auch im Textgrid-Projekt der Schwerpunkt auf exakten Metadaten liegt, zielt der hier vorgestellte Ansatz auf die originalgetreue Wiedergabe des Textes anhand der Vorlage, weshalb auf elaborierte Korrekturguidelines verzichtet wurde.

3. Die Digitale Rätoromanische Chrestomathie

Die "Rätoromanische Chrestomathie" (RC) von Caspar Decurtins, die 1891-1919 in der Zeitschrift "Romanische Forschungen" (Erlangen: Junge) erschienen ist, gilt als wichtigste Textsammlung des Rätoromanischen (Egloff & Mathieu, 1986:7). Damit ist

sie eine hervorragende Basis für die Erstellung eines rätoromanischen Textkorpus. Mit ihren etwa 8000 Seiten Text aus vier Jahrhunderten, ihrer thematischen Vielfalt, unterschiedlichen Textsorten und Genres sowie der Abdeckung der fünf Hauptidiome des Bündnerromanischen ist sie für nahezu alle sprach- und kulturwissenschaftlichen Disziplinen von außerordentlichem Interesse. Sie stimuliert lexikographisches und lexikologisches, morphologisches und syntaktisches, semantisches und textlinguistisches, literaturwissenschaftliches, volkskundliches und historisches Arbeiten. Sie ermöglicht datenbasierte Untersuchungen zu Strukturen und Textsorten und ist aufgrund ihres Varietätenreichtums von hohem Wert für diachrone (über vier Jahrhunderte reichende) und diatopische (fünf Hauptidiome umfassende) Untersuchungen, etwa zu Sprachkontakt, Sprachverwandtschaft und Sprachwandel.

3.1. Digitalisierung und OCR

Ausgangspunkt der Korpuserstellung sind die Digitalisate der RC aus der Zeitschrift "Romanische Forschungen", die von der Staats- und Universitätsbibliothek Göttingen im Rahmen des Digizeitschriften-Projekts⁵ digitalisiert und zusammen mit den in einem METS-basierten Format⁶ erstellten Metadaten zur Verfügung gestellt wurden. Um die Digitalisate für die textuelle Verarbeitung zugänglich zu machen, werden sie mittels OCR in eine maschinenlesbare Form überführt. Die hohe typographische und orthographische Vielfalt der RC stellt dabei eine besondere Herausforderung für die OCR dar, um so mehr, als der Zeichenerkennung keine angemessenen Korrekturlexika für die verschiedenen Idiome zur Verfügung stehen. Gerade die älteren Texte der Chrestomathie sind orthographisch nicht normiert, weil die Idiome des Bündnerromanischen unterschiedlichen Verschriftungsformen und -traditionen folgen. Auf Grundlage des OCR-Ergebnisses werden PDF-Dateien generiert, bei denen der erkannte Text unter dem Digitalisat positioniert wird. Das generierte PDF enthält damit nicht nur den gesamten Text, sondern auch die

³Siehe <http://www.deutschestextarchiv.de/>.

⁴Siehe <http://www.textgrid.de/>.

⁵Siehe <http://www.digizeitschriften.de>

⁶Siehe <http://gdz.sub.uni-goettingen.de/entwicklung/standardisierung/>

Positionskordinaten der einzelnen Wörter. Die Extraktion der Wörter mitsamt ihrer Positionskordinaten erfolgte mit der Software-Bibliothek PDFBox⁷. Die ausgelesenen Informationen (Wort und Position) werden in XML-Form abgelegt und stellen die Grundlage für das Highlighting-Feature in der Korrekturumgebung dar.

3.2. Der DRC-Editor

Kern des hier beschriebenen Ansatzes ist die Erstellung einer kollaborativen Korrekturumgebung, in der die Digitalisate und die mittels OCR gewonnenen Texte zusammengeführt werden. Mit dem Editor können die elektronisch eingeleseenen Texte der RC durchsucht, gelesen und bearbeitet werden.

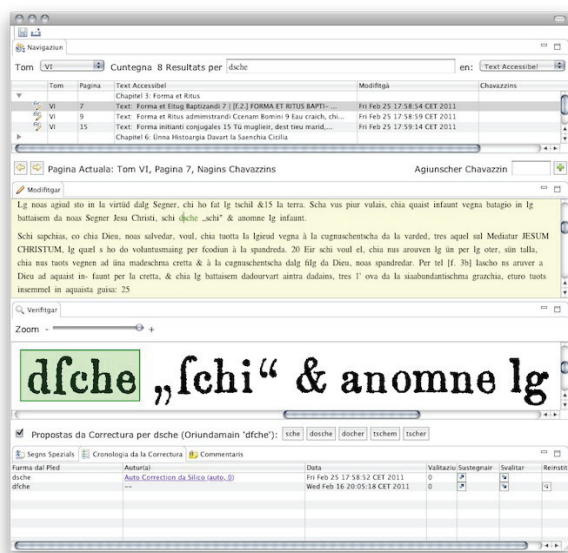


Abbildung 1: Screenshot des Editors (Beta-Version)

Die Auswahl der zu bearbeitenden Seiten erfolgt über Volltextsuche sowie über die aus dem Digizeitschriften-Projekt übernommenen Metadaten. Ziel der Bearbeitung ist die Erstellung einer fehlerfreien digitalen Textfassung, weshalb zu Vergleichszwecken stets die Originalfassung als digitales Faksimile mit angezeigt wird. Die Bildendarstellung ist dabei mit dem Text gekoppelt: Während man den Text wortweise bearbeitet, wird das jeweils korrespondierende Wort unter Nutzung der bei der OCR gewonnenen Positionskordinaten auf dem Bild hervorgehoben (siehe Abbildung 1, Bereich *Verifitgar*). Die Synchronisation von Text und

Bildkoordinaten bleibt auch bei Korrekturen bestehen, da die vorgenommenen Änderungen ebenso wie die Positionskordinaten der ursprünglichen Wortform zugeordnet werden. Über die Tastatur nicht verfügbare Sonderzeichen können über ein Auswahlfenster hinzugefügt werden.

Als zusätzliches Hilfsmittel besteht die Option zur Anzeige von Korrekturvorschlägen (siehe Abbildung 1, *Propostas da Correcturas*), die auf Grundlage von Wortlisten über die Levenshtein-Distanz, einen Algorithmus für den Stringvergleich, ermittelt werden. Da solche Wortlisten bzw. Prüffklassen derzeit nur für eines der Idiome, das Surselvische, verfügbar sind, ist zusätzlich ein automatisierter Auf- und Ausbau von Benutzerlexika geplant, indem die manuellen Korrekturen unter Nutzung der Versionierungsmechanismen der Korrekturumgebung aufgezeichnet werden. Hieraus resultiert eine stetig wachsende Liste von als korrekt bestätigten Wörtern, die einerseits als Grundlage für die Berechnung von Korrekturvorschlägen dient, andererseits dazu eingesetzt werden kann, dem Nutzer nach einer vorgenommenen Korrektur Verbesserungsvorschläge an anderen, gleichen oder ähnlichen Stellen des Textes vorzuschlagen. Sämtliche Bearbeitungen werden unter Angabe von Nutzer und Bearbeitungszeitpunkt protokolliert. Damit verbunden ist ein einfaches Bewertungs- und Wettbewerbssystem, das über die Korrekturen Buch führt.

Die Erfahrungen im laufenden Projekt haben gezeigt, dass über die reine Korrektur hinaus auch die Möglichkeit zu einer Verschlagwortung und Kommentierung nutzerseitig gewünscht ist, da dies neben erweiterten Recherchemöglichkeiten auch die Möglichkeit zur Markierung unklarer oder (im Sinne der kollaborativen Bearbeitung) strittiger Fälle bietet. In der aktuellen Beta-Version können die Daten deshalb auf Seitenebene mittels frei wählbarer Tags oder durch Hinzufügung von Freitext-Kommentaren annotiert werden. Über eine fehlerfreie Dokumentation hinaus erfolgt auf diese Weise auch eine Anreicherung der Texte. Hierbei wird die Textbasis in gewissem Sinne 'aktualisiert', indem das Wissen der Sprecher in Form von Metadaten (Schlagworte, Verweise, Nutzungskontexte) in die Texte zurückfließt.

⁷Siehe <http://pdfbox.apache.org/>.

3.3. Das DRC-Portal

Für den Datenzugriff wurde neben dem Editor eine mehrsprachige Portalseite erstellt, die als zentraler Anlaufpunkt für interessierte Nutzer dient (vgl. Abbildung 2). Über das Portal kann der DRC-Editor heruntergeladen und ein Account für dessen Benutzung angelegt werden.

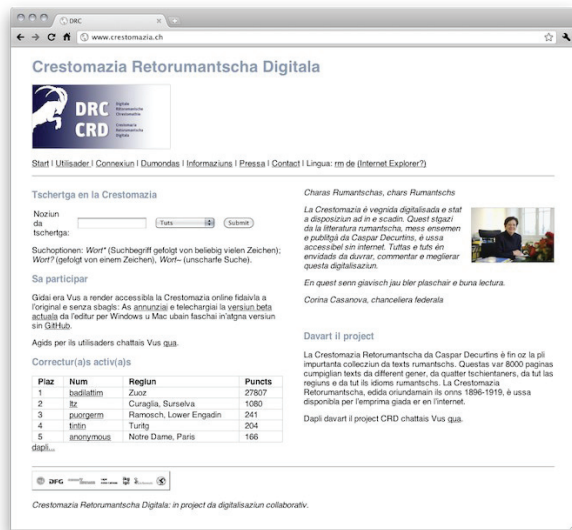


Abbildung 2: Portalseite der DRC (siehe <http://www.crestomazia.ch>)

Neben Hilfestellungen und Hinweisen zum Editor bietet die Portalseite erweiterte Recherchemöglichkeiten und enthält Hintergrundinformationen zum Projekt sowie zu ausgewählten Aspekten der bearbeiteten Daten.

3.4. Einbindung der Sprachgemeinschaft

Von zentraler Bedeutung für das hier beschriebene Vorgehen war die Frage, wie die Einbindung von Sprechern in einen kollaborativen Erschließungsprozess erfolgen kann. Um die Beteiligung einer ausreichenden Zahl von Sprechern sicherzustellen, setzten wir auf die Zusammenarbeit mit Partnern vor Ort, die neben der Presse- und Öffentlichkeitsarbeit auch eine Nutzerakquise übernehmen. Das Projekt wurde mit Hilfe der Schweizer Partner über die lokalen und überregionalen Medien propagiert. In Kombination mit einer gezielten Nutzerakquise konnte dadurch bereits für die aktuelle Beta-Version des DRC-Editors eine größere Anzahl an Nutzern gewonnen werden. So waren im August 2011 ca. 100 Nutzer angemeldet, seit dem Schaltungstermin

der DRC Anfang Juni 2011 wurde etwa ein Drittel der Texte bearbeitet.

4. Systemarchitektur

Der Natur des Vorhabens wird eine dreischichtige Systemarchitektur gerecht: Gesamtziel ist die kollaborative Produktion annotierter, textueller Daten. Diese Daten sind für alle Benutzer des Systems identisch, und können daher zentral gespeichert werden (Datenschicht). Verschiedene Nutzer sollen unabhängig voneinander auf diese Daten zugreifen und diese verändern können, wobei die Integrität der Daten gewährleistet werden muss (Logikschicht). Der Zugriff erfolgt über eine graphische Benutzerschnittstelle (Präsentationsschicht).



Abbildung 3: Grundlegende Systemarchitektur

Die Präsentationsschicht kommuniziert mit der Logikschicht und diese mit der Datenschicht. Da es keine direkte Verbindung zwischen Präsentations- und Datenschicht gibt, ist das System lose gekoppelt und erlaubt Austausch und Wiederverwendung der Schichten, etwa für eine Nutzung der Daten in anderen Kontexten.

4.1. Technologien

Aufgrund des modernen Programmierkonzepts, der hohen Modularität und Wiederverwertbarkeit durch *OSGi*⁸, der nativen GUI-Technologie sowie der Integration von Webstandards (z.B. CSS zur Gestaltung der GUI), haben wir uns für *Eclipse4*⁹ als Technologie für die Präsentationsschicht entschieden. Für eine kompakte und zugleich effiziente und kompatible Logikschicht setzen wir auf die JVM-Sprache *Scala*¹⁰. Für die Datenschicht wird mit *eXist*¹¹ eine XML-Datenbank eingesetzt. Da *eXist* über eine eingebaute Serverfunktionalität verfügt, war es zweckmäßig, die Logikschicht als Teil des Clients umzusetzen, und so keine eigenen serverseitigen Komponenten imple-

⁸Open Service Gateway Initiative, siehe <http://www.osgi.org/>.

⁹Siehe <http://eclipse.org/eclipse4/>.

¹⁰Siehe <http://www.scala-lang.org/>.

¹¹Siehe <http://exist.sourceforge.net/>.

mentieren zu müssen. Über den Datenbankserver können die Daten unabhängig von der beschriebenen Infrastruktur über standardisierte REST-Schnittstellen¹² zur Verfügung gestellt werden.

4.2. Implementierungen

Die Beta-Version des Editors ist als Eclipse-basierte Desktop-Applikation realisiert, die als Client des Datenbankservers fungiert. Der Editor wurde mit automatischen Aktualisierungen versehen, um neue Funktionalitäten und Fehlerbehebungen in der Software ohne Aufwand seitens der Nutzer bereitzustellen. Damit ergibt sich die folgende technologische Umsetzung der oben skizzierten Architektur:

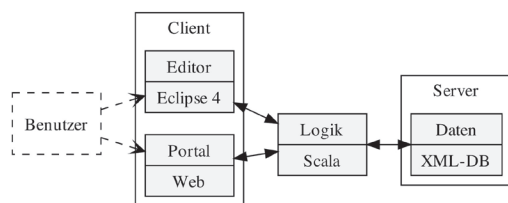


Abbildung 4: Implementierung der Architektur in der aktuellen Beta-Version

Derzeit arbeiten wir an alternativen Umsetzungen der GUI. Die aktuelle Beta-Version ermöglicht sowohl eine Weiterentwicklung zu einer Offline-fähigen Desktop-Applikation, die ohne Netzzugang verwendet werden kann und bei Bedarf die Daten mit dem Server synchronisiert, als auch die automatische Generierung einer Web-Oberfläche mithilfe des *RAP-Frameworks*¹³ (vgl. Abbildung 5).

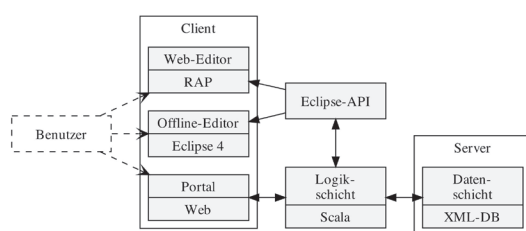


Abbildung 5: Alternative Umsetzungen der Architektur

¹²Representational State Transfer, vgl. dazu (Fielding, 2000).

¹³Rich Ajax Platform, siehe <http://eclipse.org/rap/>.

Die Software-Entwicklung erfolgte von Beginn an quelloffen; der vollständige Programmcode steht ebenso wie die jeweils aktuelle Version des Editors unter <https://github.com/spinfo/drc> frei zur Verfügung.

5. Erweiterungen

Mit der Digitalen Rätoromanischen Chrestomathie wird erstmals der digitale Zugriff auf eine größere rätoromanische Textsammlung geschaffen. Über die reine Dokumentation und Archivierung hinaus kann eine frei zugängliche RC eine Vielzahl neuer Impulse für die wissenschaftliche, mediale, edukative, aber auch private Nutzung geben. Die Möglichkeiten reichen von historischen und genealogischen Recherchen nach Personen und Ortsnamen über die kreative Auseinandersetzung durch Hinzufügung eigener Texte oder Übersetzungen, bis hin zur lexikographischen Arbeit mit der RC. Für eine Nutzung jenseits einfacher Suchanfragen ist zudem eine Annotation der Texte mit linguistischen Merkmalen geplant¹⁴. Insbesondere für eine adäquate (computer-)linguistische Nachnutzung bedarf es einer linguistischen Aufbereitung der erschlossenen Texte, da die reine Volltexterschließung nur als ein erster Schritt auf dem Weg zur Bereitstellung von computer- bzw. korpuslinguistisch ausgiebig nutzbaren Ressourcen betrachtet werden kann.

Analog zum hier beschriebenen Vorgehen soll auch die linguistische Annotation durch die Kombination automatischer und manueller Verfahren erfolgen. Um der weitgehend fehlenden orthographischen Normierung der RC zu begegnen, sollen in einem Folgeprojekt zunächst die für die fünf Hauptdiome verfügbaren lexikalischen Ressourcen digital aufbereitet werden. Auf dieser Grundlage automatisch vorgenommene Annotationen können anschließend über das entsprechend erweiterte Editor-Werkzeug durch Muttersprachler und Interessierte kollaborativ überprüft und ggf. korrigiert bzw. ergänzt werden. Das aus lexikalischer und manueller Annotation gewonnene Wissen soll mittels spezialisierter Lernverfahren zur erneuten automatischen Annotation der Texte genutzt werden.

¹⁴Vgl. dazu bspw. das Vorgehen im Projekt "Text+Berg digital" (Volk et al., 2010), siehe auch <http://www.textberg.ch>.

6. Potentiale

Durch den hier vorgestellten Ansatz werden Probleme der OCR gerade bei älteren und typographisch varianten Schriftsystemen abgefangen. Die im Vorhaben eingesetzten Erschließungs- und Auszeichnungstechniken können in der Folge auf weitere Textsammlungen des Bündnerromanischen und anderer kleiner Sprachen angewandt werden. Über das konkrete materielle Ziel der Erstellung eines rätoromanischen Textkorpus hinaus werden damit übertragbare und somit nachhaltige, kompetenzorientierte Verfahren entwickelt, die für die Tiefendigitalisierung des schriftlichen kulturellen Erbes kleinerer Sprachgemeinschaften prototypisch sind. Von besonderem Interesse ist hier auch die Möglichkeit für Mitglieder solcher Sprachgemeinschaften, über Wiki-Technologien den Erhalt des eigenen sprachlichen und kulturellen Erbes aktiv zu unterstützen.

7. Danksagung

Die Digitale Rätoromanische Chrestomathie ist ein gemeinsames Projekt der Sprachlichen Informationsverarbeitung und der Universitäts- und Stadtbibliothek Köln. Für die Organisation und Durchführung in der Schweiz konnten wir mit Dr. Florentin Lutz einen ausgewiesenen Linguisten und sehr gut vernetzten Muttersprachler gewinnen. Das DRC-Projekt wird von der Deutschen Forschungsgemeinschaft gefördert. In der Schweiz erhielt das Projekt zusätzliche finanziellen Unterstützung durch das Legat Anton Cadonau, das Institut für Kulturforschung Graubünden und das Kulturamt des Kantons Graubünden. Auch seitens der rätoromanischen Verbände und Organisationen erfuhr das Projekt regen Zuspruch und weitere Unterstützung, insbesondere durch die Lia Rumantscha¹⁵, den Dachverband der Bündnerromanen, sowie die Societad Retorumantscha¹⁶, den Trägerverein des „Dicziunari Rumantsch Grischun“, einem der vier nationalen Wörterbücher der Schweiz. All diesen Einrichtungen schulden wir unseren herzlichsten Dank.

8. Referenzen

- Decurtins, C. (1984-1986): Rätoromanische Chrestomathie. Band 1-14. Chur: Octopus-Verlag / Società Retorumantscha.
- Egloff, P., Mathieu, J. (1986): Rätoromanische Chrestomathie - Register. In: Rätoromanische Chrestomathie, Band 15. Chur: Octopus-Verlag / Società Retorumantscha.
- Fielding, R (2000): Architectural Styles and the Design of Network-based Software Architectures. Doktorarbeit, University of California, Irvine.
- Holley, R. (2009): Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers. National Library of Australia.
- Volk, M., Bubenhofer, N., Althaus, A., Bangerter, M., Furrer, L., Ruef, B. (2010): Challenges in Building a Multilingual Alpine Heritage Corpus. In: Seventh International Conference on Language Resources and Evaluation (LREC), Malta 2010.

¹⁵Siehe <http://www.liarumantscha.ch>.

¹⁶Siehe <http://www.drg.ch/main.php?l=r&a=srr>.

Ein umgekehrtes Lehnwörterbuch als Internetportal und elektronische Ressource: Lexikographische und technische Grundlagen

Peter Meyer, Stefan Engelberg

Institut für Deutsche Sprache

Mannheim

E-mail: meyer@ids-mannheim.de, engelberg@ids-mannheim.de

Abstract

Der vorliegende Beitrag stellt einen neuartigen Typ von mehrsprachiger elektronischer Ressource vor, bei dem verschiedene Lehnwörterbücher zu einem ‚umgekehrten Lehnwörterbuch‘ für eine bestimmte Gebersprache zusammengefasst werden. Ein solches Wörterbuch erlaubt es, die zu einem Etymon der Gebersprache gehörigen Lehnwörter in verschiedenen Nehmersprachen zu finden. Die Entwicklung einer solchen Webanwendung, insbesondere der zugrundeliegenden Datenbasis, ist mit zahlreichen konzeptionellen Problemen verbunden, die an der Schnittstelle zwischen lexikographischen und informatischen Themen liegen. Der Beitrag stellt diese Probleme vor dem Hintergrund wünschenswerter Funktionalitäten eines entsprechenden Internetportals dar und diskutiert einen möglichen Lösungsansatz: Die Artikel der Einzelwörterbücher werden als XML-Dokumente vorgehalten und dienen als Grundlage für die gewöhnliche Online-Ansicht dieser Wörterbücher; insbesondere für portalweite Abfragen werden aber grundlegende, standardisierte Informationen zu Lemmata und Etyma aller Portalwörterbücher samt deren Varianten und Wortbildungsprodukten (hier zusammenfassend als ‚Portalinstanzen‘ bezeichnet) sowie die verschiedenartigen Relationen zwischen diesen Portalinstanzen zusätzlich in relationalen Datenbanktabellen abgelegt, die performante und beliebig komplex strukturierte Suchabfragen gestatten.

Keywords: Lehnwörter, elektronische Lexikografie, mehrsprachige Ressource, Internetportal

1. Ein Lehnwörterbuchportal als ‚umgedrehtes Lehnwörterbuch‘

Ziel des vorgestellten Projekts ist ein Internet-Wörterbuchportal für Lehnwörterbücher, die Entlehnungen aus dem Deutschen dokumentieren. Dieses Portal ist dadurch gekennzeichnet, dass zum einen die eingestellten Wörterbücher als Einzelwerke veröffentlicht werden und zum anderen auf Portalebene komplexe Abfragen über sämtliche integrierte Wörterbücher hinweg formuliert werden können, zum Beispiel nach dem Weg einzelner deutscher Quellwörter über Mittlersprachen in die verschiedenen Zielsprachen, nach sämtlichen Lehnwörtern in bestimmten historischen Zeitspannen und geographischen Räumen, oder auch nach sämtlichen deutschen Lehnwörtern, die bestimmte Charakteristika aufweisen (z. B. Wortart, semantische Klasse). Das Portal realisiert damit – nicht in den Einzelwörterbüchern, aber in seiner Gesamtheit – als umgekehrtes Lehnwörterbuch das Konzept eines neuen

Wörterbuchtyps.¹ Während es in der Sprachkontaktlexikographie – etwa in Fremdwörterbüchern – üblich ist, Entlehnungsprozesse aus der Perspektive der Zielsprache zu beschreiben, erfasst das geplante Portal aus der Perspektive der Quellsprache die Wege, die deutscher Wortschatz in andere Sprachen genommen hat (Engelberg, 2010). Gegenwärtig wird am Institut für Deutsche Sprache (Mannheim) im Rahmen eines über 18 Monate laufenden und vom Beauftragten der Bundesregierung für Kultur und Medien geförderten Pilotprojektes die grundsätzliche Softwarearchitektur des Portals entwickelt und implementiert sowie die Integration dreier Lehnwörterbücher in das Portal vorgenommen, und zwar zu deutschen Entlehnungen im Polnischen (Vincenz & Hentschel, 2010), zu deutschen Entlehnungen im Teschener Dialekt des Polnischen (Menzel & Hentschel, 2005) und zu deutschen

¹Wiegand (2001) spricht in diesem Zusammenhang von aktiven bilateralen Sprachkontaktwörterbüchern. Wörterbücher dieses Typs sind extrem selten, vgl. auch (Engelberg, 2010). (Görlach, 2001) ist das einzige nennenswerte Beispiel.

Entlehnungen im Slovenischen (Striedter-Temps, 1963). Da das Portal auf Offenheit bezüglich der Integration weiterer Ressourcen konzipiert ist, können jederzeit weitere Lehnwörterbücher integriert werden. Entsprechende Wörterbücher zu Entlehnungen aus dem Deutschen existieren zu relativ vielen Sprachen (Englisch, Japanisch, Portugiesisch, Schwedisch, Serbokroatisch, Tok Pisin, Ukrainisch, Usbekisch, ...). Hier wären entsprechende Kooperationen anzustreben und Rechtsfragen zu klären.²

2. Nutzen eines Lehnwörterbuchportals

Das Portal soll sowohl für Laien wie für Wissenschaftler nutzbar sein. Die Laiennutzung kann über einfache Suchanfragen erfolgen, die wissenschaftliche Nutzung orientiert sich an der Möglichkeit komplexer Suchanfragen und an direkten Schnittstellen (Webservices). Dabei wird sowohl die sprachwissenschaftliche Sprachkontaktforschung wie auch die historisch, soziologisch oder anthropologisch ausgerichtete Kulturkontaktforschung Nutzen aus dem Portal ziehen.

Im Rahmen der wissenschaftlichen Nutzung soll das Portal nicht nur philologisch motivierte, interpretative Einzelstudien unterstützen, sondern durch die in ihm realisierte Kumulation von Daten auch spezifische neuartige, zum Teil quantitativ orientierte Forschungsfragen ermöglichen. Dazu gehören Untersuchungen

- zum Zusammenhang zwischen bestimmten Typen von soziokulturellen Entwicklungen (Herrschaftswechsel, Migration, Technologieschub) und Zeitverlaufstypen der Entlehnungsfrequenzen von Lexemen (wie etwa eine plötzliche oder eine eher graduelle quantitative Zunahme von Entlehnungen),
- zu Faktoren und Prozessen der Etablierung von Lehnwörtern,³
- dazu, ob verschiedene Typen des Sprachkontakts typische quantitative und zeitliche Verteilungsmuster von Lehnwörtern hervorbringen,⁴

- zur Lebensdauer von Lehnwörtern (insoweit die integrierten Wörterbücher auch das Verschwinden oder die Obsoletheit von Entlehnungen verzeichnen), abhängig von onomasiologischen, grammatischen und anderen Faktoren, vgl. etwa (Schenke, 2009; Hentschel, 2009),
- zu Lehnwortketten (z. B. Deutsch > Polnisch > Weißrussisch > Russisch > Usbekisch) im Zusammenhang mit onomasiologischen und quantitativen Faktoren,
- zu „Germanoversalien“, d. h. etwa dazu, ob bestimmte phonologische, morphologische oder semantische Eigenschaften deutscher Lexeme besonders entlehnungsfördernd sind.

3. Grundsätzliche Überlegungen zur lexikographischen Datenstruktur des Portals

Hinsichtlich der Datenorganisation des Lehnwörterbuchportals lassen sich auf einer konzeptionellen Ebene grob drei Bereiche unterscheiden:

- (1) Lexikographische Grundlage des Portals sind einzelne Lehnwörterbücher traditionellen Zuschnitts, die nach den fremdsprachigen Lehnwörtern einer bestimmten Nehmersprache lemmatisiert sind.
- (2) Um sprach- und wörterbuchübergreifende Suchen im Portal zu ermöglichen, muss über diese Datengrundlage eine möglichst dünne Zugriffsstruktur gelegt werden, die von den Idiosynkrasien der Einzelwörterbücher abstrahiert.
- (3) Für die Etyma der Gebersprache muss eine ‚Metalemmaliste‘ erstellt werden, deren Einträge jeweils über die unter Punkt 2 genannte Abstraktionsschicht untereinander und mit zugehörigen Artikeln der Einzelwörterbücher vernetzt sind.

Die folgenden Unterabschnitte stellen die in den drei genannten Bereichen auftretenden lexikographischen und technischen Anforderungen und Probleme ausführlicher dar, bevor im letzten Abschnitt die technische Umsetzung ihres Zusammenspiels erörtert wird.

²Zum Teil ist die Beschreibungssprache in diesen Wörtern die Quellsprache (z. B. Usbekisch, Portugiesisch), so dass im Falle entsprechende Übersetzungen erforderlich wären.

³ Solche Studien können auf lexikographischer und sprachübergreifender Basis Ergebnisse aus korpusbasierten Arbeiten zum lexikalischen Entrenchment von Entlehnungen komplementieren, vgl. (Chesley & Baayen, 2010).

⁴Sprachkontakttypen wären etwa (i) langandauernder Kontakt

an Bevölkerungsgrenzen (Deutsch – Slowenisch, Deutsch – Polnisch), (ii) Kontakt durch Emigration mit Sprachinselnbildung (Deutsch – Rumänisch, Deutsch – Russisch, Deutsch – Amerikanisches Englisch) und Kontakt durch Elitenaustausch (Deutsch – Japanisch, Deutsch – Russisch, Deutsch – Britisches Englisch, Deutsch – Tok Pisin).

3.1. Die Ebene der Einzelwörterbücher

Die zugrundeliegenden Lehnwörterbücher werden im Regelfall bereits existierende Werke sein, die nicht von vornherein für ein Lehnwörterbuchportal des hier diskutierten Typs entwickelt worden sind. Technische Minimalanforderung für die Verwendung im Portal ist, dass die Wörterbücher in geeigneter Form digitalisiert bzw. retrodigitalisiert als XML-Dokumente vorliegen.⁵ Auch eine Bilddigitalisierung ist denkbar, sofern zu jedem Artikel zusätzlich ein XML-Dokument mit den portalrelevanten lexikographischen Daten (und gegebenenfalls Verweisen auf Bildkoordinaten im Digitalisat) vorliegt. Angesichts der enormen Vielfalt möglicher Makro- und Mikrostrukturen in Wörterbüchern ist es nicht praktikabel, für das Portal ein festes XML-Schema vorzugeben, in das sich die XML-Repräsentationen aller Wörterbücher überführen lassen müssen. Es wird jedoch, um weitgehend automatisierte Verarbeitung zu ermöglichen, vom XML-Schema für die Einzelartikel eines jeden Wörterbuchs jeweils verlangt, dass es möglichst weitgehend von Layout- und Präsentationsaspekten abstrahiert, etwa im Sinne der TEI.dictionaries-Richtlinien; vgl. (Burnard & Bauman, 2010). Es gibt gute Gründe, die XML-Digitalisate der Ausgangswörterbücher selber nicht mit portalrelevanten Informationen anzureichern. Abgesehen von urheberrechtlichen Erwägungen und dem angestrebten Erhalt der Wörterbücher als digitalen Einzelpublikationen ist es so möglich, dass an den Einzelwörterbüchern völlig unabhängig von ihrer Nutzung im Lehnwörterbuchportal weiterhin Veränderungen und Erweiterungen von den Autoren des betreffenden Werks vorgenommen werden.

Ähnlich wie bei anderen Portalen können ganze Wörterbuchartikel oder Teile davon (XML-Dokumente bzw. XML-Fragmente) beispielsweise durch

⁵ Aus expositorischen Gründen wird hier auf der Ebene der Einzelwörterbücher durchgehend von einer XML-basierten Datenhaltung ausgegangen, so wie sie im Projekt selber tatsächlich verwendet wird. Technisch lassen sich die Mikrostrukturen von Wörterbüchern natürlich auch in relationalen Datenbankschemata abbilden, was aus Performanzgründen ratsam sein kann. Andererseits können einige moderne Datenbankmanagementsysteme (z. B. Oracle) XML-Daten mit fester Struktur intern ohnehin relational repräsentieren. Vgl. z. B. (Müller-Spitzer & Schneider, 2009) für das OWID-Portal als ein konkretes Beispiel zur texttechnologischen Umsetzung von XML-Verarbeitung in einem Wörterbuchportal.

XSL-Transformationen in eine geeignete HTML-Präsentation überführt werden. Dies ist die Grundlage für eine wörterbuchspezifische Online-Ansicht der Einzelwörterbuchartikel, vgl. (Engelberg & Müller-Spitzer, 2011) für eine ausführlichere Darstellung.⁶ Die XML-Repräsentation ermöglicht außerdem im Prinzip beliebig komplexe Suchvorgänge auf den Einzelwörterbüchern, da konkrete Informationen über Abfragesprachen wie XPath und XQuery aus den Artikeln ausgelesen werden können. Allerdings sind solche XML-basierten Abfragen häufig datenbankseitig mit hohen Verarbeitungskosten versehen und daher für performante Webanwendungen kaum praktikabel. Dies ist ein wesentlicher Grund, die für wörterbuchspezifische sowie portalweite (wörterbuchübergreifende) Suchen relevanten Informationen zusätzlich in separaten relationalen Datenbanktabellen vorzuhalten. Diese zusätzlichen Tabellen ermöglichen nicht nur ungleich performantere Datenbankabfragen, sie dienen auch, wie im folgenden ausgeführt wird, dazu, von den Spezifika der Einzelwörterbücher zu abstrahieren.

3.2. Wörterbuchübergreifende Abstraktionsschicht

Im Normalfall werden die einzelnen Lehnwörterbücher hinsichtlich ihrer Artikel- und Lemmatisierungsstruktur sowie der für Periodisierung und Lokalisierung der Entlehnung verwendeten Begriffe und Angabeformate nicht vollständig kompatibel sein. Auch hinsichtlich der zugrunde gelegten grammatischen Beschreibungssprache kann es Differenzen geben. Der hier vorzustellende Ansatz zur Lösung dieser Probleme stellt insbesondere für wörterbuchübergreifende Suchen eine eigene, relational aufbereitete Datenschicht bereit, die für das Portal relevante Informationen zu allen vorliegenden lexikalischen Einheiten aus den verschiedenen Wörterbüchern in portaleinheitlicher Weise erfasst. In einer wörterbuchübergreifenden Datenbanktabelle werden daher alle Lemmata, alle in den betreffenden Artikeln genannten (diasystematischen, ggf. auch orthographischen) Ausdrucksvarianten der Lemmata sowie sämtliche in Einzelartikeln aufgeführten Derivate

⁶ In der skizzierten Weise wird auch bei dem am Institut für deutsche Sprache entwickelten OWID-Wörterbuchportal verfahren (<http://www.owid.de/index.html>).

und Komposita der Lemmata als je eigene Entitäten – im Folgenden als ‚Portalinstanzen‘ bezeichnet – behandelt, also in jeweils einer separaten Tabellenzeile aufgeführt. Eine Tabellenzeile spezifiziert außer dem Wörterbuch, aus dem die Instanz (also das gegebene Lemma bzw. die gegebene Ausdrucksvariante, das Derivat oder Kompositum) stammt, u.a. folgende weiteren Informationen (Attribute), sofern das Wörterbuch diese zur Verfügung stellt: (a) eine räumliche, zeitliche und diasystematische Einordnung des Entlehnungsvorganges; (b) grammatische Informationen, insbesondere Wortart; (c) ggf. eine semantische/onomasiologische Kategorisierung. Außerdem muss jeweils angegeben werden, ob es sich bei der betreffenden Instanz um die Lemmavariante des zugehörigen Wörterbuchartikels handelt, so dass sich aus der Tabelle der Instanzen die Lemmalisten der Einzelwörterbücher ableiten lassen. Falls ein verwendetes Lehnwörterbuch innerhalb eines Artikels z. B. Lesarten unterscheidet, für die unterschiedliche Etymologien diskutiert werden, sind diese in je separaten Portalinstanzen zu kodieren, da von makrostrukturellen Eigenheiten der Einzelwörterbücher abstrahiert werden muss.

Bei hinreichend komplexer und rigider XML-Kodierung eines Lehnwörterbuchs können die Portalinstanzen weitestgehend automatisiert aus den Originalartikeln extrahiert werden. Die Portalinstanzen sollten keine Informationen aus den Lehnwörterbüchern duplizieren; daher enthalten sie außerdem Verweise auf den zugehörigen Artikel und gegebenenfalls auf das dem relevanten Artikelausschnitt entsprechende XML-Element, so dass sämtliche weiteren für die Instanz relevanten Informationen mechanisch aus dem Ursprungsartikel gewonnen und z.B. für eine HTML-basierte Darstellung aufbereitet werden können. Damit portalweite, wörterbuchübergreifende Suchvorgänge möglich sind, müssen zur Erstellung der Portalinstanzen die Angaben der Ausgangswörterbücher zur zeitlichen und räumlichen Einordnung des Entlehnungsvorgangs sowie grammatische Informationen in ein einheitliches konzeptuelles Schema überführt werden. Neben komplexen Technologien wie Raum- und Zeitontologien stehen für das Pilotprojekt einfachere Lösungen wie die wörterbuchspezifisch definierte Abbildung von Sprachstufenangaben auf standardisierte Jahresintervalle zur Verfügung. Auch der

Einsatz von Georeferenzierungsverfahren ist in einer späteren Ausbaustufe des Projektes denkbar, um kartographische Visualisierungen zu ermöglichen. Wichtig ist, dass Portalinstanzen mit Informationen angereichert werden können, die keinerlei Entsprechung im zugrundeliegenden Lehnwörterbuch haben. So kann jede Instanz einem Synset einer WordNet-artigen Ressource zugeordnet oder anderweitig semantisch klassifiziert werden, um Abfragen mit onomasiologischer Komponente zu ermöglichen. Schwierig ist dies sicherlich besonders in Wortschatzbereichen, aus denen intensiv und bis hin in fachsprachliche Details entlehnt wurde (z. B. Bergbau, Chemie, Religion).

Auch die Einführung von zusätzlichen Portalinstanzen kann sinnvoll sein; ist etwa ein deutsches Wort über das Polnische in das Russische gelangt, kann der womöglich im polnischen Lehnwörterbuch des Portals gar nicht verzeichnete polnische ‚Zwischenschritt‘ als eigene Portalinstanz hinzugefügt werden.

3.3. Metalemmaliste und etymologische Information

Die lexikographisch und linguistisch anspruchsvollste und zum Großteil manuell zu erstellende Datenschicht ist die Erarbeitung einer Metalemmaliste der Etyma der Gebersprache. Da Lehnwörterbücher häufig mehrere diasystematische bzw. Wortbildungsvarianten der Etyma angeben (darunter auch bloß rekonstruierte Formen) und verschiedene mögliche Etymologisierungen diskutieren, muss – auch angesichts der Probleme mit verschiedenen Transkriptionen – ein möglichst allgemeiner Ansatz gewählt werden. In der von uns gewählten Lösung werden für die in den Einzelwörterbüchern genannten Etymonformen – als *tertia comparationis* des umgekehrten Lehnwörterbuchs – jeweils ebenfalls Portalinstanzen angelegt, die in der Datenbanktabelle mit einem speziellen Attribut als (deutsche) Etymonformen gekennzeichnet werden. Im folgenden bezeichnen wir solche Portalinstanzen kurz als ‚Etymoninstanzen‘. Taucht ein deutsches Lexem in mehreren Wörterbüchern als Herkunftswort auf, wird für jedes Wörterbuch eine eigene Etymoninstanz angelegt, da die Identifikation dieser Instanzen ja erst in einem nachgelagerten lexikographischen Arbeitsschritt auf der Portalebene geschieht. Entscheidend ist daher die Identifizierung von

Gruppen „zusammengehöriger“ Etymoninstanzen. In der von uns vorgeschlagenen Datenmodellierung wird für jede solche Gruppe eine wörterbuchunabhängige Etymon-Instanz erstellt, die unter verschiedenen lexikographischen Gesichtspunkten ein besonders geeigneter Kandidat für ein Metalemma ist, also prototypischerweise ein heute noch gebräuchliches, standardsprachliches deutsches Simplex. Dieses ‚Meta-Etymon‘ kann sinnvoll insbesondere in einer Stichwortliste aller deutschen Etyma des Portals verwendet werden. Alle synchronen oder diachronen Varianten, Wortbildungsprodukte/-bestandteile usw. eines Etymons werden dann auf die im folgenden Abschnitt geschilderte Weise mit ihren zugehörigen Meta-Etyma vernetzt. Es kann wünschenswert sein, zusätzliche Meta-Etyma aufzunehmen, etwa, damit der Benutzer zu einem deutschen Simplex auch dann Entlehnungen von daraus gebildeten Komposita findet, wenn dieses Simplex selber in keinem Wörterbuch als Herkunftswort geführt wird.

4. Zur Architektur der Webanwendung

Die Einführung einer Tabelle von Portalinstanzen ermöglicht die saubere Entkopplung der Portalerstellung von der Ebene der Einzelwörterbücher. Typische portalbezogene Suchvorgänge operieren i.a. ausschließlich auf dieser Abstraktionsschicht.

4.1. Kodierung und Verwaltung der Vernetzungen zwischen Portalinstanzen

Portalinstanzen müssen untereinander vernetzt werden, etwa zur Modellierung von Wortbildungsrelationen. Eine besondere Rolle spielen etymologische Angaben, die als Vernetzungen von Portalinstanzen auf Etymoninstanzen kodiert werden können. Der häufigste Fall ist die Vernetzung von Portalinstanzen, die demselben Quellwörterbuch zugeordnet sind. Um Verkettungen von Entlehnungsvorgängen zu modellieren oder ‚Identitätsbeziehungen‘ zwischen Etymoninstanzen sowie zwischen Lemmata in sehr eng verwandten Sprachformen zu formulieren, müssen aber auch Vernetzungen zwischen aus verschiedenen Quellen stammenden Portalinstanzen angesetzt werden.

Zur Modellierung der Vernetzungen zwischen Artikeln und Instanzen könnten im Prinzip standardisierte Repräsentationsformate wie RDF und die dafür

entwickelten Speicher- und Zugriffstechnologien verwendet werden, vgl. (Hitzler, Krötzsch & Rudolph, 2009). Da aber die Vernetzungsstruktur des Portals sehr regelmäßig ist, ziehen wir eine einfachere Lösung vor, die alle Vernetzungen in einer separaten relationalen Datenbanktabelle als geordnete Paare aus einer Quell- und einer Zielinstanz repräsentiert. Jede Vernetzung von Portalinstanz P auf Portalinstanz Q wird per Attribut einem bestimmten Typ zugeordnet; unter anderem sind folgende Typen vorgesehen: (i) P ist Variante von Q (dabei können Varianten verschiedenen Typs unterschieden werden, z.B. orthographisch/synchron/diachron); (ii) P ist Derivat von Q; (iii) P ist Kompositum zu Q; (iv) P hat Q als Etymon; (v) P ist dasselbe Lexem / dieselbe Lexemvariante wie Q (wenn in einer Entlehnungskette das Lehnwort P selber wieder als Grundlage eines Entlehnungsprozesses gedient hat, wird für dieses Lehnwort eine zweite Portalinstanz Q angesetzt, die das Wort in seiner Rolle als Ausgangswort für die weitere Entlehnung repräsentiert); (vi) P gehört im jeweiligen Einzelwörterbuch zum Lemma bzw. Meta-Etymon Q.

Weitere Attribute von Vernetzungen sind die Quelle der Vernetzungsinformation sowie eine einfache, ordinalskalierte Kategorisierung der in der Quelle selber angegebenen Verlässlichkeit dieser Information.

Die Vernetzungen bilden einen gerichteten azyklischen Graphen (DAG). Bei typischen Suchvorgängen müssen im DAG Pfade von ggf. vorab nicht bekannter Länge ermittelt werden – etwa, um Entlehnungsketten zu finden oder ausgehend von einem Meta-Etymon E nach Derivaten/Varianten/... von Entlehnungen beliebiger Derivaten/Varianten/... von E zu suchen. Um performante SQL-Abfragen auf den Tabellen durchführen zu können, wird in der Vernetzungstabelle der transitive Abschluss der Vernetzungsrelationen oder eine geeignete Teilmenge davon abgebildet, d.h. es werden – zumindest auf der Ebene der Meta-Etyma und Einzelwörterbuch-Lemmata – auch ‚indirekte‘ Vernetzungen gespeichert und als solche etikettiert. Die Verwaltung der Verweisstrukturen zwischen den Datenschichten muss softwaregestützt erfolgen.⁷

⁷ Änderungen an den Einzelwörterbüchern ziehen entsprechende Änderungen in den relationalen Instanzen- und Vernetzungstabellen nach sich, die in den meisten Fällen

4.2. Präsentation

Der Benutzer kann die Einzelwörterbücher mit jeweils eigener (neben der Suchformular-/Artikelansicht ausschnittsweise angezeigten) Lemmaliste und Suchfunktionalität nutzen. Die Etymoninstanzen bilden die Grundlage für ein separates umgekehrtes Lehnwörterbuch, also das Portalwörterbuch der deutschen Herkunftswörter, dessen Lemmaliste aus den Meta-Etyma erstellt wird. Suchvorgänge in diesem Portalwörterbuch erzeugen eine Liste von Verweisen auf passende Artikel in den Einzelwörterbüchern.

5. Literatur

- Burnard, L., Bauman, S. (2010): TEI P5: Guidelines for Electronic Text Encoding and Interchange. Text Encoding Initiative. Online: <http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf>.
- Chesley, P., Baayen, R.H. (2010): Predicting new words from newer words: Lexical borrowings in French. *Linguistics* 48 (4), pp. 1343-1374.
- Engelberg, S. (2010): An inverted loanword dictionary of German loanwords in the languages of the South Pacific. In A. Dykstra & T. Schoonheim (Eds.), *Proceedings of the XIV EURALEX International Congress* (Leeuwarden, 6-10 July 2010). Ljouwert (Leeuwarden): Fryske Akademy, pp. 639-647.
- Engelberg, S., Müller-Spitzer, S. (erscheint 2011): Dictionary portals. In R. Gouws, U. Heid, W. Schweickard, & H.E. Wiegand (Eds.), *Wörterbücher / Dictionaries / Dictionnaires. Ein internationales Handbuch zur Lexikographie / An International Encyclopedia of Lexicography / Encyclopédie internationale de lexicographie*. Bd. 4. Berlin, New York: de Gruyter.
- Görlach, M. (Ed.) (2001): *A Dictionary of European Anglicisms: a Usage Dictionary of Anglicisms in Sixteen European Languages*. Oxford etc.: Oxford University Press.
- Hentschel, G. (2009): Intensität und Extensität deutsch-polnischer Sprachkontakte von den mittelalterlichen Anfängen bis ins 20. Jahrhundert am Beispiel deutscher Lehnwörter im Polnischen. In Stolz, Ch. (Ed.): *Unsere sprachlichen Nachbarn in Europa. Die Kontaktbeziehungen zwischen Deutsch und seinen Grenznachbarn*. Bochum: Brockmeyer, pp. 155-171.
- Hitzler, P., Krötzsch, M., Rudolph, S. (2009): *Foundations of Semantic Web Technologies*. Boca Raton, FL etc.: Chapman & Hall/CRC Textbooks in Computing.
- Menzel, T., Hentschel, G., unter Mitarbeit von P. Jančák und J. Balhar (2005): *Wörterbuch der deutschen Lehnwörter im Teschener Dialekt des Polnischen*. *Studia slavica Oldenburgensia*, Band 10 (2003). Oldenburg: BIS-Verlag. 2., ergänzte und korrigierte elektronische Ausgabe. Online: <http://www.bkge.de/14451.html>.
- Müller-Spitzer, C., Schneider, R. (2009): Ein XML-basiertes Datenbanksystem für digitale Wörterbücher. Ein Werkstattbericht aus dem Institut für Deutsche Sprache. *it - Information Technology* 4/2009, pp. 197-206.
- Schenke, M. (2009): Sprachliche Innovation – lokale Ursachen und globale Wirkungen. Das ‚Dynamische Sprachnetz‘. Saarbrücken: Südwestdeutscher Verlag für Hochschulschriften.
- Striedter-Temps, H. (1963): *Deutsche Lehnwörter im Slovenischen*. Wiesbaden: Harrassowitz.
- Vincenz, A. de, Hentschel, G. (2010): *Wörterbuch der deutschen Lehnwörter in der polnischen Schrift- und Standardsprache. Von den Anfängen des polnischen Schrifttums bis in die Mitte des 20. Jahrhunderts*. *Studia slavica Oldenburgensia*, Band 20. Oldenburg: BIS-Verlag. Online: <http://www.bis.uni-oldenburg.de/bis-verlag/wd1p>.
- Wiegand, H.E. (2001): Sprachkontaktwörterbücher: Typen, Funktionen, Strukturen. In: B. Iglá, P. Petkov & H.E. Wiegand (Eds.). *Theoretische und praktische Probleme der Lexikographie*. 1. Internationales Kolloquium zur Wörterbuchforschung am Institut Germanicum der St. Kliment-Ohridski-Universität Sofia, 7. bis 8. Juli 2000 (= *Germanistische Linguistik*, 161-162). Hildesheim, Zürich, New York: Georg Olms Verlag, pp. 115-224.

automatisch durch Datenbanktrigger durchgeführt werden können. Durch solche Trigger können auch Konsistenzprüfungen durchgeführt werden, die manuellen Anpassungsbedarf feststellen und melden.

Localizing A Core HPSG-based Grammar for Bulgarian

Petya Osenova

The Sofia University and ICT-BAS
25 A, Acad. G. Bonchev Str., Sofia 1113
E-mail: petya@bultreebank.org

Abstract

The paper presents the main directions, in which the localization of an HPSG-based Formal Core Grammar (called Grammar Matrix) has been performed for Bulgarian. On the one hand, the adoption process took into account the predefined theoretical schemas and their adequacy with respect to the Bulgarian language model. On the other hand, the implementation within a typological framework posited some challenges with respect to the language specific features. The grammar is being further developed, and it is envisaged to be extensively used for parsing and generation of Bulgarian texts.

Keywords: localization, core grammar, HPSG, Bulgarian

1. Introduction

Recently, a number of successful attempts have been made towards the design and application of wide-coverage grammars, which have incorporated deep linguistic knowledge and have been tested on several natural languages. Especially active in this area have been the lexicalist frameworks, such as HPSG (Head-driven Phrase Structure Grammar), LFG (Lexical-Functional Grammar) and LTAG (Lexicalized Tree Adjoining Grammar). A lot of NLP applications have been performed within HPSG-based implementation – treebanks (the LinGO Redwoods Treebank, Polish HPSG Treebank, Bulgarian HPSG-based Treebank, among others), grammar developing tools, parsers, etc.

In HPSG there already exist quite extensive implemented formal grammars – for English (Flickinger, 2000), German (Muller & Kasper, 2000), Japanese (Siegel, 2000; Siegel & Bender, 2002). They provide semantic analyses in the Minimal Recursion Semantics framework (Copestake et al., 2005). HPSG is the underlying theory of the international initiative LinGO Grammar Matrix (Bender et al., 2010; Bender et al., 2002). At the moment, precise and linguistically motivated grammars, customized on the base of the Grammar Matrix, have been or are being developed for Norwegian, French, Korean, Italian, Modern Greek, Spanish, Portuguese,

etc.¹. The most recent developments in the Grammar Matrix framework report also on successful implementation of grammars for endangered languages, such as Wambaya (Bender, 2008).

In addition to the HPSG framework and the Grammar Matrix architecture, there is also an open source software system, which support the grammar and lexicon development – LKB (Linguistic Knowledge Builder) (<http://wiki.delph-in.net/moin/LkbTop>)².

Our motivation to start the development of a Bulgarian Resource Grammar in the above-mentioned setting was as follows: there already was an HPSG-based Treebank of Bulgarian (BulTreeBank), constructed in a semi-automatic way. The knowledge within the treebank seemed to be sufficient for the construction of a wide coverage and precise formal grammar, which to parse and generate Bulgarian texts. Bulgarian is considered neither an endangered language, nor a less-processed language any more. However, it still lacks a deep linguistic grammar. Bulgarian is viewed as a “classic and exotic” language, because it combines Slavic features with Balkan Sprachbund peculiarities. These factors make Bulgarian a real challenge for the computational modeling.

¹ <http://www.delph-in.net/index.php?page=3>

² The projects DELPH-IN and Deep Thought are also closely related to the Grammar Matrix initiative.

Our preliminary supporting components were the following ones: the HPSG theoretical framework for modeling the linguistic phenomena in Bulgarian; a suitable Bulgarian corpus, which is HPSG-based, and supporting pre-processing modules; the LinGO Matrix-based Grammars software environment for encoding and integrating the suitable components; the best practices from the work on other languages. More on the current grammar model and implementation of the Bulgarian Grammar can be read in (Osenova, 2010).

2. Grammar Matrix Architecture

The Grammar Matrix (Bender et al., 2002) has been intended as a typological core for initiating the grammar writing on a specific language. It also provides a customization web interface (Bender et al., 2010). The purpose of such a core is, on the one hand, to ensure a common basis for comparing various language grammars, and thus – to focus on typological similarities and differences, and on the other hand, to speed up the process of the grammar development. Thus, it supplies the skeleton of the grammar – the type hierarchy with basic types and features as well as the basic inheritance directions. Grammar Matrix is based on the experience with several languages (predominantly English and Japanese), and it is being developed further when new languages are modeled in the framework.

In spite of supporting all the linguistic levels of representation, the Grammar Matrix aims at semantic modeling of a language. It introduces referential entities and events; semantic relations; semantic encoding and contribution of the linguistic phenomena (definiteness/indefiniteness; aspect; tense, among others). For example, the verbs, the adjectives, the adverbs and the prepositions are canonically viewed as introducing events, while nouns are considered introducing referential entities. Such an approach is a challenge for a language like Bulgarian, which grammaticalizes a lot of linguistic phenomena. Thus, the most common level of description would be the morphosyntactic level rather than the semantic one. Consequently, the balance of represented information between semantics and morphosyntax should be detected and distributed in an adequate way. Ideally, one should only inherit from Grammar Matrix types, without changing them. In real life, however, it turns out that each language challenges, and is challenged

by the Matrix model. On the one hand, Matrix predefines some phenomena too strictly, on the other – it gives possibilities for generalizations. All this is inevitable, since the ideal granularity between specificity and universality is difficult to be established.

The localization goes into several directions. First, the Grammar Matrix is implemented in accordance with some version of the HPSG theory – thus it implies certain decisions with respect to the possible analyses. However, the grammar developer in adapting the Grammar Matrix to a new language might want to apply another analysis within the language specific grammar. This is the case for Portuguese, for example. Instead of working with *head-specifier* and *head-adjunct* phrases, which are part of the standard HPSG94, the grammar adopted the more recent *head-functor* approach to these phrases. Another direction would be the preference towards the linguistic phenomena. Thus, in Portuguese the preferences concern agreement, modification and basic phrase structures, while in Modern Greek the phenomena to start with were cliticization, word order, politeness constructions. In this respect, only a common testset might ensure the implementation of common linguistic phenomena. Such a testset is briefly discussed in 3.1. Thus, depending on the preference, grammar developers might have to extend and/or change the core grammar. For example, the addition of types for contracted or missing determiners in Modern Greek, since this information influences the semantics.

Last, but not least, it is up to the grammar developer how much information to encode within the grammar, and which steps to be manipulated outside the grammar. For example, the Portuguese grammar uses a morphologically preprocessed input, while in Modern Greek grammar all the analyses are handled within the system.

3. Localization in Bulgarian

3.1. The Multilingual Testset.

The Grammar Matrix is equipped with a testset in English, which has been already translated into a number of other languages. It comprises around 100 sentences, which in the Bulgarian translated set became 178. The grammar development started with the aim this set to be covered, since it represented some very important

common phenomena. Needless to say, the translated set incorporated also a bunch of language specific phenomena, which will be discussed in more detail below. Thus, some additional test sentences have been incorporated into the common testset, which made the positive sentences 193. Also, 20 ungrammatical sentences have been included, which checked the agreement, word order of clitics, definiteness, subject control, etc. The whole set is 213 sentences, which is comparable to the testset for Portuguese in the first phase of the grammar development. The common phenomena are as follows: complementation, modification, coordination, agreement, control, quantification, negation, illocutionary force, passivization, nominalization, relative clauses, light verb constructions, etc. The types in the initial grammar are 297. It is expected that they will expand dramatically when the lexicon is enriched further. Let us comment on some localization specificities in the translated set, which made it larger in comparison to the English testset.

First of all, Bulgarian is a pro-drop language. Thus, it has always counterparts with null subjects. In the discourse, it can also omit its complements in many cases. Second, Bulgarian verbs encode aspect lexically. The English sentences often have been translated with verbs in both aspects (perfective or imperfective). When combined with the tense, the translation counterparts became even more. For example, the sentence *Abrams wondered which dog barked* might have two possibilities for the matrix verb (imperfect tense, imperfective and aorist tense, perfective), while the verb in the subordinate clause might have normally three possibilities (present tense, imperfective; aorist tense, perfective and imperfect tense, imperfective).

In some sentences more Bulgarian verb synonyms have been provided to the English one. For example, the verb *to hand* in the sentence *Abrams handed the cigarette to Browne* can be translated into at least four Bulgarian verbs – *дам* (give), *подам* (pass), *връча* (deliver), *предам* (hand in).

Next, Bulgarian has clitic counterparts to the complements as well as a clitic reduplication mechanism. Thus, translations with a clitic and a full-fledged complement have been provided to the single English one, when appropriate. Bulgarian polar questions are formed with a special question particle, which has also a

focalizing role. The modification is mostly done by the adjectives – *garden dog* (en) vs. *градинско куче* (bg, ‘garden-adjective dog’). Some alternations that are challenging for English are not relevant for Bulgarian. For example: *Browne squeezed the cat in* and *Browne squeezed in the cat* are translated in the same way: *Браун вмъкна котката* (Brown put-inside cat-the). The same holds for the well-known give-alternation: *Abrams handed Browne the cigarette* and *Abrams handed the cigarette to Browne*. The Bulgarian translation just ‘swaps’ the complements, but does not change them: *Абрамс даде на Браун цигарата* (Abrams gave to Brown cigarette-the) and *Абрамс даде цигарата на Браун* (Abrams gave cigarette-the to Brown). At the same time, the Bulgarian version of the testset provided examples for aspect/tense combinations, clitic behavior, verbal complex, agreement patterns, etc.

3.2. The Language Specific Phenomena

Concerning Bulgarian, its rich morphology seems to conflict with the requirements behind the semantic approach. Thus, the information has to be often split between the semantic phenomenon and its realization. For example, the adjectives, participles, numerals happen to have *morphologically* definite forms, while the definiteness marker is not a *semantic* property of these categories. For that reason, the most important thing in the grammar was to keep Syntactic and Semantic features separate (for example, agreement, which is separated into semantic and syntactic ones in accordance with the ideas in Kathol 1997). In this way, the definiteness operates via the **MOD(ifier)** feature. The event selects for a semantically definite:

[SYNSEM.LOCAL.HOOK.INDEX.DEF+],

but morphologically indefinite noun:

[SYNSEM.LOCAL.AGR.DEF-]

As it can be seen, the semantic feature ‘definiteness’ lies in the syntactic-semantic area of local features, and more precisely within the feature **INDEX**. The morphosyntactic one follows the same path of locality, but it is within the feature **AGR(ement)**. For example, in the phrase *старото куче* ‘old-the dog’, the adjective ‘old-the’ selects for the semantically definite, but morphologically indefinite noun ‘dog’. The analysis is linguistically sound, since the definiteness marker is considered a phrasal affix in Bulgarian, not a word one.

Other examples are the categories of tense, aspect, mood. Tense and Mood are currently encoded as a feature of **AGR.E**³.**TENSE** or **AGR.E.MOOD**, while aspect is a feature of the head **HEAD.TAM**⁴.**ASPECT**. However, in these cases at the moment there is no different contribution from semantics and morphosyntax. Thus, Grammar Matrix provides several possibilities to get the semantic information. For tense and mood the aggregated one has been chosen (**AGR.E**) in the current version, while for aspect – the separated encodings. The aggregated way is a better choice for unified syntactic-semantic analysis, while the separated representation leaves out an opportunity for different manipulation of syntactic and semantic contribution.

Thus, Bulgarian seems to require a systematic balance between the semantic contribution and the morphological marking of the same category within the overall architecture. This fact posited some difficulties in the starting design, since the categories had to be considered whether to be approached separately on both - semantic and morphological grounds, or not.

Bulgarian has a double negation mechanism (the so-called negative concord) similarly to other Slavic languages and in contrast to English. Within the proposed Grammar Matrix architecture, the negation particle has been modeled as a verb, since particles had not been presented in the Grammar Matrix, and there was no mechanism of introducing semantic relations. It scopes over the following proposition, and introduces a negation relation. At the same time the negative pronoun in the concord introduces a negative relation.

Another area, in which the rich morphology plays role, is the level of type's generalization. Very often, in Bulgarian the generalization cannot be kept at higher levels, because of the variety in the morphosyntactic behaviour types within the Bulgarian constructions. Such examples are the copula constructions. Although adjectives, adverbs and prepositions have an *event* index, they cannot share the same generalized type. Adjectives structure-share their **PNG** (person, number and gender) characteristics with the copula's **XARG** – the subject. The adverbs have to be restricted to intersective

modifiers when taken as complements. The common behaviour is that all these heads raise their semantic index to the copula, which is semantically vacuous itself. The nouns, however, have a referential index. In this case, the copula behaves like a transitive verb, which selects for its complement. No index is raised from the noun complement up to the copula. In this grammar version, 8 lexical types are introduced: two for present and past copula forms. Each of the two then is divided into four subtypes depending on the complement (present copula – noun; present copula – adjective; present copula – adverb; present copula - PP; past copula – noun; past copula – adjective; past copula – adverb; past copula - PP). The *present-past* distinction was necessary, because the past form can be in a sentence initial position, while the present form cannot.

Localization took into account the relatively free word order of Bulgarian. Thus, most of the rules include all the possible orders in spite of the canonical readings. For example, there are rules for *head-modifier* and *modifier-head*; *clitic-head* and *head-clitic*; also for the head's complement swap. The order combinations result into a proliferation of possible analyses, for whose discrimination an additional mechanism is needed. For the moment, the BulTreeBank resource is used as a discriminative tool, because it comprises the canonical and most preferred analysis per sentence.

Combining the application of the clitic rules which produce lexical signs, and the complement rules, which produce phrases, the clitic doubling examples have been successfully parsed. The incorporation of Bulgarian argument-related clitics required a new mechanism. The clitics are viewed as lexical projections of the head (i.e. operated by special rules), while the regular forms are treated as head arguments (complements) (i.e. operated by head-complement principles). The clitic does not contribute its separate semantics, because it is not a full-fledged complement. Instead, the verb incorporates clitic's contribution in its own semantics. Thus, the personal pronoun clitic lexemes have an empty relation list, while the regular pronoun forms have a pronoun relation.

Another localization, which reflects the modeling of the lexicon rather than the type hierarchy, is the representation of the lexical entries. Bulgarian is a rich-inflected language, but in contrast to other Slavic

³ **E** stands for Event.

⁴ **TAM** stands for an aggregate feature Tense, Aspect, Mood.

languages, its richness lies in the verbal system, rather than in the nominal one. Thus, two ways of morphology incorporation were possible. The first is to re-design the whole systematic and unsystematic morphology within the grammar, which would be a linguistically sound, but time-consuming step. Since Bulgarian verbs show a lot of alternations and irregularities across their grammatical categories (conjugation, tense, aspect, finite vs. infinite forms, other synthetic grammatical categories, such as imperative, etc.), the full paradigms per conjugation in the lexical types were abandoned as a generalization opportunity. Instead, the inflection classes of the morphological dictionary for Bulgarian (Popov et al., 2003), have been transferred into the grammar. Each verb type was viewed as a combination of the appropriate subparadigms from the given morphological and/or lexical categories. The set of the respective subparadigms per category was attached to each verb in the lexicon. Thus, the lexicon was also “dressed” with the morphologically specific information for the distinct verbs. The transfer of the morphosyntactic paradigms resulted into over 2600 rules for personal verbs only. Hence, the morphological work has been suppressed in the name of syntactic and semantic modeling. Also, in a lexicalist framework, such as HPSG, a large lexicon could not operate without the complete set of the morphosyntactic types and rules. Compare the morphologically poor and morphologically rich presentation of the verbs in the lexicon in *a.* and *b.*:

a.

```
ima_v1 := v_there-is_le &
[ STEM <"има">,
  SYNSEM.LKEYS.KEYREL.PRED "има_v_1_rel" ].
```

b.

```
ima_v1 := v_there-is_le &
[ STEM <"има">,
  SYNSEM [ LKEYS.KEYREL.PRED "има_v_1_rel",
            LOCAL.CAT.HEAD.MCLASS
[ FIN-PRESENT finite-present-101,
  FIN-AORIST finite-aorist-080,
  FIN-IMPERF finite-imperf-025,
  PART-IMPERF participle-imperf-024,
  PART-AORIST participle-aorist-095] ] ].
```

In case *a.* the impersonal verb *има* ‘there is’ introduces its type from which inherits the specific template

(*v_there-is_le*). Then it presents the stem, i.e. word itself, and the relation. In case *b.* there is also a morphological class (**MCLASS**), which is augmented with the respective paradigms for the relevant grammatical categories. The second one is maintained in the grammar development.

The evaluation of the current grammar version was done within the system [tdsb] (Oepen, 2001). The coverage of the first version of the grammar is as follows: 213 sentences, from which 193 grammatical ones. The average of distinct analyses is 3.73. The ambiguity of analyses is mainly due to the following factors: 1. morphological homonymy of the word forms; 2. more than one possible word order; 3. more than one possible attachment; 4. competing rules in the grammar (see more in Osenova & Simov, 2010). The first one concerns forms like the word form of the verb ‘come’: *дойде*, which is ambiguous between present tense and aorist, 2nd or 3rd person. The second one has to do with cases like: *The dog chases Brownie*, where Brownie also might be the subject in some reading. The third one considers the attachment of adjuncts at the verb level as well as at the sentence level. The last factor affects mostly the coordination rules, but also some rules for modification, where the split due to the specificities of a head allow for duplication in the remaining cases. Factor 1 requires an external disambiguation filter, which is typically done by taggers. Factor 2 also requires an additional filter to pick up the most typical reading without excluding the rest. Factor 3 considers spurious cases and requires a linguistic decision for the various types of adjuncts. Factor 4 needs a change in the grammar architecture by the grammar writer.

The Grammar Matrix based representation, in which the Bulgarian Resource Grammar should be compatible with the other grammars on a semantic level, is MRS.

4. Conclusions and Future Work

The existence of a Core Grammar proved out to be very useful in the initial steps of the grammar writing, and not only, since it provides the typological background to start with and to maintain compatibility with the other language descriptions. At the same time, depending on the purposes and tasks, the grammar writer has the possibility to re-model or even override some parts within the preliminary structure. Such developments

would give feedback to the Core Grammar developers, and would contribute to better generalizations over more languages.

The Bulgarian Resource Grammar together with the English Resource Grammar are envisaged to be used for the purposes of Machine Translation within the context of European EuroMatrixPlus project. We are using the infrastructure established within DELPH-IN and LOGON. For this task MRSes of the parallel sentences, parsed by both grammars, have been aligned on lexical and phrasal level. Transfer rules are being defined on the basis of this alignment. For example, in the alignment of MRSes for the sentence *No cat barked* in Bulgarian there will be an additional negation relation, coming from the negated verb. Otherwise, the arguments of the first negation relation coincide as well as the argument structures of the intransitive verbs. At the same time a set of valency frames for 3000 have been extracted from BulTreeBank, and will be added to the grammar lexicon. Additionally, the arguments in the valency frames have been assigned ontological classes. This step will help in selecting only one possible analysis in cases like: *John read a book* (*John* as subject and a *book* as a complement), and keeping the two possible analyses in cases like: *Abrams chased a dog* (*Abrams* as subject or complement, and the same for a *dog*).

5. Acknowledgements

The work in this paper has been supported by the Fulbright Foundation and the EU project EuroMatrix+. It profited a lot from the collaboration with Dan Flickinger (Stanford University). The author would like to thank Kiril Simov (IICT-BAS) for his valuable comments on the earlier drafts of the paper, and also the two anonymous reviewers for the very useful critical reviews.

6. References

- Bender, E. M., Drellishak, S., Fokkens, A., Poulson, L., Saleem, S. (2010): Grammar Customization. In: Research on Language and Computation, vol. 8 (1), pp. 23-72.
- Bender, E., Flickinger, D., Good, J., Sag, I. (2004): Montage: Leveraging Advances in Grammar Engineering, Linguistic Ontologies, and Mark-up for the Documentation of Underdescribed Languages. In Proceedings of the Workshop on First Steps for Language Documentation of Minority Languages: Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation, LREC 2004, Lisbon, Portugal.
- Bender, E. (2008): Evaluating a Crosslinguistic Grammar Resource: A Case Study of Wambaya. In Proceedings of ACL08: HLT, Columbus, OH.
- Copestake, A., Flickinger, D., Pollard, C., Sag, I. (2005): Minimal Recursion Semantics: An Introduction. In Research on Language and Computation (2005) 3, pp. 281–332.
- Flickinger, D. (2000): On building a more efficient grammar by exploiting types. In: Natural Language Engineering, 6 (1) (Special Issue on Efficient Processing with HPSG), pp. 15-28.
- Kathol, A. (1997): Agreement and the Syntax-Morphology Interface in HPSG. In R. Levine and G. Green (eds.) Studies in Current Phrase Structure Grammar. Cambridge University Press. pp. 223-274.
- Muller, S., Kasper, W. (2000): HPSG analysis of German. In W. Wachsler (ed.), *Verbmobil*. Foundations of speech-to-speech translation. (Artificial Intelligence ed., pp. 238-253). Berlin, Germany: Springer.
- Oepen, St. (2001): [incr tsdb()] – competence and performance laboratory. User manual. Technical Report, Saarland University, Saarbruecken, Germany.
- Osenova, P. (2010): Bulgarian Resource Grammar. Modeling Bulgarian in HPSG. Verlag Dr. Muller, pp. 71.
- Osenova, P., Simov, K. (2010): Using the linguistic knowledge in BulTreeBank for the selection of the correct parses. In: TLT Proceedings, pp. 163-174.
- Popov, D., Simov, K., Vidinska, Sv., Osenova, P. (2003): Spelling Dictionary of Bulgarian language. Nauka i izkustvo. Sofia, 2003. (in Bulgarian)
- Siegel, M. (2000): HPSG Analysis of Japanese. In: W. Wahlster (ed.): *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer Verlag.
- Siegel, M., Bender, E. (2002): Efficient deep processing of Japanese. In Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization, Taipei, Taiwan.

Poster Presentations

Autorenunterstützung für die Maschinelle Übersetzung

Melanie Siegel

Acrolinx GmbH

Rosenstr. 2, 10178 Berlin

E-mail: melanie.siegel@acrolinx.com

Abstract

Der Übersetzungsprozess der Technischen Dokumentation wird zunehmend mit Maschinellem Übersetzung (MÜ) unterstützt. Wir blicken zunächst auf die Ausgangstexte und erstellen automatisch prüfbare Regeln, mit denen diese Texte so editiert werden können, dass sie optimale Ergebnisse in der MÜ liefern. Diese Regeln basieren auf Forschungsergebnissen zur Übersetzbarkeit, auf Forschungsergebnissen zu Translation Mismatches in der MÜ und auf Experimenten.

Keywords: Machine Translation, Controlled Language

1. Einleitung

Mit der Internationalisierung des Markts für Technologien und Technologieprodukte steigt die Nachfrage nach Übersetzungen der Technischen Dokumentation. Vor allem in der Europäischen Union steigt das Bewusstsein, dass es nicht ausreicht, englischsprachige Dokumentation zu liefern, sondern dass Dokumentation in die Muttersprache der Kunden übersetzt werden muss. Diese Übersetzungen müssen schnell verfügbar, aktualisierbar, in mehreren Sprachen gleichzeitig verfügbar und von hoher Qualität sein. Gleichzeitig gibt seit einigen Jahren erhebliche technologische Fortschritte in der Maschinellen Übersetzung: Es gibt regelbasierte¹ und statistische Systeme², aber auch hybride Übersetzungsverfahren³. Diese Situation hat dazu geführt, dass Firmen mehr und mehr versuchen, ihre Übersetzungsanstrengungen mit MÜ zu unterstützen. Dabei treten allerdings eine Reihe von Problemen auf. Die Nutzer kennen die Möglichkeiten und Grenzen der MÜ nicht gut genug. Sie werden in ihren Erwartungen enttäuscht.

Um die Systeme zu testen, werden völlig ungeeignete Texte übersetzt, wie z. B. Prosa⁴.

Auch Technische Dokumentation, die an die MÜ geschickt wird, ist oft nicht von ausreichender Qualität, ebenso wenig wie Texte, die an humane Übersetzer geschickt werden. Allerdings können humane Übersetzer diesen Mangel an Qualität im Ausgangsdokument ausgleichen, während MÜ-Systeme dazu nicht in der Lage sind.

Statistische MÜ-Systeme müssen auf parallelen Daten trainiert werden. Oft werden dafür TMX-Dateien verwendet, die aus Translation Memory – Systemen herausgezogen werden. Da aber diese Daten oft unsauber sind und fehlerhafte und inkonsistente Übersetzungen enthalten, ist auch die Qualität der trainierten Übersetzung schlecht.

Wir haben uns mit der Frage beschäftigt, wie die Autoren Technischer Dokumentation darin unterstützt werden können, Dokumente für die MÜ optimal vorzubereiten, um auf diese Weise optimale Übersetzungsergebnisse zu bekommen. Das Ziel der Untersuchungen ist, die Möglichkeiten und Grenzen der MÜ genauer zu spezifizieren, daraus Handlungsoptionen für Autoren abzuleiten und diese durch automatische Verfahren zu unterstützen. Dabei gehen wir in drei Schritten vor:

- 1) Wir untersuchen die Schwierigkeiten, die ein humaner Übersetzer hat, darauf, ob sie auf MÜ-Systeme übertragbar sind.
- 2) Wir experimentieren mit automatisch prüfbaren

¹ Z.B. das System Systran (<http://www.systran.de/>), das aber jetzt auch mit statistischen Verfahren angereichert wird (Callison-Burch et al. 2009)

² Z.B. das System Moses (Koehn, 2009; Koehn et al., 2007) oder google translate (translate.google.com)

³ Z.B. Federmann et al., 2010.

⁴ Beispiel hier: Saarbrücker Zeitung vom 6.10.2009: "Vom

Leid mit der Übersetzung", von Michael Brächer. Test hier mit Auszügen aus Goethes „Erlkönig“.

Regeln der Autorenunterstützung und übersetzen Texte vor und nach der Umformulierung mit MÜ.

- 3) Wir ziehen Untersuchungen zu „Translation Mismatches“ in der MÜ heran, um Strukturen zu finden, die besonders schwer automatisch übersetzbar sind.

2. Schwierigkeiten von humanen Übersetzern – Schwierigkeiten von MÜ-Systemen

Heizmann (1994:5) erläutert den Übersetzungsprozess für humane Übersetzer: *"In our opinion, translation is basically a complex decision process. The translator has to base his or her decisions upon available information, which he or she can get from various sources."* Diese Aussage ist auch auf den Übersetzungsprozess in der MÜ übertragbar und verdeutlicht schon, dass es notwendig ist, der Maschine möglichst wenige komplexe Entscheidungsprozesse aufzubürden.

Ausgehend davon, dass ein MÜ-System einem eher unprofessionellen Übersetzer ähnlich ist, dem die Texte für die Übersetzung so vorbereitet werden sollten, dass sie einfacher übersetzbar sind, ziehen wir Parallelen vom unprofessionellen Übersetzer zum MÜ-System. Der Ausgangstext für Übersetzer wie für ein MÜ-System muss so angepasst werden, dass die Probleme möglichst umgangen werden, die der unprofessionelle Übersetzer und das MÜ-System haben:

Die Übersetzung einzelner Wörter, Phrasen und Sätze, ohne die Möglichkeit, größere Übersetzungseinheiten in Betracht zu ziehen, erfordert, dass satzübergreifende Bezüge vermieden werden müssen, wie z.B. Anaphern.

Die Unmöglichkeit der Paraphrasierung erfordert einfache Satzstrukturen ohne Ambiguitäten. Wichtig ist es auch, metaphorische Sprache zu vermeiden, da diese oft nicht einfach übersetzt werden kann, sondern Paraphrasierung erfordert.

Eine Übersetzung ohne Weltwissen führt dazu, dass Wörter mit unterschiedlichen Bedeutungen in verschiedenen Domänen (Homonyme) falsch übersetzt werden. Solche potentiell ambigen Wörter müssen vermieden werden.

Da das Spektrum von Übersetzungsvarianten potentiell größer als bei professionellen Übersetzern ist, ist eine systematische Terminologiearbeit am Ausgangstext hilfreich, die Terminologievarianten im Ausgangstext

schon mal eliminiert.

Da die MÜ ebenso wie der unprofessionelle Übersetzer wenige Hilfsmittel hat, die Hintergrundwissen zum beschriebenen Sachverhalt geben, muss die Beschreibung möglichst klar und verständlich sein. Das erfordert einfache Satzstrukturen.

3. Relevanz von automatisch prüfbar Regeln der Autorenunterstützung

In einem Experiment haben wir einige Dokumente der technischen Dokumentation mit dem MÜ-System Langenscheidt T1 übersetzen lassen. Danach haben wir die Dokumente mit einer großen Anzahl automatisch prüfbarer Regeln aus Acrolinx IQ geprüft. Die Ergebnisse der Prüfungen haben wir umgesetzt, indem wir die Ausgangstexte umformuliert haben. Diese umformulierten Texte haben wir dann wieder mit Langenscheidt T1 automatisch übersetzt und die Übersetzungen miteinander verglichen. Das Ziel dieses Experiments ist es, herauszufinden, welche Regeln der Autorenunterstützung wichtige Effekte auch für die MÜ haben. Einige dieser Regeln haben wir im vorangegangenen Abschnitt Schwierigkeiten von humanen Übersetzern – Schwierigkeiten von MÜ-Systemen schon vorgestellt. Aufgrund dieser Experimente haben wir ein Regelset zusammengestellt, das wir im nächsten Abschnitt vorstellen.

4. Erste Ergebnisse der Experimente

Rechtschreibung und Grammatik: Das Regelset für die deutschen Ausgangstexte enthält zunächst die Standard-Grammatik- und Rechtschreibregeln. Die Experimente haben klar gezeigt, dass ein MÜ-System keine sinnvollen Ergebnisse liefert, wenn der Eingabetext Rechtschreib- und Grammatikfehler enthält. Wenn ein Wort unbekannt ist, weil es falsch geschrieben ist, dann ist auch keine Übersetzung mit dem MÜ-System möglich. Allerdings führt nicht jeder Rechtschreibfehler auch zu Übersetzungsproblemen: Die Experimente haben gezeigt, dass das untersuchte MÜ-System tolerant zu alter und neuer deutscher Rechtschreibung ist – beide Varianten „muß“ und „muss“ wurden korrekt übersetzt.

Regeln zu Formatierung und Zeichensetzung: Der Gebrauch von Gedankenstrichen führt zu komplexen Sätzen im Deutschen, die Probleme bei der Übersetzung bereiten.

Regeln zum Satzbau: Beim Satzbau geht es zunächst darum, komplexe Satzstrukturen zu vermeiden. Oberstes Gebot ist hier, zu lange Sätze zu vermeiden. Komplexe Satzstrukturen entstehen durch die folgenden Konstruktionen, wie Einschübe, Hauptsatzkoordination, Trennung von Verben, eingeschachtelte Relativsätze, Schachtelsätze, Klammern, Häufung von Präpositionalphrasen, Beschreibung mehrerer Handlungen in einem Satz, umständliche Formulierungen und Bedingungssätze, die nicht mit „wenn“ eingeleitet sind. Ein anderes Problem für die MÜ sind ambige Strukturen, die durch Substantivkonstruktionen und elliptische Konstruktionen entstehen.

Regeln zur Wortwahl: Füllwörter und Floskeln sind deshalb schwierig für die MÜ, weil nicht paraphrasiert werden kann. Das MÜ-System versucht, diese Wörter zu übersetzen, obwohl ein professioneller Übersetzer sie weglassen oder umformulieren würde. Umgangssprache und bildhafte Sprache sind ebenfalls ein großes Problem. Pronomen sind dann schwierig zu übersetzen, wenn der Bezug außerhalb des Satzkontexts liegt und unklar ist. Bei der Verwendung von ambigen Wörtern kann das MÜ-System in vielen Fällen die Ambiguität nicht auflösen. Das passiert zum Beispiel bei der Verwendung von Fragewörtern in anderen Kontexten als einer Frage. Gerade ausdruckschwache Verben mit ambigem Bedeutungsspektrum sind problematisch. Der Nominalstil, bei dem Verben nominalisiert werden, kann im Englischen zu komplexen und falschen Konstruktionen führen.

5. Anwendung der Regeln, Umformulierungen und Übersetzungen

Ein wichtiger Teil der Fragestellung war aber nun, ob die Anwendung der implementierten Regeln zur Autorenunterstützung tatsächlich eine Auswirkung auf die Ergebnisse der MÜ hat. Im oben beschriebenen Experiment haben wir die aufgestellten und implementierten Regeln zur Autorenunterstützung auf zwei Dokumente angewendet und die Texte nach den Empfehlungen der Regeln umformuliert. Anschließend haben wir untersucht, welche der Regeln am häufigsten auftraten und die meisten Effekte für die Qualität der MÜ-Ausgaben hatten. Hier muss jedoch angemerkt werden, dass dieses Experiment bisher nur mit zwei

Dokumenten durchgeführt wurde, einer Anleitung zum Ausbau von Zündkerzen am Auto und einer Anleitung zur Installation einer Satellitenschüssel. Ein interessantes Ergebnis: In fast der Hälfte der Fälle konnte der Satz anhand von lexikalisch-basierten Regeln so verbessert werden, dass die Maschinelle Übersetzung gute Ergebnisse lieferte.

6. Untersuchungen zu Translation Mismatches und daraus resultierende Empfehlungen

Kameyama et al. (1991) verwendeten den Begriff "Translation Mismatches", um ein Schlüsselproblem der maschinellen Übersetzung zu beschreiben. Bei Translation Mismatches handelt es sich um Information, die in der einen am Übersetzungsprozess beteiligten Sprache explizit nicht vorhanden ist, die aber in der anderen beteiligten Sprache gebraucht wird. Der Effekt ist, dass die Information in der einen Übersetzungsrichtung verloren geht und in der anderen hinzugefügt werden muss. Das hat - wie Kameyama beschreibt - zwei wichtige Konsequenzen:

“First in translating a source language sentence, mismatches can force one to draw upon information not expressed in the sentence - information only inferrable from its context at best. Secondly, mismatches may necessitate making information explicit which is only implicit in the source sentence or its context.” (S.194)

Translation Mismatches sind für die Übersetzung eine große Herausforderung, weil Wissen, das nicht direkt sprachlich kodiert ist, inferiert werden muss. Welche Translation Mismatches relevant sind, das hängt aber stark von der Information ab, die in den beteiligten Sprachen kodiert ist. Für das Sprachpaar Deutsch-Englisch konnten wir in den Experimenten die folgenden Translation Mismatches identifizieren:

Lexikalische Mismatches. Die Bedeutung ambiger Wörter in der Ausgangssprache muss in der Zielsprache aufgelöst werden, wie z.B. bei „über“ -> „about“, „above“.

Nominalkomposita. Nach den Regeln der deutschen Rechtschreibung müssen Nominalkomposita entweder zusammen oder mit Bindestrich geschrieben werden. Wenn sie zusammengeschrieben werden, muss die Analyse der MÜ die Teile identifizieren. Das ist aber nicht immer eindeutig im Deutschen. Wenn andererseits

auch im Deutschen wie im Englischen ein Leerzeichen zwischen den Teilen des Kompositums steht, dann ist die MÜ-Analyse überfordert, weil die Beziehung zwischen den Nomen unklar bleibt. Z.B.: „bei den heutzutage verwendeten Longlife Kerzen“ - „at the nowadays used ones“

Metaphorik. Bildhafte Sprache lässt sich nicht wörtlich übertragen. Ein Beispiel aus den Experimenten: „Man ist daher leicht geneigt“ – „One is therefore slightly only still to“

Pronomen. Das Pronomen „Sie“ meint im Deutschen sowohl die 3. Person Singular als auch die 2. Person Singular, abhängig von der Großschreibung. Wenn das „Sie“ aber am Satzanfang steht, bleibt unklar, welche Variante gemeint ist. Beispiel: „Sie haben es fast geschafft“ – „her it have created almost“.

7. Zusammenfassung und nächste Schritte

Wir haben ein Regelset für die automatische Autorenunterstützung aufgestellt. Dieses Regelset basiert auf Untersuchungen zu Problemen humaner Übersetzer, auf Experimenten mit MÜ und Umformulierungen und auf Untersuchungen zu Translation Mismatches in der MÜ. In einem nächsten Schritt haben wir das entstandene Regelset in Experimenten mit verschiedenen MÜ-Systemen validiert. Die Übersetzungen wurden dieses Mal von professionellen Übersetzern und Übersetzerinnen validiert. Eine erste Auswertung der Validierungen ergab:

- Umformulierungen durch Regeln hatten keinen Einfluss auf das Ranking der Ergebnisse verschiedener MÜ-Systeme.
- Die Anzahl der klassifizierbaren Fehler der MÜ-Systeme steigt, während die Anzahl der nicht klassifizierbaren Fehler sinkt. Übersetzungen der umformulierten Texte enthalten weniger Grammatikfehler.
- Die Anzahl der korrekten Übersetzungen steigt stark.

Die Regeln für das Pre-Editing können zum Teil automatische Vorschläge für die Umformulierung geben. Wir suchen nach einem Weg, aus diesen Vorschlägen ein automatisches Pre-Editing zu erzeugen.

8. Acknowledgements

Dieses Vorhaben wird durch die TSB

Technologiestiftung Berlin aus Mitteln des Zukunftsfonds des Landes Berlin gefördert, kofinanziert von der Europäischen Union – Europäischer Fonds für Regionale Entwicklung. Investition in Ihre Zukunft!

9. Literatur

- Callison-Burch, C., Koehn, P., Monz, C., Schroeder, J. (2009): Findings of the 2009 Workshop on Statistical Machine Translation. In Proceedings of the Fourth Workshop on Statistical Machine Translation (WMT09), March.
- Drewer, P., Ziegler, W. (2011): Technische Dokumentation. Übersetzungsgerechte Texterstellung und Content-Management. Vogel-Verlag Würzburg.
- Federmann, C., Eisele, A., Uszkoreit, H., Chen, Y., Hunsicker, S., Xu, J. (2010): Further Experiments with Shallow Hybrid MT Systems. In: Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Zaidan, O. (eds.): Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, Pages 77-81, Uppsala, Sweden, ACL, Association for Computational Linguistics (ACL), 209 N. Eighth Street Stroudsburg, PA 18360 USA, 7/2010
- Heizmann, S. (1994): Human Strategies in Translation and Interpreting - what MT can Learn from Translators. Verbomobil Report 43. Universität Hildesheim.
- Kameyama, M., Ochitani, R., Peters, S. (1991): Resolving Translation Mismatches With Information Flow. In: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley: 193-200.
- Klausner, K. (2011): Einsatzmöglichkeiten kontrollierter Sprache zur Verbesserung maschineller Übersetzung. BA-Arbeit, Fachhochschule Potsdam, Januar 2011.
- Koehn, P. (2009): A Web-Based Interactive Computer Aided Translation Tool. In Proceedings of the ACL-IJCNLP 2009 Software Demonstrations, Suntec, Singapore.
- Koehn, P., Hoang, H., Birch, A. (2007): ‘Moses: Open Source Toolkit for Statistical Machine Translation’. Paper presented at the Annual Meeting of the Association for Computational Linguistics (ACL), Prague, Czech Republic.
- Siegel, M. (1997): Die maschinelle Übersetzung aufgabenorientierter japanisch-deutscher Dialoge. Lösungen für Translation Mismatches. Berlin: Logos.

Experimenting with Corpus-Based MT Approaches

Monica Gavrilă

University of Hamburg,

Vogt-Kölln Str. 30, 22527, Hamburg, Germany

E-mail: gavrila@informatik.uni-hamburg.de

Abstract

There is no doubt that in the last years corpus-based machine translation (CBMT) approaches have been in focus. Among them, the statistical MT (SMT) approach has been by far the more dominant, although the Workshop on example-based MT (EBMT) at the end of 2009 showed a revived interest the other important CBMT approach: EBMT. In this paper several MT experiments for English and Romanian are presented. In the experimental settings several parameters have been changed: the MT system, the corpus type and size, the inclusion of additional linguistic information. The results obtained by a Moses-based SMT system are compared with the ones given by *Lin-EBMT*, a linear EBMT system implemented during the research. Although the SMT systems outperforms the EBMT system in all the experiments, different behaviors of the systems have been noticed while changing the parameters in the experimental settings, which can be of interest for further research in the area.

Keywords: Machine Translation, SMT, EBMT, Moses, *Lin-EBMT*

1. Introduction

There is no doubt that in the last years corpus-based machine translation (CBMT) approaches have been in focus. Among them, the statistical machine translation (SMT) approach has been by far the more dominant. However, the Workshop on example-based MT (EBMT) at the end of 2009¹ showed a revived interest in the other important CBMT approach: EBMT.

Between these two MT approaches has always been a 'competition'. The similar and unclear definitions and the mixture of ideas make the difference between them difficult to distinguish. In order to show the advantages of one or another method, comparisons between SMT and EBMT (or hybrid) systems are found in the literature. The results, depending on the data type and on the systems considered, seemed to be positive for both approaches: (Way & Gough, 2005) and (Smith & Clark, 2009). Considering English-Romanian as language-pair, results for both SMT and EBMT systems are reported, although a comparison between the two approaches has not been made. SMT systems are presented in (Cristea, 2009) and (Ignat, 2009); results of an EBMT system are

reported in (Irimia, 2009).

In this paper several MT experiments for English (ENG) and Romanian (RON) are presented. In the experimental settings several parameters have been changed: the MT system (approach), the type and size of the corpus, the inclusion of additional part-of-speech (POS) information. The results obtained by a Moses-based SMT system are compared with the ones given by *Lin-EBMT*, a linear EBMT system implemented during the research. The same training and test data have been used for both MT systems.

The following section will briefly present both MT systems. The data used and the translation results will be described in Section 3. Additionally, a very brief analysis of the results will be made. The paper will end with conclusions and some ideas about further work.

2. System Description

In this section the two CBMT systems are briefly characterized.

The SMT system used follows the description of the baseline architecture given for the Sixth Workshop on SMT² and it is based on Moses (Koehn et al., 2007).

¹ <http://computing.dcu.ie/~mforcada/ebmt3/> - last accessed on January 2011.

² <http://www.statmt.org/wmt11/baseline.html> - last accessed on June 2011.

Moses is an SMT system that allows the user to train automatically translation models for the language pair needed, considering that the user has the necessary parallel aligned corpus. We used in our experiments SRILM (Stolcke, 2002) for building the language model and GIZA++ (Och & Ney, 2003) for obtaining the word alignment. Two changes have been done to the specifications of the Workshop on SMT: the tuning step was left out and the language model (LM) order was 3, instead of 5. Leaving out the tuning step has been motivated by results we obtained in experiments which are not the topic of this paper, while comparing different system settings: not all tests in which tuning was involved showed an improvement. We changed the LM order due to results presented in the SMART project³. *Lin-EBMT* is the EBMT system developed during the research. It is mainly based on surface forms (linear EMT system) and uses no additional linguistic resources. Due to space reasons, the main steps of the *Lin-EBMT* system - matching, alignment and recombination - are not described in detail in this paper. We will just present the main translation steps.

The test corpus is preprocessed in the same way as in specification of the Moses-based SMT system: tokenization and lowercasing. In order to reduce the search space a word index is used, a method that is often encountered in the literature, e.g. (Sumita & Iida, 1991). The information needed in the translation, such as the word-index⁴ or the GIZA++ word-alignments, is extracted prior to the translation process itself.

The main steps in *Lin-EBMT*, done for each of the input sentences of the test data, are enumerated below:

- 1) The tokens⁵ in the input, excluding punctuation, are extracted: {token₁, token₂, ..., token_n}.
- 2) Using the word-index, all sentence ids that contain at least one token from the input are considered: {sentenceId₁, ..., sentenceId_m}. The list of sentence ids contains no duplicates. The word-index is used in order to reduce the search space for the matching step. The matching procedure is run only after the search space size is decreased, by using this index.
- 3) Given the preprocessed input sentence and the list of

sentence ids {sentenceId₁, ..., sentenceId_m}, the matching between the input and the 'reduced' source language (SL) side of the corpus is done. If the input sentence is encountered in the corpus, the translation is found and the translation procedure stops. Else, the most similar sentences are extracted by using a similarity measure developed during the research. This measure is based on the longest common subsequence algorithm found in (Bergroth et al., 2000).

- 4) After obtaining the sentences which maximum cover the input, the corresponding word alignments are extracted, by considering the longest aligned target language (TL) subsequences possible.
- 5) Using the "bag of TL sequences" obtained from the alignment the output is generated by making use of a recombination matrix, a new approach for implementing this step.

More details about the *Lin-EBMT* system can be found in (Gavrila, 2011).

3. Evaluation

We used for our evaluation two corpora. The first is a sub-part of the JRC-Acquis version 2.2 (Steinberger et al., 2006), a freely available parallel corpus in 22 languages, which is formed from the European Union documents of mostly legal nature.; the latter is RoGER, a small technical manual manually created and corrected (Gavrila & Elita, 2006). The same training and test data has been used for both SMT and EBMT experiments.

In the EBMT system, matching is done on the corpus for the translation model in the SMT system and recombination on the one for the language model. Both corpora had to be saved in the format which fits the needs of each of the MT systems.

The tests on the JRC-Acquis data have been run on 897 sentences, which were not used for training. Sentences were automatically removed from different parts of the corpus to ensure a relevant lexical, syntactic and semantic coverage. Three sets of 299 sentences represent the data sets **Test 1**, **Test 2**, and **Test 3**, respectively. **Test 1+2+3** is formed from all 897 sentences. The test data has no sentence length restriction, as the training data (see Moses specification).

From RoGER, 133 sentences (**Test R**) have been randomly extracted as the test data, the rest of 2200

³ www.smart-project.eu – last accessed on June 2011.

⁴ The word-index is in fact a token index, as it contains also punctuation signs, numbers, etc.

⁵ A token is a word, a number or a punctuation sign.

sentences representing the training data. When using RoGER, POS information was considered for some of the experiments: data set **Test RwithPOS**⁶.

The obtained translations have been evaluated using two automatic evaluation metrics: BLEU (Papineni et al., 2002) and TER (Snover et al., 2006). The choice of the metrics is motivated by the available resources (software) and, for comparison reason, by the results reported in the literature. Due to lack of data and further translation possibilities, we considered the comparison with only one reference translation.

We present the evaluation scores obtained in Tables 2 and 3.

ENG-RON		
	SMT	Lin-EBMT
Test 1	0.5007	0.8071
Test 2	0.4898	0.6400
Test 3	0.5208	0.7770
Test 1+2+3	0.5023	0.7326
Test R	0.3784	0.5955
Test RwithPOS	0.4748	0.6402
RON-ENG		
	SMT	Lin-EBMT
Test 1	0.5020	0.7041
Test 2	0.3756	-
Test 3	0.4684	-
Test 1+2+3	0.4457	-
Test R	0.3465	0.5443
Test RwithPOS	0.4000	0.5490

Table 2: Evaluation Results (TER scores)

The lower the TER scores, the better the translation results. For the BLEU score the relationship between the scores and the translation quality is the opposite.

While analyzing the behavior of each of the MT system, when changing the test data-set for one corpus (i.e. JRC-Acquis) several factors have been found with a direct influence on the results, such as the number of out-of-vocabulary words, the number of test sentences directly found in the training data, sentence length or the way of extracting the training data: see **Test 1 – Test 3**. For a specific dataset (Test 2), the obtained BLEU score for the EBMT system is similar⁷ with one presented in

⁶ The POS information has been provided by the text processing web services found on: www.racai.ro/webservices/TextProcessing.aspx - last accessed on January 2011.

⁷ A one-to-one comparison is not possible, as the data is not the same.

(Irimia, 2009), where linguistic resources were used.

Considering the analysis of the behavior of each of the MT system, when changing the corpus (a larger and a smaller corpus, which fits the SMT and EBMT framework, respectively), when comparing **Test 1+2+3** and **Test R**, an improvement is found in both cases for the RoGER corpus, although usually it is stated that a large corpus is needed for SMT. This result might be in this specific case so, due to the data type. This shows the high influence of the data on the empirical approaches.

ENG-RON		
	SMT	Lin-EBMT
Test 1	0.3997	0.1335
Test 2	0.4179	0.3072
Test 3	0.3797	0.1476
Test 1+2+3	0.4015	0.2125
Test R	0.4396	0.2689
Test RwithPOS	0.3879	0.2942
RON-ENG		
	SMT	Lin-EBMT
Test 1	0.2545	0.0855
Test 2	0.5628	-
Test 3	0.4271	-
Test 1+2+3	0.4255	-
Test R	0.4765	0.2783
Test RwithPOS	0.4618	0.3624

Table 3: Evaluation Results (BLEU scores)

The results for the data with additional POS information (**Test RwithPOS**) are not conclusive, as when considering the TER scores worse results are obtained for both MT systems, but when considering the BLEU score improvement is noticed for the EBMT system.

In terms of overall BLEU and TER scores, the EBMT system is outperformed by the SMT one. Still, there are cases where the EBMT system provides a better translation, as in the example below:

Input: The EEA Joint Committee

Reference: Comitetul mixt al SEE,

SMT output: SEE Comitetului mixt,

(* ENG: EEA of the Joint Committee)

Lin-EBMT output: Comitetului mixt SEE

(* ENG: of the EEA Joint Committee)

4. Conclusions and Further Work

In this framework - system configuration and data -, in a direct comparison, the EBMT system was not able to match the performance of the SMT system, but there

were examples when its translation has been more accurate. The evaluation scores presented in this paper show how much training and test data influence the translation results. In this EBMT implementation not all the power of the approach was used, so there is room for improvement. As further work, additional information, e.g. word-order information from the TL sentences is to be extracted and used in the recombination step.

5. References

- Bergroth, L., Hakonen, H., Raita, T. (2000): A survey of longest common subsequence algorithms. In Proc. of the Seventh International Symposium on String Processing and Information Retrieval - SPIRE 2000, pp. 39-48, Spain. ISBN: 0-7695-0746-8.
- Cristea, D. (2009): Romanian language technology and resources go to Europe. Presented at the FP7 Language Technology Informative Days. URL: ftp://ftp.cordis.europa.eu/pub/fp7/ict/docs/language-technologies/cristea_en.pdf - last accessed on April 10th, 2009.
- Gavrila, M., Elita, N. (2006): Roger - un corpus paralel aliniat. In Resurse Lingvistice si Instrumente pentru Prelucrarea Limbii Romane Workshop Proceedings, pages 63-67. Workshop held in November 2006, Publisher: Ed. Univ. Alexandru Ioan Cuza, ISBN: 978-973-703-208-9.
- Gavrila, M. (2011): Constrained recombination in an example-based machine translation system. In Vincent Vondeghinste Mikel L. Forcada, Heidi Depraetere, editor, Proceedings of the EAMT-2011 Conference, pages 193-200, Leuven, Belgium, May 2011. ISBN: 9789081486118.
- Ignat, C. (2009): Improving Statistical Alignment and Translation Using Highly Multilingual Corpora. PhD thesis, INSA - LGeco- LICIA, Strasbourg, France. URL: <http://sites.google.com/site/cameliaignat/home/phd-thesis> (last accessed on August 3rd, 2009).
- Irimia, E. (2009): EBMT experiments for the English-Romanian language pair. In Proceedings of the Recent Advances in Intelligent Information Systems, pages 91-102. ISBN 978-83-60434-59-8.
- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E. (2007): Moses: Open source toolkit for statistical machine translation. In Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic.
- Och, F. J., Ney, H. (2003): A systematic comparison of various statistical alignment models. Computational Linguistics, 29(1), pp. 19-51.
- Papineni, K., Roukos, S., Ward, T., Zhu, W.-J. (2002): Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Session: Machine translation and evaluation, pp. 311-318, Philadelphia, Pennsylvania. Publisher: Association for Computational Linguistics Morristown, NJ, USA.
- Smith, J., Clark, S. (2009): EBMT for SMT: A new EBMT-SMT hybrid. In Forcada, M. L. and Way, A., editors, Proceedings of the 3rd International Workshop on Example-Based Machine Translation, pp. 3-10, Dublin, Ireland.
- Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J. (2006): A study of translation edit rate with targeted human annotation. In Proceedings of Association for Machine Translation in the Americas.
- Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., Varga, D. (2006): The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006), Genoa, Italy.
- Stolcke, A. (2002): SRILM - An extensible language modeling toolkit. In Proceedings of the International Conference on Spoken Language Processing, pp. 901-904, Denver, Colorado.
- Sumita, E., Iida, H. (1991): Experiments and prospects of example-based machine translation. In Proceedings of the 29th annual meeting on Association for Computational Linguistics, pp. 185-192, Morristown, NJ, USA. Association for Computational Linguistics.
- Way, A., Gough, N. (2005): Comparing example-based and statistical machine translation. Natural Language Engineering, 11, pp. 295-309. Cambridge University Press.

Method of POS-disambiguation Using Information about Words Co-occurrence (For Russian)

Edward Klyshinsky¹, Natalia Kochetkova², Maxim Litvinov², Vadim Maximov¹

¹Keldysh IAM

Moscow, Russia, 125047 Miusskaya sq. 4

²Moscow State Institute of Electronics and Mathematics

Moscow, Russia, 109029 B. Tryokhsvyatitelsky s. 3

E-mail: klyshinsky@itas.miem.edu.ru, natalia_k_11@mail.ru, promithias@yandex.ru, vadimmax2000@mail.ru

Abstract

The article describes the complex method of part-of-speech disambiguation for texts in Russian. The introduced method is based on the information concerning the syntactic co-occurrence of Russian words. The article also discusses the method of building such corpus. This project is partially funded by RFBR grant 10-01-00805.

Keywords: learning corpora, words co-occurrence base, POS-disambiguation

1. Introduction

Part-of-speech disambiguation is an important problem in automatic text processing. At the time there exist many systems which solve this problem. The earliest projects use rule-based methods (see, for example, Tapanainen & Voutilainen, 1994). This approach is based on the following ideas: the system is supplied with some limiting rules which forbid or allow some certain words combinations. However, this method requires a time-consuming procedure of writing the rules. Besides, though these rules provide a good result, they often leave a considerable part of text not covered. In this connection there have appeared various statistical methods of automatic generation of such rules (for example Brill, 1995).

The n-gram method uses the statistical distribution of word combination in the text. Generally, n-gram model could be written down as follows:

$$P(w_i) = \arg \max P(w_i | w_{i-1}) * \dots * P(w_i | w_{i-N}). \quad (1)$$

$P(w_i)$ is the probability of an unknown tag $\langle w_i \rangle$ occurrence, if $\langle w_{i-N} \rangle$ of the neighbours are known.

In order to avoid the problem of rare data and getting a zero probability for the occurrence of tag combination $\langle w_i | w_{i-1}, w_{i-2} \rangle$, the smoothed probability can be applied for trigram model. The smoothed trigram model contains

linear combinations of trigram, bigram and unigram probabilities:

$$P_{smooth}(w_i | w_{i-2} * w_{i-1}) = \lambda_3 * P(w_i | w_{i-2} * w_{i-1}) + \lambda_2 * P(w_i | w_{i-1}) + \lambda_1 * P(w_i) \quad (2)$$

where the sum of coefficients $\lambda_1 + \lambda_2 + \lambda_3 = 1$, $\lambda_1 > 0$, $\lambda_2 > 0$, $\lambda_3 > 0$. The values for $\lambda_1, \lambda_2, \lambda_3$ are obtained by solving the system of linear equations.

In (Zelenkov, 2005) the authors in their disambiguation model had defined an unknown tag w_i by involving not only the information on left neighbours, but also the right ones. We will use the similar approach when our system works with the trigram model. In this case the unknown tag is defined by involving the left neighbours $\langle w_{i-2}, w_{i-1}, w_i \rangle$ (3), the right ones $\langle w_i, w_{i+1}, w_{i+2} \rangle$ (4), and both the left and the right ones $\langle w_{i-1}, w_i, w_{i+1} \rangle$ (5).

However, both the rule-based and trigram models require large tagged corpora of texts. The trigram rules which do not contain the information on a lexeme reflect specific language features, but the trigrams themselves (with lexemes inside) reflect rather the lexis in use. If the texts from another knowledge domain are given, the trigrams may show considerably worse results than for the initial corpus.

According to Google researches the digital collection of English texts they possess contains 10^{12} words. The British (BNC, 2011) and America (ANC, 2011) National Corpora contain about 10^8 tagged words. According to

the information on January, 2008 Russian National Corpus (RNC, 2011) contains about $5.8 \cdot 10^6$ disambiguated words (and still remain). At present the process of filling up the latest corpora is rather frozen than active (unlike the situation for the first years of the project when it was being filled up intensively). The task of tagging (though automated) 10^{12} words, seems to be economically impracticable, and may be even unnecessary. The realization of practical applications for processing 10^9 trigrams (the quantity estimation for English language could be found in Google (2006)) will require a considerable amount of computational resources.

At present there are trigram bases accumulated that solve the problem with 94-95 % accuracy for Russian (Sokirko, 2004). The additional methods increases the quality of the disambiguation up to 97,3 % (Lyashevskaya, 2010). It is worthy to note that the application of rule-based methods requires essential time expenses. The application of trigrams demands a well-tagged corpus, and it is a costly problem too. The rule creating is also connected with a permanent work of linguists. The results of such work are never in vain, the output remains applicable to many other projects, but such results are helpless to improve the accuracy immediately. In this connection we had set a goal to develop a new method which would use results of the previous developments accumulated in this field and information from partial syntax analysis.

2. Obtaining statistical data on co-occurrence of words

It is widely acknowledged that a resolution of lexical ambiguity should be provided before a syntactic analysis. In this case it is recommended to apply methods like n-grams. However, n-gram method requires a substantial preliminary work to prepare a tagged text corpus. We have decided to develop a disambiguation method, which uses the syntactic information (obtained in the automatic mode) without carrying out full syntactical parsing. In our researches we focused on Russian.

As the practice has shown, full parsing that would provide full constructing of the tree is not required to remove the most part of a homonymy (about 90%). As it happens, it is sufficient to include the rules of words collocation in nominal and verb phrases, folding of

homogeneous parts of sentence, agreement of subject and predicate, prepositions and case government and some others, in total not exceeding 20 rules, which are described by context-free grammar. It is possible to have a more detailed look at the methods of formal description of language, for instance, in Ermakov (2002).

To solve the problems mentioned above, it is necessary to create a method of getting information on a syntactic relationship for the words which are obtained from a non-tagged corpus. Preliminary experiments have shown that in Russian language approximately 50% of words appear to be part-of-speech unambiguous (up to 80% in conversation texts, in comparison with less than 40% for news in English), i.e. there are no lexical homonyms for each of such words. So the probability to find a group of unambiguous words in a text is rather high.

The analysis of Russian sentence structure allows to determine some of its' syntactic characteristic features.

- 1) The noun phrase (NP) which follows the sole verb in the sentence is syntactically dependent on this verb.
- 2) The sole NP which opens the sentence and is followed by a verb, is syntactically subordinated to this verb.
- 3) The adjectives that are located before the first noun in the sentence, or between a verb and a noun, are syntactically subordinated to this noun.
- 4) The paragraphs 1-3 could be applied also to adverbial participles, and it is possible to consider participles instead of adjectives.

We had applied our method to the processing of several untagged corpora in Russian language. The total amount of these corpora included more than 4,2 billion of words. The text sources contain texts on various themes in Russian. The used corpora include the sources given in the Table 1.

The morphological tagging was made with the help of module of morphological analysis "Crosslator" developed by our team (Yolkeen, 2003). The volume of the databases obtained is listed in the table below. The numerator shows the detected total amount of unambiguous words with the given fixed type of syntactic relation. The denominator shows the amount of unique combinations of words of the given type.

The analysis of the results (Table 2) has shown that the selected pairs contain 22200 verbs from 26400 represented in the morphological dictionary, 55200 nouns

from 83000 and 27600 adjectives from 45300 represented in the dictionary. Such a significant amount of verbs could be explained by their low degree of ambiguity as compared with other parts of speech. A small number of adjectives could be explained by the fact that from several adjectives located immediately before a noun, only the first one was entered into the database. It should be noted that when the largest corpus had been integrated into the system, the number of lexemes has not been changed notably, but at the same time the number of pairs detected significantly increased. For example, the number of verbs has increased from 21500 up to 22200, whereas the number of unique combinations of verb + noun type has increased from 8,3 mln to 10,9. Moreover, the amount of such combinations that had occurred more than twice, has increased from 2.3 to 4 mln. Thus, it is possible to say that when a corpus contains more than one billion words, the lexis in use achieves its saturation limit, while its usage continues to change.

Source	Amount mln w/u	Source	Amount mln w/u
WebReading	3049	Lenta.ru	33
Moshkov's Library	680	Rossiyskaya Gazeta	29
RIA News	156	PCWeek RE	28
Fiction coll.	120	RBC	21
Nezavisimaya Gazeta	89	Compulent a.ru	9
		Total	4214

Table 1: Used corpora

Pair	Total, mln	>1, mln	>2, mln
V+N	243 / 10.89	237 / 5.27	235 / 4
Ger+N	40.8 / 2.76	39.3 / 1.25	38.7 / 0.91
N+Adj	67 / 2.15	66 / 1.13	65.6 / 0.9

Table 2: Obtained results

About 9 % of all word occurrences from the total amount of the corpus had been used to build a co-occurrence base. But even this percentage had appeared to be sufficient to construct a representative sample for a word co-occurrence statistics. The estimations have shown that the received word combinations contain not more than 3% of the errors mostly caused by an improper word order or neglect of some syntactically acceptable variants of collocations, deviances in projectivity and mistakes in

the text. It is necessary to stress that all results had been obtained in the shortest terms without any manual tagging of the corpus. Probably the results could be more representative, if we were to use some methods of part-of-speech disambiguation. However, the best methods give a 3-5 % error, and it would affect the accuracy of results but not noticeably. On the other hand, the sharp increase in corpus volume will allow to neglect the false alternatives at a higher level of occurrence and by these means preserve the quality level.

3. Complex Method of Disambiguation

After we had collected the co-occurrence base, which was sufficiently large, we have got all that was necessary to solve the main problem, that is, to create a method of disambiguation for texts in Russian on the basis of information on a syntactic co-occurrence of words.

Let us assume that in the sentence, which is being parsed, there are two words between which there are only several words or no words at all, and it is known that these two words could be linked by a syntactical relation. In this case, if we have other less probable variants of tagging these words, it is possible to assume that the variant with such link will be more probable. The most difficult thing is to collect a representative base of syntactic relations.

In this paper the rules shall be understood as an ordered set: $\langle \mathbf{v}_i, \mathbf{v}_{i+1}, \mathbf{v}_{i+2} \rangle$, where $\mathbf{v}_i = \langle p_w, \{pr\} \rangle$ is a short description of the word, p_w is a part of speech of the word, and $\{pr\}$ is a set of lexical parameters of the word. Thus, in such rules the lexemes of a word are not taken into account in contrast to the lexical characteristics of the word. A rule may be interpreted in different ways and can be written down as an occurrence \mathbf{v}_i with regard for its right neighbours, as an occurrence \mathbf{v}_{i+2} with regard for its left neighbours or as an occurrence \mathbf{v}_{i+1} with regard for its both neighbours. The set of rules has been obtained from the tagged corpus. Following Zelenkov (2005), we will make tagging of a word considering its right and left neighbours. In the mentioned above paper a tag of the word is defined only with regard for the nearest neighbours of current word. However, it is not necessary to produce the result that falls within the global maximum. The exhaustive search of word tagging variants is usually avoided, as it takes too much time.

As it already has been noted above, the ratio of unambiguous tokens is about 50% in Russian. In this

connection there is always a sufficient probability to find a group of two unambiguous words. Moreover, the chance grows as the length of the sentence increases. If such groups are not found while searching a global maximum, the first word in the sentence will indirectly influence even the last word. In the case such groups are present, such relationship is cancelled, and the search of global criterion can be effected over the separate fragments of the sentence. It allows to increase essentially the speed of the algorithm. So the sentence “Так думал молодой повеса, / Летя в пыли на почтовых, / Всевышней волею Зевеса / Наследник всех своих родных.” (Such were a young rake's meditations – / By will of Zeus, the high and just, / The legatee of his relations – / As horses whirled him through the dust.) can be split into three independent parts: “Так думал молодой повеса, Летя”, “Летя в пыли на почтовых” and “Всевышней волею Зевеса Наследник всех своих родных”.

Thus, we no longer consider the problem $P_{\text{sent}} = \text{argmax}(\prod_{i=1}^{n_s} P(\mathbf{v}_i | \mathbf{v}_{i-1}, \mathbf{v}_{i+2}))$, where n_s is a number of words in the sentence, but

$$P_{\text{sent}} = \prod_{i=1}^{n_f} \text{argmax}(\prod_{i=1}^{n_{fi}} P(\mathbf{v}_i | \mathbf{v}_{i-1}, \mathbf{v}_{i+2})), \text{ where } n_f \text{ is a}$$

number of fragments, n_{fi} is a number of words in the i -th fragment. According to formulas (2)-(4), we consider both left and right neighbours of the word.

We seek the optimum from the edges of the fragment towards its center. It is obvious that product of the maximal values of probabilities for each word can give a global maximum. If this is not the case, but the values obtained from two sides had come to one and the same disambiguation of the word in the middle of the fragment, than we will also consider that we have a good enough solution. If variants of disambiguation of the word in the middle of the fragment are different for two solutions, the optimization is carried out for the accessible variants until they won't achieve one and the same decision. In any case, the optimization is not carried out even for an entire fragment, not mentioning the whole sentence.

The amount of unambiguous fragments can be increased by a preliminary disambiguation using another method. We use the described above base of syntactic dependences. So, let we have a set $\{\langle \mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, p \rangle\}$,

$\mathbf{w}_i = \langle l_w, p_w, \{pr\} \rangle$ is a complete description of the word where l_w is a word's lexeme, w_1 is a key word in the word-group (for example, a verb in the pair «verb+noun»), w_2 is a preposition (if any), w_3 is a dependent word, p is a probability of word combination $\mathbf{w}_1 + \mathbf{w}_2 + \mathbf{w}_3$. In this case all rules are searched for every word of the sentence. It should be noticed that no word can participate in more than two rules. Thus, for each word it is necessary to calculate $\text{argmax}(p_1 + p_2)$, where p_1 and p_2 are the probabilities of rules containing this word in dominant and dependent position.

Actually, during the check of compatibility of the words among themselves, our system uses the following bigram model $P(w_i) = \text{argmax} P(w_i | w_{i-l})$, where l means the distance (in number of words), at which the unknown word may stand from the known one. The rule containing the given word is selected in the following way. We take the floating window containing 10 words to the right and left. The dependent word must be located within this window, the preposition must be located before the dependent word, but there must be no main word between them, the adjective must lexically agree with a noun.

4. Results of experiments and discussion

As a result of our work we had obtain the Corpus of syntactical combinations of Russian words. The relations were achieved using untagged corpora of general lexis texts containing more than 4 bln words. The tagging was carried out “on the fly”. There had been revealed about 6 mln of authentic unique word combinations which had occurred in the text more than 340 mln times. According to our estimations, the amount of errors in the obtained corpora doesn't exceed 3 %. The number of word combination can be enlarged by processing the texts of a given new domain. Though, the investigations had shown that scientific texts use other constructions which reduce the amount of sampled combinations, for example, for speech and cognition verbs. Our method extracts about 9 % of tokens from common lexis texts. But news lines give us just about 5 %. Moreover, for scientific texts this number shortened to 3 %. So the method shows different productivity for different domains. Further experiments have discovered that the received results can be used for defining the style of texts.

So the suggested method allows almost automatically

obtaining the information on word compatibility which further can be used, for instance, for parsing or at other stages of text processing. The method is also not strictly tied to the texts of a certain domain and has rather low cost of enlargement.

The estimation of the efficiency of the system with various parameters was carried out with carefully tokenized corpora that contained about 2300 words. Results were checked using Precision and Accuracy measures. The mere involving of information on word compatibility in Russian method had shown 71.98% Precision ratio and 96.75% Accuracy. This result is comparable with best results in selected area (Lee, 2010). The advantage of this method is in its' ability to be additionally adjusted to a new knowledge domain quickly and automatically (that is most important), in case a sufficiently large text corpus is available. The method gives an acceptable quality of disambiguation, unfortunately with not too large Precision.

The coverage ratio can be improved by application of trigram rules, which can be easily received, for example, from <http://aot.ru>, or by analysis of the tagged corpus in Russian (for example, <http://ruscorpora.ru>). The coverage ratio in this case has made 78%, but the accuracy has fallen to 95.6%. In Sokirko (2004) it is mentioned that the systems Inxight and Trigram provide 94.5% and 94.6% accuracy accordingly, that is comparable with the results of our system. Further improvement of coverage ratio up to 81.3 % is possible in case of the improvement of optimal decision search algorithm which is described above, but it slightly brings down the accuracy. In the current state the method is not able to show an absolute coverage, because the part-of-speech list applied in this method was not full, it contained only the following: a verb, a verbal adverb, a participle, a noun, an adjective, a preposition and an adverb. Then, there was no information on some types of relations, for example, «noun+noun». Furthermore, the information on a compatibility of some words of Russian conceptually cannot be obtained because of fundamental homonymy of certain words. For example, the word "white" can be used both as an adjective and as a noun.

Our results are applicable to some (but not all) European languages. So the extremely unambiguous English doesn't allow construct the words combinations database. Method can be applied for German or French but the

rules should be completely rewritten. Problems like verbal detachable prefixes in German and reverse words order should be taken into account.

5. References

- Tapanainen P., Voutilainen A. (1994): Tagging accurately - don't guess if you know. In Proc. of conf. on applied natural language processing, 1994.
- Brill E. (1995): Unsupervised learning of disambiguation rules for part of speech tagging. In Proceedings of the Third Workshop on Very Large Corpora, p. 1-13, 1995.
- Zelenkov Yu.G, Segalovich Yu.A., Titov V.A. (2005): Вероятностная модель снятия морфологической омонимии на основе нормализующих подстановок и позиций соседних слов. Материалы Международной конференции «Диалог'2005»
- British National Corpus (2011): <http://www.natcorp.ox.ac.uk/>
- American National Corpus (2011): <http://americannationalcorpus.org/>
- Russian National Corpus (2011): <http://www.ruscorpora.ru/>
- Google (2006): All Our N-gram are Belong to You, Google research blog, <http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html>
- Sokirko A.V., Toldova S.Yu. (2004): Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка. Материалы конференции «Корпусная лингвистика'2004»
- Lyashevskaya O. at all (2010): Оценка методов автоматического анализа текста: морфологические парсеры русского языка. Материалы Международной конференции «Диалог'2010»
- Ermaikov A.E. (2002): Неполный синтаксический анализ текста в информационно-поисковых системах. Материалы Международной конференции «Диалог'2002»
- Yolkeen S.V., Klyshinsky E.S., Steklyannikov S.E., Проблемы создания универсального морфосемантического словаря. Сб. трудов Международных конференций IEEE AIS'03 и CAD-2003, том 1, 2003. стр. 159-163.
- Lee Y.K., Haghighi A., Barzilay R. (2010) Simple Type-Level Unsupervised POS Tagging. In Proc. of EMNLP 2010

Von TMF in Richtung UML: in drei Schritten zu einem Modell des übersetzungsorientierten Fachwörterbuchs¹

Georg Löckinger

Universität Wien und Österreichische Akademie der Wissenschaften

Wien, Österreich

georg.loeckinger@univie.ac.at

Abstract

Fachübersetzer(innen) brauchen für ihre Tätigkeit maßgeschneiderte fachsprachliche Informationen. Zwischen ihrem Bedarf und den verfügbaren fachsprachlichen Ressourcen besteht jedoch eine große Diskrepanz. In meinem Dissertationsprojekt gehe ich der zentralen Forschungsfrage nach, ob sich das Fachübersetzen mit einem idealen übersetzungsorientierten Fachwörterbuch effizienter gestalten lässt. Zur Beantwortung der zentralen Forschungsfrage werden zuerst mehrere Thesen aufgestellt. Davon wird ein Modell des übersetzungsorientierten Fachwörterbuchs in zwei Detaillierungsgraden hergeleitet, das später mit „ProTerm“, einem Werkzeug für Terminologiearbeit und Textanalyse, in der Praxis experimentell erprobt werden soll. Der vorliegende Aufsatz soll einen Überblick über die bisherige Forschungsarbeit geben. Zuerst werden in knapper Form 15 Thesen vorgestellt, die auf der einschlägigen wissenschaftlichen Literatur und meiner eigenen Berufserfahrung in Fachübersetzen und Terminologiearbeit beruhen. Im Hauptteil des Aufsatzes kommt ein Modell des übersetzungsorientierten Fachwörterbuchs zur Sprache. Das Modell dient als Bindeglied zwischen den konkreten Anforderungen, die mit den 15 Thesen ausgedrückt werden, und der praktischen Umsetzung mit „ProTerm“. Der Aufsatz schließt mit einem Ausblick auf die nächsten Schritte in meinem Dissertationsprojekt ab.

Keywords: übersetzungsorientiertes Fachwörterbuch, übersetzungsorientierte Terminografie, Fachlexikografie, Fachübersetzen

1. Einleitung

Fachübersetzer(innen) hegen seit Langem den Traum von einem übersetzungsorientierten Nachschlagewerk(zeug), das ihrem Bedarf in maximalem Umfang Rechnung trägt. Den historischen Ausgangspunkt für die Beschäftigung mit der einschlägigen wissenschaftlichen Literatur bildet Tiktin (1910) mit dem klingenden Titel „Wörterbücher der Zukunft“. Auch einige andere Literaturstellen bringen die noch nicht erfüllten Träume zum Ausdruck; vgl. Hartmann (1988), Snell-Hornby (1996), de Schryver (2003). Die Diskrepanz zwischen den vorhandenen fachsprachlichen Ressourcen und dem, was Fachübersetzer(innen) benötigen, hat unter diesen zu einer gewissen Unzufriedenheit geführt. Infolgedessen begannen sie, ihre eigenen terminologischen Datenbestände und Nachschlagewerk(zeug)e zu erstellen. Somit kam zu ihrer Tätigkeit der Terminologienutzung jene der Terminologieerarbeitung hinzu.

2. Anforderungen an das übersetzungsorientierte Fachwörterbuch: 15 Thesen

Entgegen einer weitverbreiteten Meinung ist das Fachübersetzen ein komplexer Vorgang; vgl. etwa Wilss (1997). Daher hat das übersetzungsorientierte Fachwörterbuch (ü. F.) mannigfaltige Anforderungen zu erfüllen. Im Folgenden stelle ich diese in Form von 15 Thesen dar, die sich auf die wissenschaftliche Literatur und/oder eigene Argumente stützen. Die 15 Thesen leiten sich aus der empirischen Praxis des Fachübersetzens und der wissenschaftlichen Beschäftigung mit dieser Praxis ab². Sie werden einer der Kategorien „methodikbezogen“, „inhaltsbezogen“ bzw. „Darstellung und Verknüpfung der Inhalte“ zugeordnet, die sich aber – wie auch die einzelnen Thesen selbst – ergänzen und zum Teil überschneiden.

¹ Beim vorliegenden Aufsatz handelt es sich um eine erweiterte und überarbeitete deutsche Fassung von Löckinger (2011).

² Eine ausführliche Darstellung der Argumente für die einzelnen Thesen mitsamt den jeweiligen Literaturverweisen würde den Rahmen dieses Aufsatzes sprengen. Ein Literaturverzeichnis ist beim Autor erhältlich.

2.1. Methodikbezogene Anforderungen

These 1 (systematische Terminologiarbeit): Das ü. F. muss nach den Grundsätzen und Methoden der systematischen Terminologiarbeit erstellt worden sein.

These 2 (Beschreibung der angewandten Methodik): Das ü. F. muss über die (lexikografische und/oder terminografische) Methodik Aufschluss geben, die bei seiner Erstellung zum Einsatz kam.

2.2. Inhaltsbezogene Anforderungen

These 3 (Benennungen und Fachwendungen sowie ihre Äquivalente): Das ü. F. muss Benennungen, Fachwendungen und Äquivalente in Ausgangssprache und Zielsprache(n) enthalten.

These 4 (grammatikalische Informationen): Das ü. F. muss grammatikalische Informationen zu Benennungen, Fachwendungen und Äquivalenten bieten.

These 5 (Definitionen): Das ü. F. muss Definitionen der in ihm beschriebenen Begriffe enthalten.

These 6 (Kontexte): Das ü. F. muss authentische Kontexte (v. a. in der Zielsprache) bereitstellen.

These 7 (enzyklopädische Informationen): Das ü. F. muss enzyklopädische Informationen (fachgebietsbezogene Hintergrundinformationen, z. B. Angaben zur Verwendung eines bestimmten Gegenstandes) enthalten.

These 8 (multimediale Inhalte): Das ü. F. muss nach Möglichkeit und Bedarf Gebrauch von multimedialen Inhalten (Grafiken, Diagrammen, Tondateien usw.) machen.

These 9 (Anmerkungen): Das ü. F. muss mit Anmerkungen zu der in ihm enthaltenen Terminologie versehen sein, z. B. mit Hinweisen zu Übersetzungsfehlern.

2.3. Anforderungen an Darstellung und Verknüpfung der Inhalte

These 10 (elektronische Form): Um den meisten anderen Anforderungen zu entsprechen, muss das ü. F. in elektronischer Form vorliegen.

These 11 (begriffssystematische und alphabetische Ordnung): Das ü. F. muss begriffssystematisch und alphabetisch geordnet sein, um für unterschiedlichste Übersetzungsprobleme brauchbare Lösungen anzubieten.

These 12 (Darstellung von Begriffsbeziehungen): Das ü. F. muss aufzeigen, wie die einzelnen Begriffe der jeweiligen Terminologie zusammenhängen (Begriffsbe-

ziehungen in Abhängigkeit von der Strukturierung des jeweiligen Fachgebiets, z. B. Abstraktionsbeziehungen oder sequenzielle Begriffsbeziehungen).

These 13 (Nutzung von Textkorpora): Da authentische Textkorpora wertvolle fachsprachliche Informationen beinhalten, muss das ü. F. auf geeigneten Textkorpora basieren und gleichzeitig einen Zugriff auf diese bieten.

These 14 (Ergänzungen und Anpassungen durch die/den Fachübersetzer(in)): Das ü. F. muss der/dem Fachübersetzer(in) bedarfsgerechte Ergänzungen und Anpassungen ermöglichen.

These 15 (einheitliche Benutzeroberfläche): Es muss der/dem Fachübersetzer(in) möglich sein, auf die Informationen im ü. F. von einer einzigen Benutzeroberfläche aus zuzugreifen.

3. Modell des übersetzungsorientierten Fachwörterbuchs

Die 15 Thesen sollen nun in ein geeignetes Modell übergeführt werden. Da die Thesen Anforderungen an das ü. F. darstellen, die ausnahmslos auch der empirischen Praxis des Fachübersetzens entstammen, wird im Folgenden induktiv ein Modell des ü. F. entworfen.

Mit Ausnahme der Thesen 10, 14 und 15, die die Umsetzung des Modells betreffen, lassen sich sämtliche Thesen in einem Modell zusammenführen, das das ü. F. mit allen erforderlichen Inhalten beschreibt. Ausgehend von dem TMF-Modell in der internationalen Norm ISO 16642 (2003) wird das Modell des ü. F. in zwei Detaillierungsgraden vorgestellt (vgl. Budin (2002)). Insgesamt entspricht dies der Drei-Ebenen-Einteilung nach Budin und Melby (2000), die beim Projekt „SALT“ zum Einsatz kam.

Die Modellierung dient hier in zweifacher Hinsicht „als Bindeglied zwischen Empirie und Theorie“ (Budin, 1996:196): Einerseits werden die 15 Thesen induktiv zum Modell in den zwei genannten Detaillierungsgraden umformuliert, andererseits soll das Modell wiederum deduktiv in die empirische Praxis übergeführt und dort experimentell erprobt werden. Diese schrittweise Vorgangsweise hat den Vorteil, dass man sich bei der Modellierung ganz dem zu schaffenden abstrakten und implementierungsunabhängigen Modell widmen kann, ohne sich um die Einzelheiten seiner späteren technischen Umsetzung kümmern zu müssen (vgl. etwa Sager (1990)).

Nachstehend geht es um das TMF-Modell (3.1.), das Modell im ersten Detaillierungsgrad einschließlich des Modells des terminologischen Eintrags (3.2.) und das Modell im zweiten Detaillierungsgrad (Datenmodell, 3.3.). Im Mittelpunkt steht das Modell im ersten Detaillierungsgrad, da dieses bereits in ausgereifter Form vorliegt.

3.1. TMF-Modell gemäß ISO 16642 (2003)

Die internationale Norm ISO 16642 (2003) beschreibt ein Rahmenmodell für die Auszeichnung terminologischer Daten (TMF), mit dem Auszeichnungssprachen für terminologische Daten definiert werden können, die sich wiederum mit einem generischen Abbildungswerkzeug aufeinander abbilden lassen. Die ISO 16642 (2003) hat zum Ziel, die Nutzung und Weiterentwicklung von Computeranwendungen für terminologische Daten zu fördern und den Austausch terminologischer Daten zu erleichtern. Im Gegensatz dazu ist die Festlegung von Datenkategorien nicht Gegenstand dieser Norm; vgl. dazu ISO 12620 (1999).

Schematisch ergibt das TMF-Modell folgendes Bild:

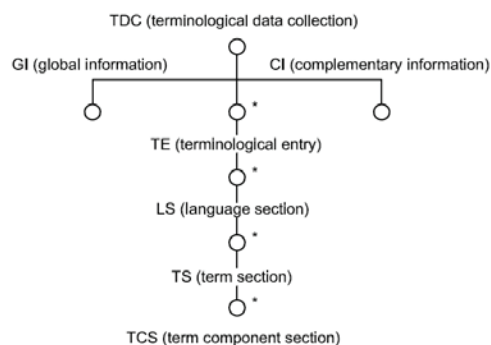


Bild 1: Schematische Darstellung des TMF-Modells aus ISO 16642 (2003).

Die oben abgebildeten Bestandteile des Modells lassen sich wie folgt beschreiben (von oben nach unten, von links nach rechts; vgl. DIN 2330 (1993), DIN 2342 (2004), ISO 16642 (2003)):

TDC (terminological data collection): oberste Stufe, die alle zu einem terminologischen Datenbestand gehörenden Informationen umfasst;

GI (global information): globale Informationen = administrative und technische Angaben, die sich auf den gesamten terminologischen Datenbestand beziehen;

CI (complementary information): zusätzliche Informa-

tionen = Angaben, die über jene in den terminologischen Einträgen hinausgehen und üblicherweise von mehreren terminologischen Einträgen aus angesprochen werden;

TE (terminological entry): terminologischer Eintrag (Eintragungsebene), d. h., jener Teil eines terminologischen Datenbestands, der terminologische Daten zu einem einzigen Begriff oder zu mehreren quasiäquivalenten Begriffen enthält;

LS (language section): Sprachebene, d. h., jener Teil eines terminologischen Eintrags, in dem sich terminologische Daten in einer Sprache befinden;

TS (term section): Benennungsebene, d. h., jener Teil der Sprachebene, der terminologische Daten zu einer oder mehreren Benennungen bzw. Fachwendungen umfasst;

TCS (term component section): unterste Stufe, die (nicht) bedeutungstragende Einheiten von Benennungen bzw. Fachwendungen beschreibt.

3.2. Das Modell des übersetzungsorientierten Fachwörterbuchs

Als Grundlage dient das Modell eines terminologischen Eintrags von Mayer (1998). Dieses wird nach den Erfordernissen meines Dissertationsprojektes so angepasst und erweitert, dass daraus ein Modell des ü. F. in zwei Detaillierungsgraden resultiert (Modell im ersten Detaillierungsgrad, Modell im zweiten Detaillierungsgrad = Datenmodell).

3.2.1. Das Modell des terminologischen Eintrags

Gemäß dem derzeitigen Stand der Forschung zur terminografischen Modellierung muss das Modell des terminologischen Eintrags folgenden fünf Kriterien entsprechen: Begriffsorientierung (vgl. etwa ISO 16642 (2003)), Benennungsautonomie (vgl. etwa Schmitz (2001)), Elementarität (vgl. etwa ISO/PRF 26162 (2010)), Granularität (vgl. etwa Schmitz (2001)) und Wiederholbarkeit (vgl. etwa ISO/PRF 26162 (2010)). Von Belang sind hier ferner die drei oben genannten Ebenen des TMF-Modells (Eintragungsebene, Sprachebene und Benennungsebene).

Die nachstehenden Datenkategorien leiten sich entweder aus den 15 Thesen oder aus dem derzeitigen Stand der Forschung zur terminografischen Modellierung ab (vgl. insbesondere ISO 12620 (1999) und das ISO-Datenkategorienverzeichnis „ISocat“ unter www.isocat.org). Mit einem hochgestellten Pluszeichen („+“) versehene Bezeichnungen beziehen sich auf Da-

tenkategorien, die auf einer oder mehreren der drei oben genannten Ebenen Datenelemente enthalten können. Ein hochgestelltes „^W“ zeigt an, dass die jeweilige Datenkategorie innerhalb der Ebene, auf der sie genannt wird, wiederholbar sein muss.

Die Eintragungsebene umfasst folgende Datenkategorien: enzyklopädische Informationen⁺, multimediale Inhalte^W, Anmerkung^{+W}, Position des Begriffs (wenn nur ein Begriff), Quellenangabe^{+W}, administrative Angaben^{+W}. Auf der Sprachebene befinden sich folgende Datenkategorien: Definition (wenn nur ein Begriff) bzw. Definition^W (wenn mehrere quasiäquivalente Begriffe), enzyklopädische Informationen⁺, Anmerkung^{+W}, Position der Begriffe^W (wenn mehrere quasiäquivalente Begriffe), Quellenangabe^{+W}, administrative Angaben^{+W}. Die Benennungsebene schließlich besteht aus den Datenkategorien Benennung/Fachwendung/Äquivalent^W, grammatikalische Informationen^W, Kontext^W, enzyklopädische Informationen⁺, Anmerkung^{+W}, Quellenangabe^{+W}, administrative Angaben^{+W}.

3.2.2. Das Modell im ersten Detaillierungsgrad

Das Modell im ersten Detaillierungsgrad, dessen Herzstück das oben erläuterte Modell des terminologischen Eintrags bildet, sieht grob so aus:

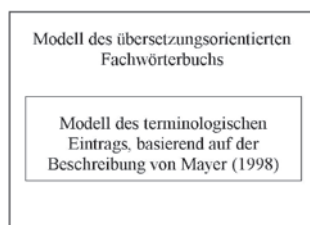


Bild 2: Überblicksartige schematische Darstellung des Modells im ersten Detaillierungsgrad.

Zu den drei bereits erwähnten Ebenen (Eintragungsebene, Sprachebene, Benennungsebene) kommen noch die zwei Bestandteile „globale Informationen“ und „zusätzliche Informationen“ aus dem TMF-Modell in ISO 16642 (2003) hinzu. Die erforderlichen Datenkategorien leiten sich erneut entweder aus den 15 Thesen ab oder ergeben sich aus dem derzeitigen Stand der Forschung zur terminografischen Modellierung; vgl. insbesondere ISO 12620 (1999), ISO 16642 (2003), ISO/PRF 26162 (2010), aber auch ISO 1951 (2007). Folglich handelt es sich bei den globalen Informationen um administrative und technische Angaben, während zu den zusätzlichen

Informationen Begriffspläne, Meta-Informationen zum ü. F., multimediale Inhalte, alphabetische Auszüge aus der terminologischen Datenbasis, bibliografische Angaben, Textkorpora, Quellenangaben und administrative Angaben zählen.

Das Modell im ersten Detaillierungsgrad lässt sich im Einzelnen wie folgt darstellen:

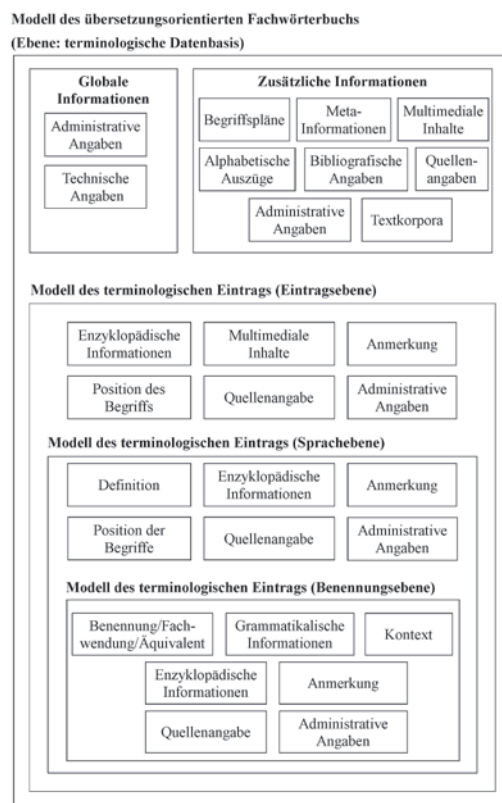


Bild 3: Genauere schematische Darstellung des Modells im ersten Detaillierungsgrad.

3.2.3. Das Modell im zweiten Detaillierungsgrad (Datenmodell)

Aus dem oben erörterten und abgebildeten Modell im ersten Detaillierungsgrad soll ein Datenmodell entwickelt werden, das später in einer empirischen Untersuchung mit „ProTerm“ praktisch umgesetzt und experimentell erprobt wird. Hierbei kommt die objektorientierte Modellierungssprache „Unified Modeling Language“ (UML) zum Einsatz. Diese wird in den einschlägigen internationalen Normen verwendet (vgl. ISO 16642 (2003) und ISO/PRF 26162 (2010)) und bietet sich vor allem dann an, wenn ein Datenmodell in Form einer relationalen Datenbank umgesetzt werden soll. UML-Modelle sind jedoch implementierungsunabhängig und können technisch auch anders umgesetzt werden.

Das UML-Modell befindet sich im Entwurfsstadium und kann daher an dieser Stelle nicht veröffentlicht werden. Der aktuelle Entwurf kann auf Anfrage zur Verfügung gestellt werden.

4. Ausblick

Der nächste Schritt nach einer etwaigen Verfeinerung des Modells im ersten Detaillierungsgrad wird darin bestehen, ein Datenmodell in Form eines UML-Diagramms zu entwerfen, das sich für die Umsetzung mit „ProTerm“ eignet. Eine empirische Untersuchung wird zeigen, ob das Modell dem Bedarf von Fachübersetzerinnen und Fachübersetzern in maximalem Umfang Rechnung tragen und eine Antwort auf die zentrale Forschungsfrage geben kann. Das Modell des ü. F. ist unabhängig von einem bestimmten Fachgebiet oder einer bestimmten Sprachenkombination. Für die empirische Untersuchung wird das Fachgebiet Terrorismus, Terrorismusabwehr und Terrorismusbekämpfung in den Sprachen Deutsch und Englisch herangezogen. Mit der Terminologie dieses Fachgebiets habe ich mich sowohl wissenschaftlich als auch in der Berufspraxis eingehend beschäftigt.

5. Literatur

- Budin, G. (1996): Wissensorganisation und Terminologie: Die Komplexität und Dynamik wissenschaftlicher Informations- und Kommunikationsprozesse. Tübingen: Narr.
- Budin, G. (2002): Der Zugang zu mehrsprachigen terminologischen Ressourcen – Probleme und Lösungsmöglichkeiten. In K.-D. Schmitz, F. Mayer & J. Zeumer (Hg.), *eTerminology. Professionelle Terminologearbeit im Zeitalter des Internet – Akten des Symposiums*, Köln, 12.-13. April 2002. Köln: Deutscher Terminologie-Tag e.V., S. 185–200.
- Budin, G., Melby, A. (2000): Accessibility of Multilingual Terminological Resources – Current Problems and Prospects for the Future. In A. Zampolli et al. (Hg.), *Proceedings of the Second International Conference on Language Resources and Evaluation*, volume II. Athens, S. 837–844.
- DIN 2342 (2004). *Begriffe der Terminologielehre (Entwurf)*.
- DIN 2330 (1993). *Begriffe und Benennungen – Allgemeine Grundsätze*.
- Hartmann, R. R. K. (1988): *The Learner's Dictionary: Traum oder Wirklichkeit?* In K. Hyldgaard-Jensen & A. Zettersten (Hg.), *Symposium on Lexicography III. Proceedings of the Third International Symposium on Lexicography*, May 14–16, 1986 at the University of Copenhagen. Tübingen: Niemeyer, S. 215–235.
- ISO 12620 (1999). *Computer applications in terminology – Data categories*.
- ISO 16642 (2003). *Computer applications in terminology – Terminological markup framework*.
- ISO 1951 (2007). *Presentation/representation of entries in dictionaries – Requirements, recommendations and information*.
- ISO/PRF 26162 (2010). *Systems to manage terminology, knowledge and content – Design, implementation and maintenance of Terminology Management Systems*.
- Löckinger, G. (2011): *User-Oriented Data Modelling in Terminography: State-of-the-Art Research on the Needs of Special Language Translators*. In T. Gornostay & A. Vasiljevs (Hg.), *NEALT Proceedings Series Vol. 12. Proceedings of the NODALIDA 2011 workshop, CHAT 2011: Creation, Harmonization and Application of Terminology Resources*, May 11, 2011, Riga, Latvia. Northern European Association for Language Technology, S. 44–47.
- Mayer, F. (1998): *Eintragsmodelle für terminologische Datenbanken. Ein Beitrag zur übersetzungsorientierten Terminographie*. Tübingen: Narr.
- Sager, J. C. (1990): *A Practical Course in Terminology Processing*. Amsterdam: Benjamins.
- Schmitz, K.-D. (2001): *Systeme zur Terminologieverwaltung. Funktionsprinzipien, Systemtypen und Auswahlkriterien (online edition)*. *technische kommunikation*, 23(2), S. 34–39.
- de Schryver, G.-M. (2003): *Lexicographers' Dreams in the Electronic-Dictionary Age*. *International Journal of Lexicography*, 16(2), S. 143–199.
- Snell-Hornby, M. (1996): *The translator's dictionary – An academic dream?* In M. Snell-Hornby (Hg.), *Translation und Text. Ausgewählte Vorträge*. Wien: WUV-Universitätsverlag, S. 90–96.
- Tiktin, H. (1910): *Wörterbücher der Zukunft*. *Germanisch-romanische Monatsschrift*, II, S. 243–253.
- Wilss, W. (1997): *Übersetzen als wissensbasierte Tätigkeit*. In G. Budin & E. Oeser (Hg.), *Beiträge zur Terminologie und Wissenstechnik*. Wien: TermNet, S. 151–168.

Annotating for Precision and Recall in Speech Act Variation: The Case of Directives in the *Spoken Turkish Corpus*

Şükriye Ruhi^a, Thomas Schmidt^b, Kai Wörner^b, Kerem Eryılmaz^c

^a Middle East Technical University, ^b Hamburg University

^a Dept. of Foreign Language, Education, Faculty of Education, 06800 Ankara

^b SFB 538 'Mehrsprachigkeit' Max Brauer-Allee 60, D-22765 Hamburg

^c Dept. of Cognitive Science, Graduate School of Informatics, 06800 Ankara

E-mail: sukruh@metu.edu.tr, thomas.schmidt@uni-hamburg.de, kai.woerner@uni-hamburg.de,
keryilmaz@gmail.com

Abstract

Speech act realizations pose special difficulties in search during annotation and pragmatics research based on corpora in spite of the fact that their various forms may be relatively formulaic. Focusing on spoken corpora, this paper concerns the generation of discourse analytical annotation schemes that can address not only variation in speech act annotation but also variation in dialog and interaction structure coding. The major arguments in the paper are that (1) enriching the metadata features of corpus design can act as useful aids in speech act annotation; and that (2) sociopragmatic annotation and corpus-oriented pragmatics research can be enhanced by incorporating (semi-)automated linguistic annotations that rely both on bottom-up discovery procedures and the more top-down, linguistic categorizations based on the literature in traditional approaches to pragmatics research. The paper illustrates implementations of enriched metadata and pragmatic annotation with examples drawn from directives in the demo version of the *Spoken Turkish Corpus*, and presents a qualitative assessment of the annotation procedures.

Keywords: speech act annotation, variation, spoken Turkish, precision, metadata

1. Speech acts as a challenge for corpus annotation

Speech act realizations are notorious for the special difficulties they pose in search both during annotation and pragmatics research based on corpora, in spite of the fact that their various forms may be relatively formulaic, hence amenable to (semi-)automatic annotation. Sociopragmatic annotation involves significant difficulties in the very process of identifying categories and units of pragmatic phenomena such as variation in manifestations of speech acts and the identification of conversational segments (Archer, Culpeper & Davies, 2008:635). As underscored by Schmidt and Wörner, this makes pragmatics research conducted on corpora “heuristic” in nature in that the relationship between theory and corpus analysis is bi-directional (2009:4). This is all the more so in the identification of speech acts, as function only partially follows form.

To illustrate this with a short excerpt from a naturally occurring speech event, the utterance *iki çay* ‘two teas’ may be describing the number of cups of tea one has had. But followed by *tamam hocam* “okey deferential address term”, the noun phrase would achieve the illocutionary force of a request when uttered to a service provider. It goes without saying that the initial utterance can occur with *please* as a politeness marker, which would certainly increase its chance of being identified as a request. Communications, however, do not always exhibit such pre-fabricated forms. Thus their recall in corpora would require the analyst to increase the number of search expressions infinitely. Even so, that would not guarantee full recall; neither would it filter false cases. This situation goes against the advantage of using corpora for the study of variation and largely limits the derivation of qualitative and quantitative conclusions from corpora. In this paper we argue that annotation for studying variation in speech act realizations can be improved by (1) enriching metadata coding during the construction stage

of a corpus; and (2) by implementing (semi-)automated annotation for sociopragmatic features of communications that rely both on bottom-up discovery procedures and top-down, linguistic categorizations based on traditional approaches to pragmatics research (e.g. annotation of socially and discursively significant verbal and non-verbal phenomena and non-phonological units such as multi-word expressions and changes in tone of voice). The argumentation is based on insights from Multidimensional Analysis (Biber, 1995) and vocabulary-based identification of discourse units (Csomay, Jones & Keck, 2007), and the fact that pragmatic phenomena in conversational management (e.g., illocutionary force indicating devices, address terms, and politeness formulae) tend to form constellations of ‘traces’ in discourse. Annotating such traces can add “precision” and improve “recall” (Jucker et al. 2008) in searching for variation in speech acts. The main thrust of the paper is that speech events and discourse level units exhibit such verbal and non-verbal clusters, and that annotating such units can provide insights for further discursive coding. Below, we explain the procedures for these two approaches to annotation with illustrations from the demo version of the *Spoken Turkish Corpus* (STC), which currently comprises 44,962 words from a selection of recordings in conversational settings, service encounters, and radio archives (STC employs EXMARaLDA corpus construction tools (Schmidt, 2004), along with a web-based corpus management system.).

2. Metadata construction in the transcription and annotation workflow of STC

Besides constructing a metadata system for domain, interactional goal and speaker features, we maintain that the inclusion of speech acts and conversational topics as part of the metadata features of a corpus is a significant tool for tracing variation in speech acts in a systematic manner, as topical variation can impact their performance beyond the influence of domain and setting features. Viewed from another perspective, spoken texts are slippery resources of language in terms of domain and setting categorization such that they are often characterized by shifts in interactional goals. A service encounter in a shop, for example, can easily turn into a

chat. Thus, if a communicative event were classified only for its domain of interaction, one would risk the chance of tracing subtle differences within the same domain along several dimensions. The simultaneous annotation of topics and speech acts during the compilation of the recordings and during their transcription can address the concern for achieving maximal retrieval of tokens of a speech act. It enables a bottom-up approach to search for variation through control for topic and speech acts, as manifestations of the act may not exhibit structures noted in the literature. It also allows for a corpus-driven categorization of speech acts that may not have been investigated at all in the particular language. The stages in this procedure in the construction of STC are outlined below:

- 1) Noting of local and global topics, and the communication related activities by recorders (e.g. studying for an exam)
- 2) Checking of topics and additions during the transfer of the recording to the corpus management system
- 3) Stages in transcription:
 - a. Initial step: basic transcription of recording for verbal and non-verbal; editing of topics and addition of speech act metadata
 - b. First check: Checking the transcription for verbal and non-verbal events; editing of topics and speech act metadata
 - c. Second check: Checking the transcription for verbal and non-verbal events; editing of topics and speech act metadata

To achieve a higher level of reliability in transcription, a different transcriber is responsible for the annotation in each step in (3), and differences in transcription are handled through consultation. Stages (1) and (3a) ideally involve the same person so that the transcriber has an intuitive grasp of the topical content and the affective tone of the communication. This procedure has the added advantage of detecting regional variation with more precision. It also renders possible the construction of sub-corpora for initial pilot annotation not only through control for domain but also for topic and speech act, thus enhancing the likelihood of retrieval of a greater variety of tokens in a more economical manner. Naturally, this workflow taps into native speaker intuitions on speech act performance, but it is a viable methodological

procedure in linguistics because it harnesses intuitions in a context-sensitive environment during text processing. Figure 1 displays a select number of the metadata features of one communication in STC (Note that topics are written in Turkish, and that the term *requests* is used instead of *directives* because the former was a more transparent term for the transcribers in step (3a) above:

063_090622_00020 (5 Speakers, 1 Transcription) Browse online	
Date recorded	2009-06-22T18:00:00
Domain	Conversations among family and/or relatives
Duration	1640
Genre	Conversation between family and/or relatives
Physical space	Home
project-name	ODT-STD
Relations	ZEY000073 is mother of ISA000058. ISA000058 is elder brother of CAG000125. ZEY000073 is mother of CAG000125.
Speech acts	Leaves taking, Thanking, Well wishes/congratulations, Refusals (as a response to a request), Requests, Advising, Criticizing, Offering
Topics	Dertleşme, Üniversite eğitimi sonrası hakkında, futbol oynama, Aile içi iletişim sorunları, akşam için plan yapma, Arapçanın ağızları arasında farklar, Yurt arkadaşları ile iletişim ve paylaşma, Hava durumu, Bebek bakımı, Yemek yapma, Matematik çalışma

Figure 1: Partial metadata for a communication in STC

3. Annotation procedure for speech acts

Speech act annotation in STC is being implemented with Sextant (Wörner, n.d.), which also allows searches to be conducted with EXAKT. The search for tokens of directives employs a snowballing technique in developing regular expressions, and is similar to what Kohlen (2008:21) describes as “structural eclecticism”.

The annotation procedure starts off with the identification of forms that have been identified as being representative of directives in Turkish. Regular expressions based on these forms have been developed, and the development of tag sets is done according to the syntactic and/or lexical features of the head act. But instead of tagging only the head act, the full act is further coded by placing opening and closing tags for the relevant head act (see, Examples 1 and 2). This will allow further detailed tagging of the act in later stages of annotation.

The regular expressions are enriched based on tokens detected first in the sub-corpora of service encounters both by examining the larger context of the tokens recalled in initial searches and by manually investigating specific communications that are marked for directives in the corpus metadata. However, this procedure does not

allow elliptical directives and hints to be recalled automatically. Based on the idea that a directive is ideally part of an adjacency pair, the search for ‘hidden’ manifestations of the act is conducted through the presence of address terms and a select number of minimal responses, including lexical and non-lexical backchannels (e.g. *tamam* ‘okey/enough/full’, *ha?*, *hm*), which turned out to collate frequently with directives. Searches were thus conducted separately for these responses, and tokens that did not collocate with directives or form the head act itself were eliminated from the annotation (as is the case with *tamam*).

Example (1) shows the co-occurrence of *tamam* with an elliptical request (tag code: RNp), which could not be recalled with a regular expression (The head act is marked in bold). It is noteworthy that the sequence manifests the presence of the discourse marker *şimdi* ‘now’, which marks the speech act boundary, and illustrates how both minimal responses (*tamam*) and discourse markers collocate with the head act.

(1)

Speakers	Interaction	Translation
XAM000066	<i>şimdi</i> ((0.3)) <i>T.C. kimlik numarası</i> ((0.2)) <i>ve öncelikli olarak</i> ((0.1)) <i>ev adresinizi</i> DIL000065 <i>tamam.</i> ((0.3)) ((filling in a form,10.8))	RNp-openNow your Turkish ID number and first you home address ((XXX))RNp-close
DIL000065	okey.	

Example (2) is an illustration from a service encounter. The head act has a verb with the future in the past. In isolation the utterance could be a manifestation of a representative. However, the collocation of the utterance with *buyrun* ‘lit. command’ (idiomatic equivalent: Welcome) disambiguates it as a request.

(2)

Speakers	Interaction	Translation
MEH000222	((0.3)) <i>buyrun.</i>	welcome.
MED000112	<i>iyi günler!</i>	good day!
MEH000222	<i>neresi olacak?</i>	where is it to be? (idiomatic equivalent: where to?)
MED000112	<i>Dikili'ye bilet alacaktım.</i>	RImpFul-openI was going to get a ticket for Dikili RImpFul-close

Such collocations allow us to form a list of (semi-)formulaic conversational management units, which should be tagged as pragmatic markers for directives. In the demo version of STC, *tamam* 'okey' is the item that exhibits the highest frequency. A search on the occurrence of the item was therefore conducted to check its collocation with directives. The search yielded 298 tokens, 20 of which were related to directives. In 8 instances, the item is a supportive move for the directive head act. In 2 recalls it was the head act itself to close off a conversational topic, while the remaining tokens were responses to a verbal or non-verbal request or part of the response to questions asking for advice/opinion. Amongst these we find the supportive function of *tamam* as a compliance gainer to be especially significant since the literature on directives in Turkish does not identify this function. Within these recalls, *tamam* collocates with 6 requests of the kind illustrated in Example (2). This suggests that *tamam* can function to disambiguate representatives from requests and can be used to retrieve elliptical directives and hints. Although the full description of the pragmatics of *tamam* needs to be refined, we can say that in its semantically bleached use, it appears in topic closures, it functions as a backchannel to check comprehension, and is used as an agreement marker or as a pre-sequence to disagreement. In this regard, we can say that *tamam* is a pragmatic marker in its non-literal use and needs to be tagged accordingly.

4. Conclusions

This paper touches only upon the disambiguating capacity of lexical pragmatic markers, but the distribution of *tamam* supports the claim that discourse segmentation and conversational structure annotation can use the clues provided by such 'traces'. The functional description of *tamam* naturally raises the question as to coding principles for such items, including politeness formulae. While non-lexical backchannels may not be too problematic, the classification and coding of pragmatic markers is a fuzzy area. At this stage, we propose that a semantic-based, broad categorization be made to distinguish lexical and non-lexical markers, interjections and discourse markers, and discourse particles.

Our experience in testing the effect of pragmatic markers on recall of speech acts suggests that it is possible to

envision generic level schemes for speech act annotation. These would proceed first with a bottom-up approach, in which (multi-word) pragmatic markers, backchannels and non-verbal cues such as a classification of activity types (e.g., handing over money) are tagged. It is likely that such a venture will reveal commonalities between speech acts beyond what may be gleaned from the current pragmatics literature on speech act manifestations.

5. Acknowledgements

This paper was supported by TÜBİTAK, grant no. 108K283, and METU, grant no. BAP-05-03-2011-001.

6. References

- Archer, D., Culpeper, J., Davies, M. (2008): Pragmatic annotation. In A. Lüdeling & M. Kytö, Merja (Eds.), *Corpus Linguistics: An International Handbook*, Vol. I. Berlin/New York: Walter de Gruyter, pp. 613-642.
- Biber, D. (1995): *Dimensions of Register Variation*. New York: Cambridge University Press.
- Csomay, E., Jones, J.K., Keck, C. (2007): Introduction to the identification and analysis of vocabulary-based discourse units. In D. Biber, U. Connor & T.A. Upton (Eds.), *Discourse on the Move. Using Corpus Analysis to Describe Discourse Structure*. Amsterdam/ Philadelphia: John Benjamins, pp. 155-173
- Jucker, A., Schneider, G., Taavitsainen, I., Breustedt, B. (2008): "Fishing" for compliments. Precision and recall in corpus-linguistic compliment research. In A. Jucker & I. Taavitsainen (Eds.), *Speech Acts in the History of English*. Amsterdam/Philadelphia: Benjamins, pp. 273-294.
- Kohnen, T. (2008): Historical corpus pragmatics: Focus on speech acts and texts. In A. Jucker & I. Taavitsainen (Eds.), *Speech Acts in the History of English*. Amsterdam/Philadelphia: Benjamins, pp. 13-36.
- Schmidt, T. (2004): Transcribing and Annotating Spoken Language with EXMARALDA. In *Proceedings of the LREC-Workshop on XML Based Richly Annotated Corpora*, Lisbon 2004. Paris: ELRA, pp. 69-74.
- Schmidt, T., Wörner, K. (2009): EXMARALDA – creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics*, 19(4), pp. 565-582.
- Spoken Turkish Corpus. <http://std.metu.edu.tr/en/>
- Wörner, K. n.d. Sextant tagger. <http://www.exmaralda.org/sextant/sextanttagger.pdf>

The SoSaBiEC Corpus: Social Structure and Bilinguality in Everyday Conversation

Veronika Ries¹, Andy Lücking²

¹Universität Bielefeld, BMBF Projekt Linguistic Networks

²Goethe-Universität Frankfurt am Main

E-mail: Veronika.Ries@uni-bielefeld.de, Luecking@em.uni-frankfurt.de

Abstract

The SoSaBiEC corpus is comprised audio recordings of everyday interactions between familiar subjects. Thus, the material the corpus is based on is not gained in task-oriented dialogue under strict experimental control; rather, it is made up of spontaneous conversations. We describe the raw data and the annotations that constitute the corpus. Speech is transcribed at the level of words. Dialogue act oriented codings constitute a functional, qualitative annotation level. The corpus so far provides an empirical basis for studying social aspects of unrestricted language use in a familiar context.

Keywords: bilinguality, social relationships, spontaneous dialogue, annotation

1. Introduction

From the point of view of the methodology of psycholinguistic research on speech production unconstrained responding behavior of participants is problematic: it is known as “the problem of exuberant responding” and it is to be avoided by means of some sort of controlled elicitation in an experimental setting (Bock 1996:407; see also Pickering & Garrod 2004:169). In addition, elicitations are usually bound up with a certain task the participants of the experimental study have to accomplish. Of course, each experimental set-up that obeys to the general “avoid-exuberant-responding”-design and is therefore appropriate to study and test the conditions underlying speech production in a controlled way. However, when studying human-to-human face-to-face dialogue (or multi-logue, in case of more than two interlocutors), elicited communication behavior hinders the unfolding of spontaneous utterances and task-independent dialogue management. Task-oriented dialogue is known to be plan-based (Litman & Allen, 1987). The domain knowledge the interlocutors have of the task-domain together with the difference between their current state and the target state (defined in terms of the task to be accomplished) provides a structuring of dialogue states: the way from the current dialogue state to the target state is operationalized as a

sequence of sub-tasks, each of these sub-tasks is part of a plan that has to be worked off sequentially in order to reach the target state. Plan-based accounts to dialogue provide a functional account to dialogue and have been successfully applied in computational dialogue systems for, e.g., timetable enquiries (Young & Proctor, 1989). At least partly due to the neat status of task-oriented conversational settings, respective study designs have been paradigmatic in linguistic research on dialogue. Task-oriented dialogues, inter alia, pre-determine the following conversational ingredients:

- they define a dialogue goal and thereby a terminal dialogue state;
- they constrain the topics the interlocutors talk about to a high degree (up to move type predictability, modulo repairs etc.);
- they are cooperative rather than competitive;
- the dialogue goal determines the social relationship of the interlocutors (for instance, whether they have equal communicative rights or whether task-knowledge is asymmetrically distributed) and it does so regardless of the actual relationships that might obtain between the interlocutors;
- they are unilingual.

Each of the ingredients above is lacking in spontaneous, everyday conversation. Does this mean that spontaneous,

everyday conversations also lack any structure of dialogue management? Answers to this question are in general given on the grounds of armchair theorizing or case studies. The feasibility of empirical approaches is simply hindered by the lack of respective data. The afore-given list can be extended by a further feature, namely the fact that it is easier to gather task-oriented dialogue data in experimental settings than to collect rampant spontaneous dialogue data. We have some spontaneous dialogue data that lack each of the task-based features listed above – see section 2 for a description. We focus on the latter two aspects here, namely social structure and bilingualism. The social dimension of language use, for instance, social deixis, is a well-known fact in pragmatics (Anderson & Keenan, 1985; Levinson, 2008). The influence of social structure on the structure of lexica has also been reported (Mehler, 2008). Yet, there is no account that scales the macroscopic level of language communities down to the microscopic level of dialogue. The data collected in SoSaBiEC aims at exactly this level of granularity of social structure and language structure: how does the social relationship between interlocutors affect the structure of their dialogue lexicon?

A special characteristic of SoSaBiEC is bilingualism. The subjects recorded speak Russian as well as German, and they make use of both languages in one and same dialogue. What dialogical functions performed by the two languages seems to depend at least partly on who the addressees are, that is, on the social relationship between the interlocutors (Ries, to appear). This qualitative observation will be operationalized in terms of quantitative analyses that focus on the relationship-dependent, functional use of languages (cf. the outlook given in section 4).

According to the bi-partition of corpora – primary or raw data are coupled with secondary or annotation data (loosely related to Lemnitzer & Zinsmeister, 2006:40) – the following two sections describe the data material (section 2) and its annotation (section 3) in terms of functional dialogue acts. In the last section, we sketch some research question we will address by means of SoSaBiEC in the very near future.

2. Primary Data

The primary data are made up of audio recordings of everyday conversations (Ries, to appear). The recorded subjects all know each other, most of them are even related. The observations focus on natural language use, and in particular on bilingual language use. The compiled corpus is authentic because the researcher, who recorded, is herself a member of the observed speech community. The speaker gave their consent for recording at any time and without prior notice. So the recordings were taken spontaneously and at real events, such as birthday parties. For recording a digital recorder without microphone was used, so that it was without attracting too much attention. They include telephone calls and face-to-face conversations. The length of the conversations varies from about three minutes up to three hours. Depending on the topic of the conversation the number of the involved speakers differs: from two up to four speakers. In sum, there are about 300 minutes of data material covering six participants. Altogether ten conversations have been recorded. Four conversations have been analysed in detail and annotated because the participant constellation is obvious and definite: the participants come under the category parent-child or sibling. The six participants come from two families, not known to each other. As working basis for the qualitative analysis the recordings were transcribed. By way of illustration, an excerpt of the transcribed data is given:

```
01 F: NAME
    A: guten abend.
    F: hallo?
    A: hallo guten abend
05 F: nabend (.) hallo
    A: na wie gehts bei euch?
    F: gut
    A: gut?
    F: ja.
10 A: на что вы смотрели что к чему там?
    F: ja а что там?
```

This is a sequence of a telephone call between father F and his daughter A. The conversation starts in German and initiated by daughter A there is an alternation into Russian (line 10). The qualitative analysis showed that through this language switch speaker introduced the first

topic of the telephone call and so managed the conversation opening. Results such as the described one are the main content of the annotation.

3. Annotation

The utterances produced by the participants have been transcribed using the Praat tool (<http://www.fon.hum.uva.nl/praat/>) on the level of orthographic words. That means, that no phonetic features like accent or defective pronunciations are coded. However, spoken language exhibits regularities of its own kind, regularities we accounted for in speech transcription. Most prominently, words that are separated in written language might get fused into phonetic word in spoken language. A common example in German already part of the standard of the language is “zum”, a melting of the preposition “zu” and the dative article “dem”. Meltings of this pattern are much more frequent in spoken German than acknowledged in standard German. The English language knows hard-wired combinations like “I’m” which usually is not resolved to the full-fledged standard form “I am”. The annotation takes care for these demands in providing respective adaptations of annotations to spoken language. In order to reveal the dialogue-related functions performed by the utterances, we employed a dialogue act-based coding of contributions. Here, we follow the ISOCat (www.isocat.org) initiative for standardization of dialogue act annotations outlined by Bunt et al. (2010). To be able to talk about dialogue-related functions and natural bilingual language use, language alternations regarding their functions and roles in the current discourse were annotated. The important factor annotated is the function of the involved languages and the observed language alternations: That means to annotate each language switch and its meaning on the level of conversation, for example the conversation opening. The differentiation by speakers is crucial for the examination of a connection between language use and social structure. The functional annotation labels have been derived from qualitative, ethnomethodological analyses by an expert researcher. The annotations made by this vey researcher can be regarded as having the privileged status as “gold standard” since part of the expert’s knowledge is not only the pre- and posthistory of the data recorded, but also familiarity with the subjects involved, a kind of

knowledge rather exclusive to our expert. However, since the annotation are a compromise between the qualitative and quantitative methods and methodologies that are brought together in this kind of research, we want to assess whether the ethnomethodological, functional annotation can be reproduced to a sufficient degree by other annotators. For this reason, we applied a reliability assessment in term of inter-rater agreement of two raters’ annotations of a subset (one conversation) of the data. We use the agreement coefficient AC1 developed by Gwet (2001). The annotation of dialogue acts result in an AC1 of 0.61, the rating of function result in an AC1 of 0.78. Two observations can be made: firstly, the functional dialogue annotation is reproducible -- an outcome of 0.78 is regarded as "substantial" by Rietveld and van Hout (1993); secondly, the standardised dialogue act annotation scheme tailored for task-oriented dialogues can be applied with less agreement than the functional scheme custom-build to more unconstrained everyday conversations. We take this as further evidence for the validity of the distinction of different types argued for in the introduction.

4. Outlook

So far, we finished data collection and annotation of the subset of SoSaBiEC data that interests us first, namely the data that involve parent-child and sibling dialogues. The next step is to test our undirected hypothesis by means of mapping the annotation data on a variant of the dialogue lexicon model of Mehler, Lücking, and Weiß (2010). This model provides a graph-theoretical framework for classifying dialogue networks according to their structural similarity. Applying such quantitative measure onto mostly qualitative data allows not only to study whether social structure imprints on language structure in human dialogue, but in particular to *measure* if there is a traceable influence at all.

5. Acknowledgments

Funding of this work by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung) is gratefully acknowledged. We also want to thank Barbara Job and Alexander Mehler for discussion and support.

6. References

- Anderson, S. R., Keenan, E. L. (1985): "Deixis". In: *Language Typology and Syntactic Description*. Ed. by Timothy Shopen. Vol. III. Cambridge: Cambridge University Press. Chap. 5, pp. 259–308.
- Bock, K. (1996): "Language Production: Methods and Methodologies". In: *Psychonomic Bulletin & Review* 3.4, pp. 395–421.
- Bunt, H. et al. (May 21, 2010): "Towards an ISO Standard for Dialogue Act Annotation". In: *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Ed. by Nicoletta Calzolari (Conference Chair) et al. Valletta, Malta: European Language Resources Association (ELRA).
- Cohen, J. (1960): "A Coefficient of Agreement for Nominal Scales". In: *Educational and Psychological Measurement* 20, pp. 37–46.
- Gwet, K. (2001): *Handbook of Inter-Rater Reliability*. Gaithersburg, MD: STATAXIS Publishing Company.
- Lemnitzer, L., Zinsmeister, H. (2006): *Korpuslinguistik. Eine Einführung*. Tübingen: Gunter Narr Verlag.
- Levinson, S. C. (2008): "Deixis". In: *The Handbook of Pragmatics*. Blackwell Publishing Ltd, pp. 97–121.
- Litman, D. J., Allen, J. F. (1987): "A plan recognition model for subdialogues in conversations". In: *Cognitive Science* 11.2, pp. 163–200.
- Mehler, A. (2008): "On the Impact of Community Structure on SelfOrganizing Lexical Networks". In: *Proceedings of the 7th Evolution of Language Conference (Evolang 2008)*. Ed. By Andrew D. M. Smith, Kenny Smith, and Ramon Ferrer i Cancho. Barcelona: World Scientific, pp. 227–234.
- Mehler, A., Lücking, A., Weiß, P. (2010): "A Network Model of Interpersonal Alignment in Dialogue". In: *Entropy* 12.6, pp. 1440–1483.
doi: 10.3390/e12061440.
- Pickering, M. J. and Garrod, S. (2004): "Toward a Mechanistic Psychology of Dialogue". In: *Behavioral and Brain Sciences* 27.2, pp. 169–190.
- Ries, V. (2011): "da=kommt das=so quer rein. Sprachgebrauch und Spracheinstellungen Russlanddeutscher in Deutschland". PhD thesis. Universität Bielefeld.
- Rietveld, T. van Hout, R. (1993): *Statistical Techniques for the Study of Language and Language Behavior*. Berlin ; New York: Mouton de Gruyter.
- Young, S. J., Proctor, C. E. (1989): "The design and implementation of dialogue control in voice operated database inquiry systems". In: *Computer Speech and Language* 3.4, pp. 329–353.
doi: 10.1016/0885-2308(89)90002-8.

DIL, ein zweisprachiges Online-Fachwörterbuch der Linguistik (Deutsch-Italienisch)

Carolina Flinz

Universität Pisa

E-mail: c.flinz@ec.unipi.it

Abstract

DIL ist ein deutsch-italienisches Online-Fachwörterbuch der Linguistik. Es ist ein offenes Wörterbuch und mit diesem Beitrag wird für eine mögliche Zusammenarbeit, Kollaboration plädiert. DIL ist noch im Aufbau begriffen; zur Zeit ist nur die Sektion DaF komplett veröffentlicht, auch wenn andere Sektionen in Bearbeitung sind. Die Sektion LEX (Lexikographie), die zur Veröffentlichung ansteht, wird zusammen mit den wichtigsten Eigenschaften des Wörterbuches präsentiert.

Keywords: Fachwörterbuch, Linguistik, zweisprachig, deutsch-italienisch, Online-Wörterbuch

1. Einleitung

DIL (*Dizionario tedesco-italiano di terminologia linguistica / deutsch-italienisches Fachwörterbuch der Linguistik*) ist ein online Wörterbuch, das Lemmata aus dem Bereich der Linguistik und einiger ihrer Nachbardisziplinen auflistet. Es ist ein offenes Wörterbuch, nach dem Muster von *Wikipedia*, bzw. *Glottopedia*, um eine mögliche Beteiligung von Experten der unterschiedlichen Disziplinen zu fördern.

Im Handel und im Online-Medium existieren heute mehrere deutsche¹ und italienische² Wörterbücher der Linguistik aber kein einziges Fachwörterbuch für das Sprachenpaar deutsch-italienisch. Hingegen ist der Bedarf an einem solchen „Instrument“ in Italien, sowohl für die universitäre Didaktik als auch für die Forschung, sehr stark: in einem Zeitraum, wo das Fach „Deutsche Linguistik“ als Folge einer Universitätsreform (1999) einen starken Aufschwung erlebt hat, könnte DIL für die wissenschaftliche Kommunikation von großer Relevanz sein³. DIL könnte nämlich eine große Hilfe für die Suche

nach Äquivalenten von deutschsprachigen linguistischen Fachtermini sein.

DIL ist ein Projekt des Deutschen Instituts der Fakultät *Lingue e Letterature Straniere* der Universität Pisa (daf, 2004:37), das 2008 online veröffentlicht worden ist (http://www.humnet.unipi.it/dott_linggensac/glossword) und an dem weiterhin gearbeitet wird. Es handelt sich um ein monolemmatisches Fachwörterbuch⁴ (Wiegand, 1996:46): die Lemmata sind in deutscher Sprache, während die Kommentarsprache Italienisch ist.

Ziele dieses Beitrags sind:

- 1) durch einen kurzen Überblick die wichtigsten Eigenschaften des Wörterbuches vorzustellen, wie Makro- und Mikrostruktur des Wörterbuches, Lemmabestand und Kriterien;
- 2) zu ähnlichen Arbeiten und zukünftigen Kollaborationen an diesem Projekt anzuregen, insbesondere für die geplante Sektion der Computerlinguistik;
- 3) die gerade neu erstellte Sektion LEX (Lexikographie) vorzustellen.

2. Makro- und Mikrostruktur

Die Makrostruktur und die Mikrostruktur von DIL wurden natürlich von der Funktion des Wörterbuches und der intendierten Benutzergruppe beeinflusst⁵. Die Erkundung der Benutzerbedürfnisse wurde mit Hilfe von

¹ Vgl. u.a. Bußmann, 2002; Conrad, 1985; Crystal, 1993; Ducrot & Todorov, 1975; Dubois, 1979; Glück, 2000; Heupel, 1973; Lewandowki, 1994; Meier & Meier, 1979; Stammerjohann, 1975; Ulrich, 2002.

² Vgl. u.a. Bußmann, 2007; Cardona, 1988; Casadei, 1991; Ceppellini, 1999; Courtes & Greimas, 1986; Crystal, 1993; Ducrot & Todorov, 1972; Severino, 1937; Simone, 1969.

³ Die Relevanz von Fachwörterbüchern für die wissenschaftliche Kommunikation war Thema vieler lexikographischer Arbeiten: vgl. u.a. Wiegand, 1988; Pileegard, 1994; Schaeder & Bergenholtz, 1994; Bergenholtz & Tarp, 1995; Hoffmann & Kalverkämper & Wiegand, 1998.

⁴ Eine bilinguistische Ergänzung des Wörterbuches ist nicht ausgeschlossen.

Fragebögen, die sowohl im Printmedium als auch im Onlineformat gesendet wurden, und einer Analyse der möglichen Benutzersituationen erforscht⁶. Jeder Benutzer kann weiterhin den Fragebogen von der Homepage aufrufen und beantworten, so dass ein ständiger Kontakt mit dem Benutzer vorhanden ist.

DIL wendet sich im Allgemeinen an ein heterogenes Publikum: es ist sowohl für Experten als auch für Laien gedacht, so dass die potentiellen Benutzer sowohl Lerner und Lehrender in den Bereichen Germanistik, Romanistik, Linguistik oder Deutsch / Italienisch als Fremdsprache sein können als auch Lehrbuchautoren, Lexikographen oder Fachakademiker. Das Online Medium, dank seiner Flexibilität, ist von großem Vorteil in dieser Hinsicht.

DIL kann nämlich in folgenden Benutzungssituationen verwendet werden:

- 1) Der Benutzer sucht bestimmte fachliche Informationen, und das Wörterbuch, laut seiner Werkzeugnatur, erfüllt das Bedürfnis;
- 2) Der Benutzer greift zum Wörterbuch, um ein Kommunikationsproblem in der Textproduktion, Textrezeption oder Übersetzung zu lösen. DIL erfüllt deswegen mehrere Funktionen: es kann sowohl für aktive / produktive als auch passive / rezeptive Tätigkeiten verwendet werden.
 - a. Der italophone Benutzer (primärer Benutzer) wird es als dekodierendes Wörterbuch für die Herübersetzung verwenden, d.h. wenn er ein deutsches Fachwort verstehen will oder dessen Übersetzung sucht, oder wenn er spezifischere Informationen braucht und sich weiter informieren und weiterbilden möchte;
 - b. Der deutschsprachige Benutzer wird es als enkodierendes Wörterbuch für die Hinproduktion benutzen, d.h. wenn er ins Italienische übersetzt und Fachtexte in italienischer Sprache erstellt.

Die Makrostruktur von DIL vereinigt sowohl Eigenschaften der linguistischen Printwörterbücher (1.)

⁵ Vgl. u.a. Storrer & Harriehausen, 1998; Barz, 2005.

⁶ Für einen Überblick über mögliche Techniken zur (Benutzerbedürfnissen-Erforschung: meglio zur Erforschung von Benutzerbedürfnissen) vgl. u.a. Barz, 2005; Ripfel & Wiegand, 1988; Schaefer & Bergenholtz, 1994; Wiegand, 1977.

als auch der Onlinewörterbücher (2.):

1. Die Strukturierung der Umtexte im Printmedium beeinflusste den aktuellen Stand. DIL verfügt nämlich über folgende nach wissenschaftlichen Kriterien verfasste Umtexte: Einleitung, Abkürzungsverzeichnis, Benutzerhinweise, Redaktionsnormen, Register der Einträge⁷;
2. Die Vorteile der Online-Wörterbücher wurden auch zum größten Teil ausgenutzt:
 - a. Neue Einträge und neue Sektionen können sehr schnell veröffentlicht werden;
 - b. DIL kann ständig erneuert, ergänzt und korrigiert werden;
 - c. Es verfügt über ein klar strukturiertes Menu, in dem die wichtigsten Umtexte verlinkt sind, so dass der Benutzer schnell die gewünschten Informationen erreichen kann;
 - d. Es verwendet sowohl interne⁸ als auch externe⁹ Hyperlinks;
 - e. Es bietet dem Benutzer nützliche Informationen, wie die TOP 10 (vgl. u.a. die „zuletzt gesuchten“ oder „die am meisten geklickten Lemmata“);
 - f. Es bietet wichtige Instrumente, wie die Suchmaschine, die Feedbackseite, das Login Feld etc.

Die Mikrostruktur von DIL bietet sowohl sprachliche als auch sachliche Informationen und ist auf der Grundlage, dass der Erst-Adressat der italophone Benutzer ist, strukturiert worden. Jeder Eintrag wird von folgenden Angaben komplettiert:

- 1) grammatische Angaben (Genus und Numerus);
- 2) das Äquivalent / die Äquivalente in italienischer Sprache;
- 3) die Markierung als Information zum fachspezifischen Bereich des Lemmas;
- 4) die enzyklopädische Definition;
- 5) Beispiele¹⁰;

⁷ Eine empirische Analyse linguistischer Online-Fachwörterbücher zeigte, wie „unwissenschaftlich“ oft Online-Wörterbücher mit Umtexten umgehen. Nur 45% der analysierten Werkzeuge verfügte über solche Texte und nur in seltenen Ausnahmen wurden wissenschaftlichen Kriterien gefolgt (Flinz, 2010:72)

⁸ Der Benutzer kann von einem Eintrag zu thematisch verbundenen Lemmata springen.

⁹ Es sind sowohl sprachliche Wörterbücher, wie *Canno.net* und *Grammis*, als auch sachliche, wie *Glottopedia* und *DLM*, verlinkt.

¹⁰ Alle Lemmata folgen im Prinzip dem gleichen Schema, da

- 6) Angaben zur Paradigmatik, wie Synonyme; thematisch verbundene Lemmata;
- 7) bibliographische Angaben.

3. Lemmabestand und Kriterien

Der Lemmabestand von DIL kann nur „eingeschätzt“ werden. Die Gründe dafür können wie folgt zusammengefasst werden:

- 1) erstens handelt es sich um ein Online-Wörterbuch, das noch in Projekt und Testphase ist;
- 2) zweitens soll das Werk, wie es sein Format vorgibt, nicht als etwas Statisches und Vollendetes gesehen werden, sondern in ständiger Erweiterung und Erneuerung. Aus einem Vergleich der existierenden linguistischen Fachwörterbücher kann aber eine ungefähre Zahl von ca. 2.000 Lemmata ausgerechnet werden, die allerdings ständig erweitert oder geändert werden kann.

Primärquellen waren allgemeine Wörterbücher der Linguistik (deutsch- wie italienischsprachige), sowie spezifische deutsche und italienische Glossare der Disziplin Lexikographie und Fachlexikographie. Es wurden Quellen sowohl im gedruckten als auch im Online-Medium herangezogen. Sekundärquellen waren Handbücher aus dem Bereich der jeweiligen Disziplin (für die Sektion **Lex** waren es zum Beispiel Standardwerke der Disziplin Lexikographie und Fachlexikographie) sowohl in deutscher als auch in italienischer Sprache.

Hauptkriterien für die Auswahl der Lemmata sind Frequenz und Relevanz (Bergenholtz, 1989:775)¹¹:

- 1) Es wurde eine entsprechende Analyse der existierenden lexikographischen Wörterbücher sowohl im Print- als auch im Onlineformat hinsichtlich der dort aufgeführten Lemmata des jeweiligen Bereiches durchgeführt;
- 2) Es wurde ein kleiner Korpus von Fachtexten des betreffenden Faches hergestellt. Die im Endregister enthaltenen Termini wurden in Excell-Tabellen

eingetragen, und die entstehenden Listen wurden auf Grund von Frequenzkriterien verglichen. Die aus diesem Prozess entstehende Endliste wurde zusätzlich auf der Basis des Relevanzkriteriums ergänzt.

Die Einträge sind in strikt alphabetischer Reihenfolge, und die typischen Nachteile dieser Ordnung können dank des Online-Formats teilweise aufgehoben werden, da die begriffssystematischen Zusammenhänge durch verlinkte Verweise oft verdeutlicht werden.

Das Wörterbuch enthält zurzeit eine vollständige Sektion (DaF) mit 240 Einträgen, während andere Bereiche in Erarbeitung sind:

- 1) Historische Syntax;
- 2) Wortbildung;
- 3) Textlinguistik;
- 4) Fachsprachen.

Eine neu erstellte Sektion LEX (Lexikographie) wurde gerade fertiggestellt und steht zur Veröffentlichung an. Sie enthält Lemmata aus dem Bereich der Lexikographie und Fachlexikographie sowie Metalexikographie und Metafachlexikographie.

4. Die Sektion: LEX

Die Sektion LEX wird voraussichtlich ca. 120 Einträge (Stand Juni 2011) enthalten, die sich auf die wichtigsten Aspekte des Fachbereiches der Lexikographie konzentrieren. Es wird dabei auf folgende Themen Aufmerksamkeit gelegt:

- a. Lexikographie;
- b. Fachlexikographie;
- c. Wörterbuchforschung;
- d. Wörterbuchtypologie;
- e. Wörterbuchbenutzer und Wörterbuchbenutzung;
- f. Wörterbuchfunktionen;
- g. lexikographische Kriterien;
- h. Makrostruktur;
- i. Umtexte;
- j. Mediostruktur;
- k. Mikrostruktur.

Im Folgenden wird ein Beispiel eines Eintrags aus dem Bereich **Lex** gezeigt. Es kann als Muster für die Erarbeitung von neuen Einträgen gelten. Jeder Autor kann die produzierten Lemmata an die Redaktion des Wörterbuches senden; nach der redaktionellen Prüfung wird der Eintrag veröffentlicht und mit der Abkürzung des Autorennamens vermerkt.

die Standardisierung der Mikrostruktur eine wichtige Voraussetzung war. Da aber Beispiele nur in bestimmten Kontexten behilflich sind, wurden sie nur gelegentlich eingefügt.

¹¹ Korpusanalysen, im Sinne von automatischen Analysen von Textkorpora mit anschließender Korpusauswertung (Frequenzwerte) wurde bis jetzt ausgeschlossen. Jedoch wäre es interessant zu sehen, inwiefern eine solche Analyse mit einer Integrierung des Relevanzkriteriums die erhaltenen Ergebnisse widerspiegeln könnte oder nicht.

Fachlexikographie

(LEX); (die, PLURAL UNÜBLICH)

Lessicografia specialistica

Termine che indica l'attività scientifica, il cui obiettivo primario è la produzione (ossia la pianificazione, la redazione, la revisione, la stampa ecc.) di dizionari specialistici linguistici, di dizionari enciclopedici e dizionari linguistico-enciclopedici. Pertanto si distingue in:

- lessicografia specialistica linguistica;
- lessicografia specialistica enciclopedica;
- lessicografia linguistica-enciclopedia.

Da coloro che la praticano la lessicografia specialistica pretende non solo una solida base lessicografica, ma anche un competenza nel linguaggio specialistico oggetto del dizionario.

La lessicografia specialistica si attesta come disciplina scientifica a partire da metà degli anni '90, infatti fino ad allora l'interesse per i dizionari scientifici era marginale e spesso si lamentava la mancanza di lavori che si occupassero in modo scientifico di dizionari specialistici. I primi articoli pubblicati trattano ancora la disciplina in modo generico; il primo articolo ufficiale viene considerato quello di Wiegand, H. E. (1988): *Was ist eigentlich Fachlexikographie?*. Dagli anni '90' vengono organizzati le prime giornate di studio e vengono pubblicate le prime bibliografie. (cf)

Vedi anche: [Lexikographie](#)

Fonte: SCHAEFER, B. – BERGENHOLTZ, H. (1994): *Fachlexikographie. Fachwissen und seine Repräsentation in Wörterbüchern.*, WIEGAND, H.E. (1988): *Was ist eigentlich Fachlexikographie?*. In Munske, H.H. – Von Polenz, P. – Reichmann, O. – Hildebrandt, R. (Hrsg.) 1988: S. 729-790., WIEGAND, H.E. (1976): *Die Wahrheit der Wörterbücher. In: Probleme der Lexikologie und Lexikographie. Jahrbuch 1975 des Instituts für deutsche Sprache, Düsseldorf (Sprache der Gegenwart XXXIX).* S. 347-371., WIEGAND, H.E. (1988): *Was ist eigentlich Fachlexikographie?*. In Munske, H.H. – Von Polenz, P. – Reichmann, O. – Hildebrandt, R. (Hrsg.) 1988: S. 729-790., PILEGAARD, M. (1994): *Bilingual LSP Dictionaries. User benefit correlates with elaborateness of „explanation“.* In: Bergenholtz, H. - Schaefer, B. 1994. S. 211-228., KUCERA, A. (1984): *Aus der Werkstatt der praktischen Verlagslexikographie. Übersetzungswörterbücher der Fachsprachen.* In: Mitteilungen für Dolmetscher und Übersetzer 1/30. S. 3-6.

Bild 1: Das Lemma "Fachlexikographie"

5. Literatur

- Abel, A. (2006): Elektronische Wörterbücher: Neue Wege und Tendenzen. In San Vincente, F. (Ed.) Akten der Tagung "Lessicografia bilingue e Traduzione: metodi, strumenti e approcci attuali" (Forlì, 17.-18.11.2005). Polimetrica Publisher (Open Access Publications). S. 35-56.
- Almind, R. (2005): Designing Internet Dictionaries. *Hermes*, 34, S. 37-54.
- Barz, I., Bergenholtz, H., Korhonen, J. (2005): Schreiben, Verstehen, Übersetzen, Lernen. Zu ein- und zweisprachigen Wörterbüchern mit Deutsch. Frankfurt a. M.: Peter Lang.
- Bergenholtz, H. (1989): Probleme der Selektion im allgemeinen einsprachigen Wörterbuch. In Hausmann, F. J. et al. (Hg). *Wörterbücher: ein internationales Handbuch zur Lexikographie*. Band 1. Berlin & New York: de Gruyter. S. 773-779.
- Bergenholtz, H., Tarp, S. (1995): *Manuel of LSP lexikography. Preparation of LSP dictionaries - problems and suggested solutions.* Amsterdam, Netherlands & Philadelphia: J. Benjamins.
- Foschi-Albert, M., Hepp, M. (2004): Zum Projekt: Bausteine zu einem deutsch-italienischen Wörterbuch der Linguistik. In *daf Werkstatt*, 4, S. 43-69.
- Hoffmann, L., Kalverkämper, H., Wiegand, H.E (Eds.) (1999): *Fachsprachen. Handbücher zur Sprach- und Kommunikationswissenschaft (HSK 14.2.)*. Berlin & New York: de Gruyter.
- Pilegaard, M. (1994): *Bilingual LSP Dictionaries. User benefit correlates with elaborateness of „explanation“.* In Bergenholtz, H. & Schaefer, B. S. 211-228.
- Schaefer, B., Bergenholtz, H. (1994): *Fachlexikographie. Fachwissen und seine Repräsentation in Wörterbüchern*. Tübingen: G. Narr.
- Ripfel M., Wiegand, H.E. (1988): Wörterbuchbenutzungsforschung. Ein kritischer Bericht. In *Studien zur Neuhochdeutschen Lexikographie VI*. 2. Teilb. S. 482-520.
- Storrer, A., Harriehausen, B. (1998): *Hypermedia für Lexikon und Grammatik*. Tübingen: G. Narr.
- Wiegand, H.E. (1977): *Nachdenken über Wörterbücher. Aktuelle Probleme*. In Drosdowski, H., Henne, H. & Wiegand, H.E. *Nachdenken über Wörterbücher*. Mannheim: Bibliographisches Institut / Dudenverlag. S. 51-102.
- Wiegand, H.E. (Ed.) (1996): *Wörterbücher in der Diskussion II. Vorträge aus dem Heidelberger Lexikographie-Kolloquium*. Tübingen: Lexicographica Series Major 70.
- Wiegand, H.E. (1988): *Was ist eigentlich Fachlexikographie?*. In Munske, H.H., Von Polenz, P. & Reichmann, O. & Hildebrandt, R. (Hg.). *Deutscher Wortschatz. Lexikologische Studien*. Berlin & New York: de Gruyter. S. 729-790.

Knowledge Extraction and Representation: the EcoLexicon Methodology

Pilar León Araúz, Arianne Reimerink

Department of Translation and Interpreting, University of Granada

Buenucesos 11, 18002, Granada, Spain

E-mail: pleon@ugr.es, arianne@ugr.es

Abstract

EcoLexicon, a multilingual terminological knowledge base (TKB) on the environment, provides an internally coherent information system which aims at covering a wide range of specialized linguistic and conceptual needs. Knowledge is extracted through corpus analysis. Then it is represented and contextualized in several dynamic and interrelated information modules. This methodology solves two challenges derived from multidimensionality: 1) it offers a qualitative criterion to represent specialized concepts according to recent research on situated cognition (Barsalou, 2009), and 2) it is a quantitative and efficient solution to the problem of information overload.

Keywords: knowledge extraction, knowledge representation, EcoLexicon, multidimensionality, context

1. Introduction

EcoLexicon¹ is a multilingual knowledge base on the environment. So far it has 3,283 concepts and 14,695 terms in Spanish, English and German. Currently, two more languages are being added: Modern Greek and Russian. It is aimed at users such as translators, technical writers, environmental experts, etc., which can access it through a friendly visual interface with different modules devoted to both conceptual, linguistic, and graphical information.

In this paper, we will focus on some of the steps applied to extract and represent conceptual knowledge in EcoLexicon. According to Meyer et al. (1992), terminological knowledge bases (TKBs) should reflect conceptual structures in a similar way to how concepts relate in the human mind. The organization of semantic information in the brain should thus underlie any theoretical assumption concerning the retrieval and acquisition of specialized knowledge concepts as well as the design of specialized knowledge resources (Faber, 2010). In Section 2, we explain how knowledge is extracted through corpus analysis. In Section 3, we show how conceptual knowledge is represented and contextualized in dynamic and interrelated networks.

2. Conceptual Knowledge Extraction

According to corpus-based studies, when a term is studied in its linguistic context, information about its meaning and its use can be extracted (Meyer & Mackintosh, 1996). In EcoLexicon, the corpus consists of specialized (e.g. scientific journal articles, thesis, etc.), semi-specialized texts (textbooks, manuals, etc.) and texts for the general public, all in the multidisciplinary domain of the environment. Each language has a separate corpus and the knowledge is extracted bottom-up from each of the corpora. The underlying ontology is language independent and based on the knowledge extracted from all the corpora. The extraction of conceptual knowledge combines direct term searches and knowledge pattern (KP) analysis. According to many studies on the subject, KPs are considered one of the most reliable methods for knowledge extraction (Barrière, 2004). Normally, the most recurrent knowledge patterns (KPs) for each conceptual relation identified in previous research are used to find related term pairs (Auger & Barrière, 2008). Afterwards, these terms are used for direct term searches to find new KPs and relations. Therefore, the methodology consists of the cyclic repetition of both procedures.

When searching for the term EROSION, conceptual concordances show how different KPs convey different

¹ <http://ecolexicon.ugr.es>

relations with other specialized concepts. The main relations are *caused_by*, *affects*, *has_location* and *has_result*, which highlight the procedural nature of the concept and the important role played by non-hierarchical relations.

In Figure 1, EROSION is related to its diverse kinds of

agents, such as STORM SURGE (1, 7), WAVE ACTION (2, 13), RAIN (3), CONSTRUCTION PROJECTS (6) and HUMAN-INDUCED FACTORS (11). They can be retrieved thanks to all KPs expressing the relation *caused_by*, such as *resultant* (1), *agent for* (2, 3), *due to* (6, 7), and *responsible for* (11).

```

Caused_by
1 Alabama. Significant storm surge and resultant beach erosion were associated with Ivan's landfall. However,
2 nd climate on the Castellon coast, the main agent for erosion is wave action, and this is therefore responsi
3 f a stream. The first factor, rain, is the agent for erosion, but the degree of erosion is governed by oth
4 rts (SW) and semiarid steppe (BS). Wind can also cause erosion and deposition in environments where sediments
5 ety. Reflection of waves from a jetty may also cause erosion of adjacent shorelines. However, erosion furthe
6 ostial zone management. However, in some cases coastal erosion can be due to construction projects that a
7 tude of about 0.3 M m3 per year. Acute erosion Acute erosion due to storm surges (waves and water levels at
8 er. Mangrove removal is also reported to cause coastal erosion and change sedimentation patterns and shoreline
9 [edit] Erosion surface runoff is one of the causes of erosion of the earth's surface. Reduced crop product
10 pes. Local disturbances, for instance by flood-induced erosion, redistribution of sediment or accumulation of
11 ors and human-induced factors, responsible for coastal erosion and highlight the time and space patterns withi
12 cess is typical of a cyclical process of storm-caused erosion in winter, followed by progradation owing
13 can cause excessive wave action that can lead to beach erosion. Trash dumped from boats can be washed up onto
14 that have reached base level develop broad valleys by erosion caused by meandering channels. The stream chain

Affects
15 ing these sensitive creatures. In some cases, coastal erosion can have adverse effects on water quality and h
16 use of dredged material to restore beaches damaged by erosion. EPA works with the U.S. Coast Guard to regul
17 reasonable points, though when push comes to shove and erosion threatens buildings, traditional beach maintena
18 ks and arches found on irregular rocky coastlines; and erosion provides the material which forms deltas and ba
19 near the base of the cliff. This process undercuts and erosion causes the cliffs to retreat landward.

Has_location
21 ed by the position of sand accumulation and beach erosion around littoral barriers. A coastal structure i
22 hes. Kuenen (1950) estimates that beach and cliff erosion along all coasts of the world totals about 0.12g
23 ce and divergence of wave energy over an offshore bar. erosion downdrift of a structure such as a groin, sudd
24 proportional to the longshore transport rate, and erosion takes place downdrift at about the same rate. T

Has_result
25 Excessive loads of silt and other sediments caused by erosion can suffocate bottom-dwelling plants and animal
26 islands or coral reefs. Primary coasts are created by erosion (the wearing away of soil or rock), deposition
27 \par transported. Beach material is also derived from erosion of the coastal formations caused by waves
28 ed to the passage of the ice. Shorelines produced by erosion of glacial till deposits differ markedly from
29 beaches and marshes, are being formed as a result of erosion also transportation of unconsolidated material
30 ion of the seashore and a rise in s.l.w. The results of erosion could lead to further seawater intrusion that c
31 fs are deposited in landslide debris. In this cliffs, erosion of softer material has created bays. The expect
32 s of steep systems, a sea-level rise may cause coastal erosion resulting in profile steepening, and therefore
    
```

Figure 1: Non-hierarchical relations associated with EROSION

```

Is_a
33 vided by the area (A) of the drainage basin (L) Erosion is the natural process of removal of soil by wa
34 in the Netherlands, geomorphological processes such as erosion, transport and sedimentation of sandy materials
35 BURY AND DUXBURY, 1996). coastal processes such as erosion and accretion are site-specific, season-specific
36 these catchments include: stormwater impacts such as erosion, channelisation, sediment deposition and sediment

Type_of
37 eroded by shallow overland flow (sheet, rill and gully erosion) and delivered to the drainage network. Channel
38 m the great local relief, the result of differential erosion by glacier ice. Figure 9-20 includes two sche
39 ing flood events, the dikes are subject to the lateral erosion of the river trying to reoccupy its former coa
40 d enlarges these small channels and generates headward erosion directed towards the aggrading active channel (
41 out five percent of the material on most beaches, wave erosion of rocky coasts is usually slow, even where the
42 of the earth's land surface is dominated by fluvial erosion. Lakes that do occur are threatened with either
43 ind climate, topography and surface roughness. wind erosion risk applies only when soils are dry and not c
44 opportional to the steepness of the land surface, water erosion is in proportion to the shear stress exerted by
45 lay to become both wider and deeper over time. Glacial erosion also results in a change in the valley's cross-
46 dominate in periglacial environments: nivation; eolian erosion and deposition; and fluvial erosion and deposit
47 erosion processes. 215 CHAPTER 13 EQUATIONS: SEDIMENT Erosion caused by rainfall and runoff is computed with
48 givers to simulate cross-shore beach, berm, and dune erosion produced by storm waves and water levels. The l
49 uctures constructed to date have resulted in shoreline erosion in their lee. Furthermore, the key environmen
    
```

Figure 2: Hierarchical relations associated with EROSION

This relation can also be conveyed through compound names such as *flood-induced* (10) or *storm-caused* (12) and any expression containing *cause* as a verb or noun: *one of the causes of* (9), *cause* (4, 5, 8) and *caused by* (14). EROSION is also linked to the patients it *affects*, such as WATER (15), SEDIMENTS (16), and BEACHES (17). However, the affected entities, or patients, are often equivalent to locations (eg. if EROSION *affects* BEACHES it actually *takes place at* the BEACH). The difference lies in the kind of KPs linking the propositions. The *affects* relation is often reflected through the preposition *of* (10) or verbs like *threatens* (18), *damaged by* (17) or *provides* (19), whereas the *has_location* relation is conveyed through prepositions linked to directions (*around*, 21; *along*, 22; *downdrift*, 23) or spatial expressions such as *takes place* (24). In this way, EROSION appears linked to the following locations: LITTORAL BARRIERS (21), COASTS (22) and STRUCTURES (23). *Result* is an essential

dimension in the description of any process, since it also has certain effects, which can be the creation of a new entity (SEDIMENTS, 25; MARSHES, 29; BAYS, 31) or the beginning of another process (SEAWATER INTRUSION, 31; PROFILE STEEPENING, 32).

All these related concepts are quite heterogeneous. They belong to different paradigms in terms of category membership or hierarchical range. For instance, some of the agents of EROSION are natural (WIND, WAVE ACTION) or artificial (JETTY, MANGROVE REMOVAL) and others are general concepts (STORM) or very specific (MEANDERING CHANNEL). This explains why knowledge extraction must still be performed manually, but it also illustrates one of the major problems in knowledge representation: multidimensionality (Rogers, 2004).

This is better exemplified in the concordances in Figure 2, since multidimensionality is most often codified in the *is_a* relation. In the scientific discourse community,

HYDROLOGY (Figure 5).

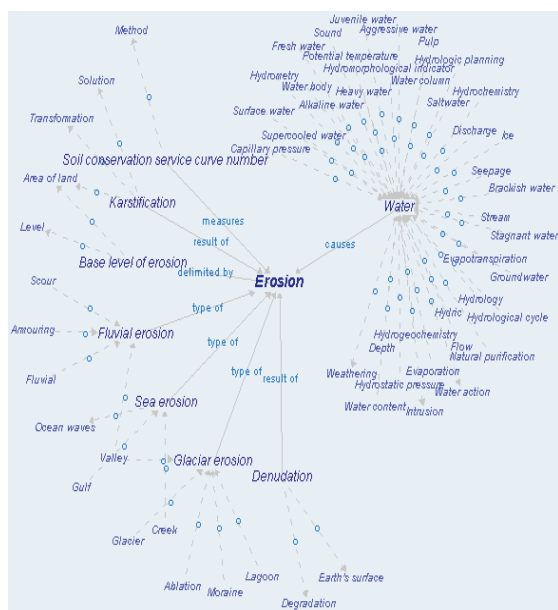


Figure 5: EROSION in the HYDROLOGY domain

Comparing both networks and especially focusing on EROSION and WATER, the following conclusions can be drawn. The number of conceptual relations changes from one network to another, as EROSION is not equally relevant in both domains. EROSION is a prototypical concept of the GEOLOGY domain, this is why it shows more propositions. Nevertheless, since it is also strongly linked with WATER, the HYDROLOGY domain is also essential in the representation of EROSION. Relation types do not substantially change from one network to the other, but the GEOLOGY domain shows a greater number of *type_of* relations. This is due to the fact that the HYDROLOGY domain only includes types of EROSION whose agent is WATER, such as FLUVIAL EROSION and GLACIER EROSION. The GEOLOGY domain includes those and others, such as WIND EROSION, SHEET EROSION, ANTHROPIC EROSION, etc. The GEOLOGY domain, on the other hand, also includes concepts that are not related to HYDROLOGY such as ATTRITION because there is no WATER involved.

On the contrary, WATER displays more relations in the HYDROLOGY domain. This is caused by the fact that WATER is a much more prototypical concept in HYDROLOGY. Therefore, its first hierarchical level shows more concepts. For example, in GEOLOGY, there are less WATER subtypes because the network only shows those that are related to the geological cycle (MAGMATIC WATER, METAMORPHIC WATER, etc.). In HYDROLOGY,

there are more WATER subtypes related to the hydrological cycle itself (SURFACE WATER, GROUNDWATER, etc.). Even the shape of each network illustrates the prototypical effects of WATER or EROSION. In Figure 4, EROSION is displayed in a radial structure that shows it as a central concept in GEOLOGY, whereas in Figure 5, the asymmetric shape of the network implies that, more than EROSION, WATER is the prototypical concept of HYDROLOGY.

4. Acknowledgements

This research has been carried out in project FFI2011-22397/FILO funded by the Spanish Ministry of Science and Innovation.

5. References

- Auger, A., Barrière, C. (2008): Pattern-based approaches to semantic relation extraction: A state-of-the-art. Special Issue on Pattern-Based Approaches to Semantic Relation Extraction, Terminology, 14(1), pp. 1–19
- Barrière, C. (2004): Knowledge-rich contexts discovery. In Proceedings of the 17th Canadian Conference on Artificial Intelligence (AI'2004). May 17–19, London, Ontario, Canada.
- Barsalou, L.W. (2009): Simulation, situated conceptualization and prediction. Philosophical Transactions of the Royal Society of London: Biological Sciences, 364, pp. 1281–1289.
- Faber, P. (2010): Conceptual modelling in specialized knowledge resources. In XII International Conference Cognitive Modelling in Linguistics. September, Dubrovnik.
- León Araúz, P., Faber, P. (2010): Natural and contextual constraints for domain-specific relations. In Proceedings of Semantic relations. Theory and Applications. 18–21 May, Valetta, Malta.
- Meyer, I., Mackintosh, K. (1996): The corpus from a terminographer's viewpoint. International Journal of Corpus Linguistics, 1(2), pp. 257–285.
- Meyer, I., Bowker, L., Eck, K. (1992): COGNITERM: An experiment in building a knowledge-based term bank. In Proceedings of Euralex '92, pp. 159–172.
- Rogers, M. (2004): Multidimensionality in concepts systems: a bilingual textual perspective. Terminology, 10(2), pp. 215–240.

Processing Multilingual Customer Contacts via Social Media

Michaela Geierhos, Yeong Su Lee, Matthias Bargel

Center for Information and Language Processing (CIS)

Ludwig Maximilian University of Munich

Geschwister-Scholl-Platz 1, D-80539 München, Germany

E-mail: micha@cis.uni-muenchen.de, yeong@cis.uni-muenchen.de, matthias@cis.uni-muenchen.de

Abstract

Within this paper, we will describe a new approach to customer interaction management by integrating social networking channels into existing business processes. Until now, contact center agents still read these messages and forward them to the persons in charge of customer's in the company. But with the introduction of Web 2.0 and social networking clients are more likely to communicate with the companies via Facebook and Twitter instead of filling data in contact forms or sending e-mail requests. In order to maintain an active communication with international clients via social media, the multilingual consumer contacts have to be categorized and then automatically assigned to the corresponding business processes (e.g. technical service, shipping, marketing, and accounting). This allows the company to follow general trends in customer opinions on the Internet, but also record two-sided communication for customer relationship management.

Keywords: classification of multilingual customer contacts, contact center application support, social media business integration

1. Introduction

Considering that Facebook alone had more than 750 million active users¹ in August 2011 it becomes apparent that Facebook currently is the most preferred medium by consumers and companies alike. Since many businesses are moving to online communities as a means of communicating directly with their customers, social media has to be explored as an additional communication channel between individuals and companies. While the English speaking consumers on Facebook are more likely to respond to communication rather than to initiate communication with an organization (Browne et al., 2009), the German speaking community in turn directly contacts the companies. Therefore, some German enterprises already have regularly updated Facebook pages for customer service and support, e.g. Telekom. Using the traditional communication channels such as telephone and e-mail, there are already established approaches and systems to incoming requests. They are used by companies to manage all client contacts through a variety of mediums such as telephone, fax, letter, e-mail, and online live chat. Contact center agents are therefore

responsible to assign all customer requests to internal business processes. However, social networking has not yet been integrated into customer interaction management tools.

1.1. Related Work

With the growth of social media, companies and customers now use sites such as Facebook and Twitter to share information and provide support. More and more integrated cross-platform campaigns are dealing with product opinion mining or providing web traffic statistics to analyze customer behavior. There is a plenty of commercial solutions, of varying quality, for these tasks, e.g. GoogleAlerts, BuzzStream, Sysomos, Alterian, Visible Technologies, and Radian6.

The current trend goes to development of virtual contact centers integrating company's fan profiles on social networking sites. This virtual contact center processes the customer contacts and forwards them to company's service and support team. For instance, Eptica provides a commercial tool for customer interaction management via Facebook.

Other monitoring systems try to predict election results (Gryc & Moilanen, 2010) or success of movies and music

¹ <http://www.facebook.com/press/info.php?statistics>

(Krauss et al., 2008) by using scientific analysis of opinion polls or doing sentiment analysis on special web blogs or online forum discussions. Another relevant issue is the topic and theme identification as well as sentiment detection. Since blogs consist of news or messages dealing with various topics, blog content has to be divided into several topic clusters (Pal & Saha, 2010).

1.2. Towards a Multilingual Social Media Customer Service

Our proposed solution towards a web monitoring and customer interaction management system is quite simple. We focus on a modular architecture fully configurable for all components integrated in its work-flow (e.g. software, data streams, and human agents for customer service). Our first prototype, originally designed for processing customer messages posted on social networking sites about mobile-phone specific issues, can also deal with other topics and use different text types such as e-mails, blogs, RSS feeds etc. Unlike the commercial monitoring systems mentioned above, we concentrate on a linguistic, rule-based approach for message classification and product name recognition. One of its core innovations is its paraphrasing module for intra- and inter-lingual product name variations because of different national and international spelling rules or habits. By mapping product name variations to an international canonical form, our system allows for answering questions like *Which statements are made about this mobile phone in which languages/in which social networks/in which countries?* Its product name paraphrasing engine is designed in such a way that standard variants are assigned automatically, regular variants are assigned semi-automatically and idiosyncratic variants can be added manually. Moreover, our system can be adapted according to user's language needs, i.e. the application can be easily extended on further natural languages. Until now, our prototype can deal with three very different languages: German, Greek, and Korean. It therefore provides maximum flexibility to service providers by enabling multiple services with only one system.

2. System Overview

Since customers first share their problems with a social networking community before directly addressing the company, the social networking site will be the interface

between customer and company. For instance, Facebook users post on the wall of a telecommunication company messages concerning tariffs, technical malfunction or bugs of its products, positive and negative feedback. The collector should download every n seconds (e.g. 10 sec) data from the monitored social networking site. Above all it should be possible to choose the social networking site, especially the business pages, to be monitored. This can be configured by updating the collector's settings. In order to retrieve data from Facebook, we use its graph API. Then customer messages will be stored in a database. After simplifying their structure², the requests have to be categorized by the classification module. During the classification process, we assign both content and semantic tags (cf. Sect. 3.2) as features to the user posts before re-storing them in a database. According to the tags the messages are assigned to the corresponding business process. This $n : 1$ relationship is modeled in the contact center interface before passing these messages as e-mail requests to the customer interaction management tool used in contact centers. Finally, the pre-classified e-mails are automatically forwarded to the persons in charge of customer services. Those agents reply to the client requests and their responses will be delivered via e-mail to the contact center before being transformed into social network messages and sent back to the Facebook wall. Afterwards, the Facebook user can read his answer.

3. Linguistic Processing of Customer Contacts

Within the customer requests, we try to discover relationships between clients and products, customers and technical problems, products and features that will be used for classification purposes. We are aware of the fact that many products (mobile phones, chargers, headsets, batteries, software, and operating systems) are sold in different countries under the same or under different names. Our system stores a unique international ID for each product. Product names and their paraphrases are language specific. Our prototype normalizes found product names to the international ID.

² For example, Facebook wall posts are represented as structured data that can easily be retrieved from Facebook graph API. We simplify this data format before using it for extraction and classification purposes.

3.1. International Product Name Paraphrasing

Our first approach to product name paraphrasing was to use *paraphrasing classes*. Much as verbs are inflected according to their inflection class, product names were inflected according to their paraphrasing class. Yet, paraphrasing classes had to be assigned manually and quite many classes were needed. Therefore, we decided to use a simplified system: Each product or manufacturer name is stored in a canonical form: Thus, a name of the type *glofiish g500* is stored in the form *glofiish-g-500*, even if *glofiish g-500* or *glofiish g500* should be more frequent. The minus characters tell our system where a new part of the product name begins. A product or manufacturer name has *permutations*: In German *o2 online tarif* has the permutation *tarif o2 online*. Standard permutations are added automatically: A product or manufacturer name with three parts has the standard permutation *123*. German tariff names of the type *o2 online tarif* have the standard permutations *312* and *23 von 1* as in *online tarif von O2 (online tariff by o2)*.

Apart from their canonical name and its variants, product names can also have spelling variants. Thus, *android* has the spelling variants *androit*, *antroid*, *antroit*, *andorid*, *adroid*, *andoid* and *andoit*. (These are some of the most frequent ways *android* is actually spelt in the customer messages.) For each spelling variant, our system automatically generates all paraphrases that exist according to the standard and the manually added permutations of the canonical name. I.e. the paraphrases of the mobile phone name *e-ten glofiishg-500* include *e-ten klofisch-g-500*, *e-ten klofisch-g 500*, *e-ten klofisch g-500*, etc.

Apart from spelling variants, product names can also have lexical variants. The mobile phone *tct mobile one-touch-v-770-a* has the lexical variant *playboy-phone*. The regular permutation transformations are not applied to lexical variants. But lexical variants and their manufacturer-based variants (e.g. *tct playboy-phone* and *playboy-phone*) are, of course, paraphrases, too.

3.2. Grammar-based classification

Grammar experts can create any number of content and sentiment classifiers. A classifier's grammar consists of a set of positive constraints and a set of negative constraints. To classify a message, our system simply applies the grammars of all its classifier objects to the

message. If a content classifier's grammar matches, its tag is added to the message's content tags. Sentiment classification works analogously with the exception that exactly one tag is assigned.

Content and sentiment classifiers are language and URL specific: A classifier has exactly one language and a set of URLs. It will only be applied to messages that have the same language and that stem from one of the URLs in the classifier's set of URLs. In general, content tags and product list are independent of each other. But many classifiers will have constraints that require that a product (or other entity) of a certain type be mentioned. Thus, a classifier that assigns the tag *phone available?* (e.g. to the message *When will the new iPhone be released?*) would probably include the mobile phone grammar in its constraints by using the special term `\mobile_phone`.

4. Discussion

4.1. No statistical approach

We think that the fact that contact center agents can invent new tags and assign new or old tags to (badly) classified messages, if they mark the strings that are supposed to justify the assignment of the tag, is a good reason for not using a statistical approach. If we used a statistical approach, human work would be necessary at some point of the development process: Some algorithm would have to be trained. In our approach, the human work is done in the customer management process. This way, two things are achieved in one step: The customer's request is answered and the classification algorithm is enhanced. The system is being enhanced while it is used. There is no need to interrupt the customer interaction in order to train it on new data that data specialists have created. Besides, manual intervention is much more straightforward and transparent, if a grammar of the type described above is used than it would be with a statistical algorithm. Our system is flexible in the sense that it can easily be modified in such a way that very specific requirements are met. If, e.g., a future user of our tool (a company that wants to interact with its customers) should want to assign every message that has the word *hotline* in it a certain tag – such as *hotline problem* –, then this requirement can be met by simply adding the line *hotline* to the positive constraints of the classifier called *hotline problem*.

4.2. Applying the DRY principle

Our prototype follows the DRY principle (Don't repeat yourself (Murrell, 2009:35)): Changes are only made in one place. An example: the Korean variants of the mobile phone name with the international ID *google-nexus-s* include *google nexus s*, *google nexus-s*, *nexus s*, *nexus-s*, *구글 넥서스 에스*, *구글 넥서스에스*, *구글 넥서스s*, *구글 넥서스-s*, *넥서스 에스*, *넥서스에스*, *넥서스 s*, *넥서스-s*, *구글 nexus s*, *구글 nexus-s*, *google 넥서스에스*, *구글의 넥서스에스*. This phenomenon is represented in our system as follows: The Korean producer name corresponding to the international ID *google* has the variants *google* and *구글*. The Korean mobile phone name with the international ID *nexus-s* has the variants *nexus-s*, *넥서스에스* and *넥서스-s*. This is the only information our users have to store in order to make the system generate these and many other variants. Our tool generates *google nexus s*, *구글 넥서스 s* and similar variants using the general rule that in any permutation of a product name any minus character may be replaced by a space character. It generates *넥서스에스*, *넥서스 s* and similar variants using the general rule that the producer name may be omitted. And our system generates *구글의 넥서스에스* using the two Korean variants of the producer name and the general rule that phone names can have the form <producer name>의 <product name>. (의 is a genitive affix, i.e. *구글의 넥서스에스* literally means *Google's Nexus S* or *Nexus S by Google*.)

We might, of course, add the general rule to our product name paraphrasing engine that any part of a Korean product name may be spelt either with Latin or with Hangul characters – according to several sets of transliteration conventions that are used in parallel.

Any change in a producer, tariff or product name object, such as the Korean mobile phone name with the international ID *nexus-s*, has implications for the grammars of the message classifiers: Newly generated variants of the product name must be matched by all instances of `\mobile_phone` in all grammars. For efficiency reasons, we compile all product names, tariff names, producer names, message classification grammars, sentiment classification grammars, and so on, into one single function. This function is very efficient, because it doesn't do much more than apply one very large, compiled regular expression. The compiling and reloading of this function is done in the background, so

the users of our tool do not need to know anything about it. They don't even have to understand the word *compile*. They just need to know that the system sometimes needs a few seconds to be able to use changed objects.

5. Conclusion and Outlook

Within this paper, we described a new technical service dealing with the integration of social networking channels into customer interaction management tools. Mining social networks for classification purposes is no novelty; providing an assignment of customer messages to business processes instead of classifying them in topics did not exist before. Above all, our system features effective named entity recognition because of its name paraphrasing mechanism dealing with different types of misspellings in both intra- and interlingual names of tariffs, products, manufacturers and providers. Future research will expand upon this study, investigating other social networking sites and additional companies across a range of non-telecommunication products or services.

6. Acknowledgements

This work was supported by grant no. KF2713701ED0 awarded by the German Federal Ministry of Economics and Technology.

7. References

- Browne, R., Clements, E., Harris, R., Baxter, S. (2009): Business and consumer communication via online social networks: a preliminary investigation. In ANZMAC 2009.
- Gryc, W., Moilanen, K. (2010): Leveraging Textual Sentiment Analysis with Social Network Modeling: Sentiment Analysis of Political Blogs in the 2008 U.S. Presidential Election. In Proceedings of the From Text to Political Positions Workshop (T2PP 2010), Vrije Universiteit, Amsterdam, April 9–10 2010.
- Krauss, J., Nann, S., Simon, D., Fischbach, K., Gloor, P.A. (2008): Predicting Movie Success and Academy Awards Through Sentiment and Social Network Analysis. In ECIS 2008.
- Murrell, P. (2009): Introduction to Data Technologies. Auckland, New Zealand.
- Pal, J.K., Saha, A. (2010): Identifying Themes in Social Media and Detecting Sentiments. Technical Report HPL-2010-50, HP Laboratories.

ATLAS – A Robust Multilingual Platform for the Web

Diman Karagiozov*, **Svetla Koeva****, **Maciej Ogrodniczuk*****, **Cristina Vertan******

* Tetracom Interactive Solutions Ltd., ** Bulgarian Academy of Sciences,

*** Polish Academy of Sciences, **** University of Hamburg,

*Tetracom LTd. Sofia, Bulgaria, **52 Shipchenski prohod, bl. 17 Sofia 1113 Bulgaria,

ul. J.K. Ordonia 2101-237 Warszawa, Poland, *Von-Melle Park 6 20146 Hamburg, Germany

E-mail: diman@tetracom.com, svetla@dcl.bas.bg, maciej.ogrodniczuk@gmail.com,

cristina.vertan@uni-hamburg.de

Abstract

This paper presents a novel multilingual framework integrating linguistic services around a Web-based content management system. The language tools provide semantic foundation for advanced CMS functions such as machine translation, automatic categorization or text summarization. The tools are integrated into processing chains on the basis of UIMA architecture and using uniform annotation model. The CMS is used to prepare two sample online services illustrating the advantages of applying language technology to content administration.

Keywords: content management system, language processing chains, UIMA, language technology

1. Introduction

During the last years, the number of applications which are entirely Web-based, or offer at least some Web front-end has grown dramatically. As a response to the need of managing all this data, a new type of system appeared: the Web-content management system. In this article we will refer to these type of system as WCMS.

Existent WCMS focus on storage of documents in databases and provide mostly full-text search functionality. These types of systems have limited applicability, due to two reasons:

- data available online is often multilingual, and
- documents within a CMS are semantically related (share some common knowledge, or belong to similar topics)

Shortly currently available CMS do not exploit modern techniques from information technology like text mining, semantic Web or machine translation.

The ICT PSP EU project ATLAS¹ – Applied Technology

for Language-Aided CMS aims at filling this gap by providing three innovative Web services within a WCMS. These three Web services: i-Librarian, EUDocLib and i-Publisher are not only thematically different but offer also different levels of intelligent information processing. The ATLAS WCMS makes use of state-of-the art text technology methods in order to extract information and cluster documents according to a given hierarchy. A text summarization module and a machine translation engine are embedded as well as a cross-lingual semantic search engine (Belogay et al., 2011).

The cross-lingual search engine implements Semantic Web technology: the document content is represented as RDF triples and the search index is built up from these triples.

The RDF representation of documents collects not only metadata information about the whole file but also exploits linguistic analysis of the document and store as well the mapping of the file on some ontological concept. This paper presents the architecture of the ATLAS system with particular focus on the language processing components to be embedded aiming to show how robust NLP (natural language processing) tools can be wrapped in a common framework.

¹ The work reported here was carried out within the Applied Technology for Language-Aided CMS project co-funded by the European Commission under the Information and Communications Technologies (ICT) Policy Support Programme (Grant Agreement No 250467). The authors would like to thank all representatives of project partners for their contribution

2. Language resources in the ATLAS System

The linguistic diversity in the project is a challenge not to be neglected: the languages belong to four language families and involve three alphabets. To our knowledge it is the first WCMS which will offer solutions for documents written in languages from Central and South-Eastern Europe.

Whilst the standardised development of tools for widespread languages as English and German is more common, the situation is quite different when involving languages from Central and South Eastern Europe (see <http://www.c-phil.uni-hamburg.de/view/Main/LrecWorkshop2010>).

Tools with different processing depth, different output formats and sometimes very particular approach are current state of the art in the language technology map of the above-mentioned area (Degórski, Marcińczuk & Przepiórkowski, 2008). One of the innovative issues in project ATLAS is the integration of linguistically and technologically heterogeneous language tools within a common framework.

The following description presents the steps taken in order to provide such common representation.

- Starting from the fixed desiderata to include text summarisation, automatic document classification, machine translation and cross-lingual information retrieval the minimal list of tools required by such engines which can be provided by all languages involved in the project has been collated and includes:

- tokeniser,
- sentence boundary detector,
- paragraph boundary detector,
- lemmatizer,
- PoS Tagger,
- NP (noun phrase) chunker,
- NE (named entity) extractor.

Some of these tools are not completely available for particular languages (e.g. NP chunker for Croatian) but can be developed within the project. Regarding the NE extractor the following entities have been agreed upon: persons, dates, time, location and currency.

- The annotation levels in the texts and the minimal features to be annotated have been defined: Paragraph, Sentence, Token, NP and NE. In order to

provide a common representation all linguistic information regarding lemma, PoS etc. have been agreed to be provided at the token level. For a token following features are retained:

- begin – an integer representing the offset of the first character of the token,
 - end – an integer representing the offset of the last character of the token,
 - pos – a string representing the morphosyntactic tag (PoS, gender, number) associated with the token,
 - lemma – a string containing the lemma of the token.
- For each of the above-mentioned tools the list of additional linguistic features to be represented (if necessary and available) have been defined, e.g. *antecedentBegin* and *antecedentEnd* representing the offset of the first and respectively the last character of the referent in an NP. This feature is necessary for processing German NPs and is therefore included as optional in the NP annotation frame.

A glossary of tagsets delivered by each tool is also maintained, ensuring cross-lingual processing.

Each of the language tools can be included as primitive engine, i.e. part of an UIMA aggregate engine, but also as an aggregate engine. In this way any language component can reuse results produced by a particular tool and exploit its full functionality if required.

3. Language Processing chains

One of the goals of the ATLAS WCMS is to offer documented language processing chains (LPCs) for text annotation. A processing chain for a given language includes a number of existing tools, adjusted and/or fine-tuned to ensure their interoperability. In most respects a language processing chain does not require development of new software modules but rather combining existing tools.

Most of the basic linguistic tools (sentence splitters, stopword filters, tokenizers, lemmatizers, part-of-speech taggers) for languages in scope of our interest have already existed as standalone offline applications.

The multilinguality of the system services requires high level of accuracy of each monolingual language chain – simple example is that a word with part-of-speech tag

ambiguity in one language may correspond to an unambiguous word in the other language.

The complexity grows at the level of structure and sense ambiguity differs among languages. Thus the high precision and performance of language specific chains predefines to the great extend the quality of the system as a whole.

For example the Bulgarian PoS tagger has been developed as a modified version of the Brill tagger applying a rule-based approach and techniques for the optimization leading to the 98.3% precision (Koeva, 2007). The large Bulgarian grammar dictionary used for the lemmatization is implemented as acyclic and deterministic finite-state automata to ensure a very fast dictionary look-up.

The language processing chains have been fine-tuned and adjusted to facilitate integration into a common UIMA framework. Other tools (such as noun phrase extractors or named entity recognizers) had to be implemented or multilingually ported.

The annotation produced by the chain along with additional tools (e.g. frequency counters) results in higher-level functions such as detection of keywords and phrases along with improbable phrases from the analyzed content, and utilisation of more sophisticated user functionality deserves complex linguistic functions as multilingual text summarisation and machine translation. UIMA is a pluggable component architecture and software framework designed especially for the analysis of unstructured content and its transformation into structured information. Apart from offering common components (e.g. the type system for document and text annotations) it builds on the concept of analysis engines (in our case, language specific components) taking form of primitive engines which can wrap up NLP (natural language processing) tools adding annotations aggregate engines which define the sequence of execution of chained primitives.

Making the tools chainable requires ensuring their interoperability on various levels. Firstly, compatibility of formats of linguistic information is maintained within the defined scope of required annotation (Ogrodniczuk & Karagiozov, 2011).

The UIMA type system requires development of a uniform representation model which helps to normalize heterogeneous annotations of the component NLP tools.

With ATLAS it covers properties vital for further processing of the annotated data, e.g. lemma, values for attributes such as gender, number and case for tokens necessary to run coreference module to be subsequently used for text summarisation, categorization and machine translation.

To facilitate introduction of further levels of annotation a general markable type has been introduced, carrying subtype and reference to another markable object. This way new annotation concepts can be tested and later included into the core model.

4. Integration of language processing chains in ATLAS

The language chains are used in order to extract relevant information such as named entities and keywords from the documents stored within the ATLAS WCMS. Additionally they provide the baseline for further engines: Text summarization, Clustering and Machine translation (Koehn et al., 2007) and as such they are the foundation of the enhanced ATLAS platform.

The core online service of the ATLAS platform is i-Publisher, a powerful Web-based instrument for creating, running and managing content-driven Web sites. It integrates the language-based technology to improve content navigation e.g. by interlinking documents based on extracted phrases, words and names, providing short summaries and suggested categorization concepts.

Currently two different thematic content-driven Web sites, i-Librarian and EUDocLib, are being built on top of ATLAS platform, using i-Publisher as content management layer. i-Librarian is intended to be a user-oriented web site which allows visitors to maintain a personal workspace for storing, sharing and publishing various types of documents and have them automatically categorized into appropriate subject categories, summarized and annotated with important words, phrases and names.

EUDocLib is planned as a publicly accessible repository of EU legal documents from the EUR-LEX collection with enhanced navigation and multilingual access.

An important aspect of ATLAS System is that all three services operate in a multilingual setting. Similar functionality will be implemented within the project for Bulgarian, Croatian, German, English, German, Greek, Polish and Romanian. The architecture of the system is

modular and allows anytime a new language extension. It is an asynchronous architecture based on queue processing of requests (see Figure 1)

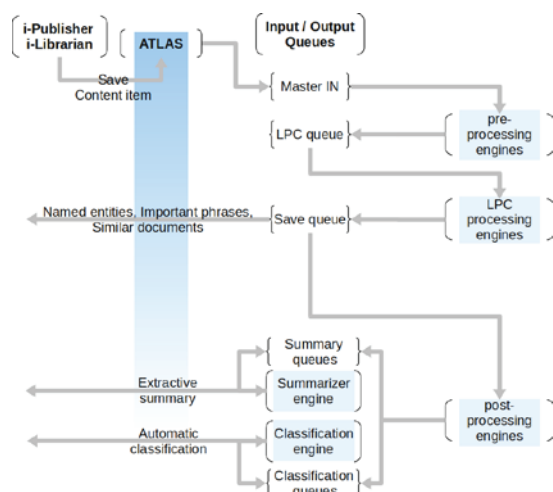


Figure 1: Linguistic processing support in ATLAS System

5. Conclusions

In this paper we present an architecture which opens the door to standardized multilingual online processing of language and it offers localized demonstration tools built on top of the linguistic modules.

The framework is ready for integration of new types of tools and new languages to provide wider online coverage of the needful linguistic services in a standardized manner. New versions of the online services are planned to be launched in the beginning of 2012.

6. References

- Belogay, A., Čavar, D., Cristal, D., Karagiozov, D., Koeva, S., Nikolov, R., Ogrodniczuk, M., Przepiórkowski, A., Raxis P., Vertan C. (to appear): i-Publisher, i-Librarian and EUDocLib – linguistic services for the Web. In: Proceedings of the 8th Practical Applications in Language and Computers (PALC 2011) conference. University of Łódź, Poland, 13-15 April 2011
- Degórski, Ł., Marcińczuk, M., Przepiórkowski, A. (2008): Definition extraction using a sequential combination of baseline grammars and machine learning classifiers. In: Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008. ELRA, Marrakech, http://nlp.ipipan.waw.pl/~adamp/Papers/2008-lrec-lt4el/213_paper.pdf
- Koehn, P., Hoang H., Birch A., Callison-Burch, C., Federico M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar O., Constantin, A., Herbst, E. (2007): Moses: Open Source Toolkit for Statistical Machine Translation. In: ACL (ed.) Annual Meeting of the Association for Computational Linguistics, (ACL), demonstration session. Prague, <http://acl.ldc.upenn.edu/P/P07/P07-2045.pdf>
- Koeva, S. (2007): Multi-word Term Extraction for Bulgarian. In: Piskorski, J., Pouliquen, B., Steinberger, R., Tanev, H. (eds.) Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, pp. 59–66. Association for Computational Linguistics, Prague, Czech Republic, June 2007. <http://www.aclweb.org/anthology/W/W07/W07-1708>
- Ogrodniczuk, M., Karagiozov, D. (to appear): ATLAS – The Multilingual Language Processing Platform. In: Proceedings of the 27th Conference of the Spanish Society for Natural Language Processing. University of Huelva, Spain, 5-7 September 2011

Multilingual Corpora at the Hamburg Centre for Language Corpora

Hanna Hedeland, Timm Lehmberg, Thomas Schmidt, Kai Wörner

Hamburger Zentrum für Sprachkorpora (HZSK)

Max Brauer-Allee 60

D-22765 Hamburg

E-mail: hanna.hedeland@uni-hamburg.de, timm.lehmberg@uni-hamburg.de, thomas.schmidt@uni-hamburg.de,
kai.wörner@uni-hamburg.de

Abstract

We give an overview of the content and the technical background of a number of corpora which were developed in various projects of the Research Centre on Multilingualism (SFB 538) between 1999 and 2011 and which are now made available to the scientific community via the Hamburg Centre for Language Corpora.

Keywords: corpora, spoken language, multilingualism, digital infrastructures

1. Introduction

In this paper, we give an overview of the content and the technical background of a number of corpora which were developed in various projects of the Research Centre on Multilingualism (SFB 538) between 1999 and 2011 and which are now made available to the scientific community via the Hamburg Centre for Language Corpora.

Between 1999 and 2011, the Research Centre on Multilingualism (SFB 538) brought together researchers investigating various aspects of multilingualism focussing either on the language development of multilingual individuals, on communication in multilingual societies, or on diachronic change of languages in multilingual settings. Without exception, the projects of the Centre worked empirically, basing their analyses on corpora of spoken or written language. Over the years, an extensive and diverse data collection was thus built up consisting of language acquisition and attrition corpora, interpreting corpora, parallel (translation) corpora, corpora with a sociolinguistic design and historical corpora.

Since corpus creation, management and analysis were thus crucial to the work of the Research Centre, a project was set up in June 2000 with the aim of designing and implementing methods for the computer-assisted processing of multilingual language data. One major

result of that project is EXMARaLDA, a system for setting up and analysing spoken language corpora (Schmidt & Wörner, 2009, Schmidt et al., this volume). The focus of this paper will be on the spoken language corpora of the Research Centre which were either created or curated with the help of EXMARaLDA.

2. Overview of corpora

As the list of resources in the appendix shows, altogether 31 resources constructed at the SFB 538 were transferred to the inventory of the Hamburg Centre for Language Corpora. 27 of these are spoken language corpora, 3 are corpora of modern written language, and one is a corpus of historical written language. More specifically, we are dealing with the following resource types:

- Language acquisition corpora which document the acquisition of two first languages or a second language. Most of these corpora are longitudinal studies of child language in different bilingual language combinations (German-French, German-Portuguese, German-Spanish, German-Turkish), but other corpus designs (e.g. cross-sectional studies) and other speaker types (e.g. adult learners or monolingual children) are also present.
- Language attrition corpora which document the development of a “weaker” language in adult bilinguals. Three different language combinations

(German-Polish, German-Italian, German-French) are involved.

- Interpreting corpora which document consecutive and simultaneous interpreting involving trained and ad-hoc interpreters for different language combinations (German-Portuguese, German-Turkish, German-Russian, German-Polish, German-Romanian) and in different settings (doctor-patient communication and expert discussion).
- Corpora with a sociolinguistic corpus design whose data are stratified according to biographic characteristics (e.g. age) of the speakers and/or their regional provenance. This comprises a corpus documenting Faroese-Danish bilingualism on the Faroese Islands and a corpus documenting the use of Catalan in different districts of Barcelona.
- Parallel and comparable corpora in which originals and translations of texts are aligned or which consist of original texts from specific genres in different languages.

The entirety of spoken language resources amounts to approximately 5500 transcriptions with approximately 5.5 million transcribed words (not counting secondary annotations).

3. Data model

The spoken language corpora, while sharing the common theme of multilingualism, are still highly heterogeneous with respect to many parameters. As far as their content is concerned, they do not only cover a spectrum of fourteen different languages, but also greatly differ with respect to the recorded discourse types (e.g. interviews, free conversation, expert discussion, classroom discourse, semi-controlled settings, and institutional discourse). Even more variation is to be found with respect to the research interests pursued with the help of the corpora and, consequently, the methodology used to record, transcribe and annotate the data. To begin with, either only audio or both video and audio data are recorded, depending on whether or not non-verbal behavior plays a role for analysis (as is the case, for example, for data of young children). As some projects focused their research on syntactic aspects of language, while others were interested in phonological properties or discourse structures, different systems were applied in

transcribing (e.g. orthographic vs. phonetic transcription or complete vs. selective transcription) and annotating (e.g. prosodic annotations, annotation of code switches) the data.

The challenge in representing the corpora on a common technical basis was thus to find a degree of abstraction which, on the one hand, allows operations common to all resources (such as time alignment of transcription and media) to be carried out efficiently on a unified structure, but, on the other hand, also makes it possible to apply theory or resource specific functions (such as segmentation according to a specific model) to the data. A data model based on annotation graphs (Bird & Liberman, 2001), but supplemented with additional semantic specifications and structural constraints, turned out to be suitable for this task (Schmidt, 2005).

4. Data curation

The construction of a non-negligible part of the resources had been completed or started before EXMARaLDA was available as a working system. A number of legacy software tools (syncWriter, HIAT-DOS, LAPSUS, WordBase) was used for the construction of these corpora resulting in data for which there was hardly a chance of sustainable maintenance. The resources therefore had to be converted to EXMARaLDA in a laborious process described in detail in Schmidt & Bennöhr (2007).

From about 2003 onwards, all projects used EXMARaLDA or other compatible tools (e.g. Praat) for corpus construction. Although these resources were much easier to process once they had been completed, there was still a considerable amount of data curation to be done before they could be published. This involved various completeness and consistency checks on the transcription and annotation data and the construction of valid metadata descriptions for all parts of a resource.

5. Data dissemination

Completed resources are made available to interested users via the WWW¹ through several methods:

- A hypermedia representation of transcriptions, annotations, recording and metadata allows users to browse corpora online (see figure 1).

¹ <http://www.corpora.uni-hamburg.de>

The screenshot shows the HAMATAC interface. At the top, it says 'HAMATAC - MT_091209_David [Prev] [Next]'. Below that is a play button and a 00:00 timer. On the left, there's a 'Tier display' section with checkboxes for 'disfluency', 'nn', 'pho', and 'v'. Below that is a 'Files' section listing various transcription and annotation formats like RTF Partitur, PDF Partitur, POS tags, EXMARaLDA Basic Transcription, EXMARaLDA Segmented Transcription, ELAN Annotation File, TEI transcription, PRAAT TextGrid, FOLKER transcription, CHAT transcription, and AG XML. The main area shows five numbered transcription segments (1-5) with speaker labels (Dav [v], Dav [disfluency], Ruf [v]) and time-coded annotations in parentheses, such as ((0,4s)) hallo ((lacht)) ((1,0s)) ich wollte Ihnen ganz gerne den Weg erklären ((0,8s)) ähm ((1,1s)) ich befir ((0,2s)) ja.

Figure 1: Hypermedia representation of a transcription from the Hamburg Map Task Corpus (HAMATAC)

- Resources can be downloaded in the EXMARaLDA format and then edited and queried with the system's tools (Partitur-Editor for editing transcriptions, Coma for editing and querying metadata, EXAKT for querying transcription and annotation data).
- Queries via EXAKT can also be carried out on remote data, i.e. without downloading the resource first, or through a web interface, i.e. without the need to install local software first.
- A number of export formats are offered for each annotation file making it possible to edit or query the data also with non-EXMARaLDA tools. Most importantly, most data are also available in the CHAT format of the CHILDES system, as ELAN annotation files, as Praat TextGrids and as TEI files.

Access to all corpora is password protected. The process for obtaining a password varies from resource to resource, but always requires the data owner's consent. Due to privacy protection issues, a part of the spoken resources can only be made accessible in the form of transcriptions, not audio or video recordings.

6. Future plans

In order to cater for the long term archiving and availability of the data beyond the finite funding period of the Research Centre, in January 2011 the Hamburg Centre for Language Corpora (HZSK, <http://www.corpora.uni-hamburg.de>) was set up. This institution is intended to provide a permanent basis not

only for the corpora and tools referred to in this paper, but also for further resources existing or under construction at the University of Hamburg. The HZSK is part of the CLARIN-D network and will, in the years to come, integrate its resources into this infrastructure by providing protocols for metadata harvesting, assigning PIDs to resources, allowing for single-sign-on mechanisms and implementing interfaces as defined by CLARIN for access to metadata and annotations.

7. References

- Bird, S., Liberman, M. (2001): A formal framework for linguistic annotation. In: *Speech Communication* (33), pp. 23-60.
- Schmidt, T. (2005): *Computergestützte Transkription - Modellierung und Visualisierung gesprochener Sprache mit texttechnologischen Mitteln*. Frankfurt a. M.: Peter Lang.
- Schmidt, T., Bennöhr, J. (2008): Rescuing Legacy Data. In: *Language Documentation and Conservation* (2), pp. 109-129.
- Schmidt, T., Wörner, K. (2009): EXMARaLDA – Creating, analysing and sharing spoken language corpora for pragmatic research. In: *Pragmatics* 19(4), pp. 565-582.

Appendix: List of resources

Corpus name Project / Data Owner Type	Short description	Language(s)	Size
Spoken resources			
HABLA (Hamburg Adult Bilingual Language) E11 / Tanja Kupisch spoken/audio/exmaralda	Audio recordings of semi-spontaneous interviews (elicited grammaticality judgments and production data are collected from the same speakers)	deu, fra, ita	169 communications 127 speakers 737797 transcribed words 169 transcriptions
DUFDE (Deutscher und Französischer doppelter Erstspracherwerb) E2 / Jürgen Meisel spoken/video/exmaralda	Video recordings (longitudinal study) of seven French-German bilingual children aged between 1 year;6 months and 6 years;11 months (+some later recordings).	deu, fra	562 communications 14 speakers ca. 1000000 transcribed words 849 transcriptions
BIPODE (Bilingualer Portugiesisch-Deutscher Erstspracherwerb) E2 / Jürgen Meisel spoken/video/exmaralda	Video recordings (longitudinal study) of three Portuguese-German bilingual children aged between 1 year;6 months and 5 years;6 months.	deu, por	250 communications 48 speakers ca. 250000 transcribed words 227 transcriptions
CHILD-L2 E2 / Jürgen Meisel spoken/video/exmaralda	Video recordings of children which start acquiring French or German as a second language at the age of three or four years.	deu, fra	181 communications 69 speakers 376114 transcribed words 181 transcriptions
ZISA (Zweitspracherwerb Italienischer und Spanischer Arbeiter) E2 / Jürgen Meisel spoken/audio/exmaralda	Recordings of adult L2-German-learners	deu	101 communications 5 speakers 119667 transcribed words 100 transcriptions
BUSDE (Baskischer und Spanischer doppelter Erstspracherwerb) E2 / Jürgen Meisel spoken/video/other	Longitudinal language acquisition study on bilingual Basque-Spanish children	eus, spa	<i>unknown</i>
PAIDUS (Parameterfixierung im Deutschen und Spanischen) E3 / Conxita Lleó spoken/audio/exmaralda	Audio recordings of monolingual children.	deu, spa	253 communications 66 speakers 166976 transcribed words 253 transcriptions
PHONBLA Longitudinalstudie Hamburg E3 / Conxita Lleó spoken/audio+video/exmaralda	Longitudinal data of Spanish/German bilingual children	deu, spa	413 communications 61 speakers 303792 transcribed words 413 transcriptions
PHONBLA Querschnittsstudie Madrid E3 / Conxita Lleó spoken/audio+video/exmaralda	Cross sectional study of bilingual German-Spanish L1 acquisition	deu, spa	113 communications 34 speakers 56722 transcribed words 113 transcriptions
PEDES (Phonologie-Erwerb Deutsch-Spanisch als Erste Sprachen) E3 / Conxita Lleó spoken/audio/exmaralda	Longitudinal data of Spanish/German bilingual children	deu, spa	127 communications 21 speakers 101292 transcribed words 127 transcriptions
PHON-CL2 E3 / Conxita Lleó spoken/audio/exmaralda	Recordings of German subjects/children who have learned (or are learning) Spanish after the age of two	deu, spa	26 communications 22 speakers 17412 transcribed words 26 transcriptions
PHONMAS E3 / Conxita Lleó spoken/audio/exmaralda	Recordings of monolingual Spanish children (as comparable data for Madrid-PhonBLA)	spa	49 communications 4 speakers 3067 transcribed words 49 transcriptions
TÜ_DE-CL2-Korpus E4 / Monika Rothweiler spoken/video/exmaralda	Video recordings of (spontaneous and elicited language) of eight bilingual children with Turkish as their first language	deu	112 communications 19 speakers 348292 transcribed words 112 transcriptions
TÜ_DE-L1-Korpus E4 / Monika Rothweiler spoken/audio/exmaralda	Video recordings of (spontaneous and elicited language) of twelve bilingual children with Turkish as their first language	tur	12 communications 22 speakers 13 transcriptions

Rehbein-ENDFAS/Rehbein-SKOBI-Korpus E5 / Jochen Rehbein spoken/audio/exmaralda	Audio recordings of evocative field experiments with Turkish and German monolingual and Turkish/German bilingual children.	deu, tur	1017 communications 523 speakers 289012 transcribed words 836 transcriptions
ENDFAS/SKOBI Gold Standard E5 / Jochen Rehbein spoken/audio/exmaralda	Audio recordings of Turkish and German monolingual and Turkish/German bilingual children. Demo Excerpt from the larger Rehbein-ENDFAS/Rehbein-SKOBI-Korpus	deu, tur	3 communications 8 speakers 4862 transcribed words 3 transcriptions
Catalan in a bilingual context H6 / Conxita Lleó spoken/audio/exmaralda	Prompted, read and spontaneous speech data of Catalan speakers from Barcelona, stratified according to district and age of speakers	cat	225 communications 234 speakers 187967 transcribed words 875 transcriptions
Hamburg Corpus of Polish in Germany H8 / Bernhard Brehmer spoken/audio/exmaralda	Audio recordings of bilingual (Polish and German) and monolingual (Polish) adults (16-46 years). Recordings of semi-spontaneous data (3 topics) and renarration of a picture story (from 'Vater und Sohn')	pol	354 communications 94 speakers ca. 350000 transcribed words 358 transcriptions
Hamburg Corpus of Argentinean Spanish (HaCASpa) H9 / Christoph Gabriel spoken/audio/exmaralda	Recordings of spontaneous speech and laboratory data of speakers of Porteño Spanish in Argentina (read speech, story retelling, read question-answer pairs, intonation questionnaires, free interviews); 7 experiments altogether.	spa	259 communications 63 speakers 141321 transcribed words 261 transcriptions
Dolmetschen im Krankenhaus K2 / Kristin Bührig Bernd Meyer spoken/audio/exmaralda	Monolingual and interpreted doctor-patient communication in hospitals	deu, por, tur	91 communications 189 speakers 165689 transcribed words 92 transcriptions
SkandSemiko (Skandinavische Semikommunikation) K5 / Kurt Braunmüller spoken/audio/exmaralda	Radio recordings, recordings of group discussions and classroom discourse with speakers of two or more Scandinavian languages (Swedish, Danish, Norwegian) interacting.	dan, nor, swe	162 communications 515 speakers 269945 transcribed words 74 transcriptions
CoSi (Consecutive and Simultaneous Interpreting) K6 / Bernd Meyer spoken/audio+video/exmaralda	Recordings of simultaneously and consecutively interpreted lectures	deu, por	3 communications 8 speakers 35432 transcribed words 5 transcriptions
FADAC Hamburg (Faroese Danish Corpus Hamburg) K8 / Kurt Braunmüller spoken/audio/exmaralda	Recordings of semi-structured interviews in Faroese and Danish with bilingual speakers living on the Faroe Islands.	dan, fao	92 communications 82 speakers 440194 transcribed words 92 transcriptions
ALCEBLA T4 / Conxita Lleó spoken/audio/exmaralda	Recordings of Spanish-German bilingual children living in Germany and attending the Spanish complementary school at the first level	deu, spa	66 communications 23 speakers 36717 transcribed words 66 transcriptions
Simuliertes Dolmetschen im Krankenhaus T5 / Kristin Bührig, Bernd Meyer spoken/audio+video/exmaralda	Simulations of interpreted doctor-patient communication.	deu, pol, ron, rus	4 communications 12 speakers 4018 transcribed words 4 transcriptions
EXMARaLDA Demo Corpus Z2 / Hamburger Zentrum für Sprachkorpora spoken/audio+video/exmaralda	A selection of short audio and video recordings in different languages for demonstration of the EXMARaLDA system	deu, eng, fra, ita, nor, pol, spa, swe, tur, vie	19 communications 50 speakers 11659 transcribed words 19 transcriptions
Hamburg Map Task Corpus Z2 / Hamburger Zentrum für Sprachkorpora spoken/audio/exmaralda	Audio recordings of map tasks with advanced learners of German	deu	24 communications 26 speakers 24409 transcribed words 24 transcriptions

Written resources			
HaCOSSA (Hamburg Corpus of Old Swedish with Syntactic Annotations) H3 / Kurt Braunmüller written/tei	Bible translations, religious and secular prose, law texts, non-fiction literature (geographical, theological, historic, natural science), diploma.	dan, deu, isl, lat, nob, swe	35 texts
Covert translation: popular science K4 / Juliane House written/tei	Translation corpora of original texts with translations and comparable texts from the genre popular scientific prose	deu, eng	114 texts 500446 words
Covert Translation: business communication (old) K4 / Juliane House written/tei	Translation corpora of original texts with translations and comparable texts from the genre external business communication	deu, eng	119 texts 169154 words
Covert Translation: business communication (new) K4 / Juliane House written/tei	Translation corpora of original texts with translations and comparable texts from the genre external business communication	deu, eng	198 texts

The English Passive and the German Learner – Compiling an Annotated Learner Corpus to Investigate the Importance of Educational Settings

Verena Möller, Ulrich Heid

Universität Hildesheim

Institut für Informationswissenschaft und Sprachtechnologie

- Sprachtechnologie / Computerlinguistik -

Marienburger Platz 22

31141 Hildesheim

E-mail: verena.moeller@uni-hildesheim.de, ulrich.heid@uni-hildesheim.de

Abstract

In the south of Germany, a number of changes have recently been effected with respect to the possible environments in which pupils in primary and secondary schools learn/acquire English. The current co-existence of various educational settings allows for investigation of the effects that each of these settings has on the structure of learners' interlanguage. As different text types are used as input in the various educational environments which have been created in secondary schools, the English passive has been chosen as a diagnostic criterion for the analysis of the learners' production of written text. The present article describes the compilation of a corpus of teaching materials and a learner corpus. It outlines the procedures involved in annotating metadata, esp. those obtained from questionnaires and psychological tests. Tools for linguistic annotation (POS-taggers and a parser) are compared with respect to their effectiveness in dealing with data from students after 6-10 years of instruction and/or immersion.

Keywords: second language acquisition, learner corpus, metadata, POS-tagging, parsing

1. Co-Existence of Educational Settings

In recent years, a number of changes in the educational system in Baden-Württemberg (Germany) have been effected, some of them directly related to language learning and acquisition. In addition to English as a Foreign Language (EFL) lessons in secondary schools, more and more CLIL (content and language integrated learning) programmes have been established. CLIL learners are taught History and Biology, as well as a combination of Geography, Economics and Politics in English during certain years specified by the curriculum. In addition, 'immersive-reflective' lessons (IRL) have been introduced at the primary level. These focus on situational context and communication, while at the same time allowing for reflection on language whenever this is deemed necessary.

Due to the current co-existence of various educational settings, it is timely to compile a learner corpus in order

to investigate the effects of educational settings on the interlanguages of the following four groups of learners:

- 1) participants in EFL, but neither IRL nor CLIL;
- 2) participants in EFL and IRL, but not CLIL;
- 3) participants in EFL and CLIL, but not IRL;
- 4) participants in EFL, CLIL and IRL.

All learners participating in the study described below are in Year 11, i. e. they have entered the final stage of their school career.

2. The Passive and the German Learner

To test the impact of educational settings on the learner groups outlined above, grammatical structures need to be analysed with respect to the question which ones will most likely occur with different frequency in the types of input that is available to these learners. For the purpose of the present study, the English passive has been chosen as an indicator.

Being exposed to scientifically-oriented writing, CLIL learners receive input from a genre that differs from those

used in EFL classes. Based on the findings of Svartvik (1966), this genre may be assumed to contain a relatively larger number of passive structures. This will be tested on a corpus of teaching materials. It is likely that passive constructions will also occur with higher frequency in the written output of CLIL learners.

Different types of *be Ved* constructions, i. e. central passives with solely verbal features and semi-passives carrying verbal as well as adjectival characteristics, are included into an analysis of teaching materials and of written learner language. Questions of verb valency are also taken into account.

3. The Teaching Materials Corpus

3.1. Input and Norm: TMCinp and TMCref

To determine whether or not the various groups of learners are indeed exposed to different types of written input, a corpus of teaching materials (TMC) is being compiled. It includes written material for learners from Year 7 onwards, as both CLIL and the treatment of the English passive start in that year.

The TMC serves two purposes: On the one hand, it compares input from EFL lessons to input from CLIL by means of an input subcorpus (TMCinp). An analysis of Year 7-10 materials for both groups will establish whether or not passive structures do indeed occur with higher frequency in CLIL materials than in EFL materials.

Secondly, the TMC represents a reference norm. All four groups of learners take the same EFL exams at the end of their school career. Hence a target norm, against which the learners' written performance at that stage can be measured, is defined by compiling a reference subcorpus (TMCref). TMCref comprises Year 11-12 materials designed for use in the EFL classroom.

The overall structure of the TMC is presented in Fig. 1.

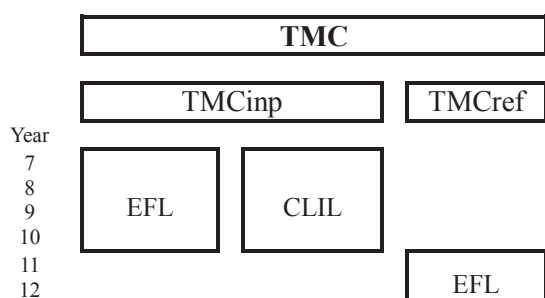


Figure 1: Teaching Materials Corpus (TMC)

3.2. Metadata

The TMC is, amongst others, annotated with the following metadata to enable efficient querying:

- learning environment (EFL vs. CLIL);
- publisher and title;
- targeted age group;
- type of material (textbook, workbook, newspaper, fiction not included into textbooks etc.);
- genre.

The TMC includes written text as well as supplementary information and instructions referring to written text only, rather than to film sequences or listening comprehension exercises that may accompany textbooks. Skills files, which are used to acquire the techniques and vocabulary needed for various types of text production, are also excluded from the TMC.

3.3. POS-Tagging and Parsing

To linguistically annotate the TMC, the English versions of TreeTagger (Schmid, 1994) and of the MATE parser (Bohnet, 2010) were used. TreeTagger is a stochastic part-of-speech (POS) tagger that uses annotated reference texts, lexical entries (word form, lemma, POS), word endings and three-word windows (two items left of the candidate) as an input. It performs lemmatization together with tagging (U Penn Treebank tagset, 36 tags). MATE is a trainable dependency parser (trained on the U Penn Treebank). Both tools also perform sentence tokenization. Having the TMC tagged, lemmatized and parsed, we expect to be able to extract occurrences of passives with good precision and recall.

4. Learner Corpus: Data Elicitation

4.1. Personal Data

If a difference in the use of passive constructions in learner text is to be attributed to a specific educational setting, it is inevitable to make sure that all groups of learners are comparable with respect to a number of individual parameters. The collection of these personal data centres around two methods – a questionnaire and psychological testing. In the questionnaire, learners are asked to provide information e. g. on age, sex, mother tongue, learning environment, etc. (cf. sec. 5.1.). Moreover, information on cognitive capacities and motivation needs to be gathered by means of

psychological testing. Participation in CLIL lessons is not compulsory and there is room for the possibility that learners opt for these programmes because they possess better overall or language-related cognitive skills, or a higher level of motivation.

The intelligence test used in this study (PSB-R 6-13, Horn, 2003) provides information on the two cognitive factors mentioned above, along with individual scales on lexical fluency in German and language-related logical thinking. Data from a pilot study with 28 subjects (cf. Table 1) suggest that the most reasonable procedure will be to sort participants into two groups according to the scores attained on the scales for overall and language-related cognitive capacities (SW 100-109/IQ 100-114 vs. SW 110-119/IQ 115-129).

SW	General (PSB-R 6-13 GL)	Lang.-related (PSB-R 6-13 V)
100-109	14	19
110-119	10	9
<100 or >119	4	0

Table 1: Pilot study – cognitive skills

The psychological test related to motivational factors (FLM 7-13, Petermann & Winkel, 2007) provides, amongst others, information on orientation towards performance and success as well as perseverance and effort. The study aims at learners with an average motivation (T-score 40-60), allowing for a margin on both sides (T-score 36-64). The results of the pilot study show that 23/24 out of 28 learners fall into this category for the two scales.

4.2. Learner Text Data

Learners are invited to write two short argumentative essays within a time frame of about 70 minutes. Students at this level are used to this kind of task, as it is widely practised throughout the years preceding their final exams. Learners key in their texts using a simple editor without a spellchecker. However, they are allowed to use a printed version of a monolingual dictionary.

Some of the essay topics to choose from involve passive constructions, others do not. The following enumeration lists the topics most frequently chosen:

- 1) In order to fight teenage drinking, the legal drinking age should be raised to 21. (18 essays)

- 2) In Germany, the education system offers equality of opportunity to everyone, rich or poor. (9 essays)
- 3) Privacy is a thing of the past. (9 essays)
- 4) The death penalty should be reintroduced in Germany. (9 essays)

In the pilot study, the average number of words produced in one essay was 308, resulting in a corpus of slightly more than 17,000 words.

4.3. Experimental Data

A study on the *International Corpus of Learner English (ICLE)* has revealed a marked underuse of the English passive even in more advanced German learners (cf. Granger, 2009). It can therefore be assumed that this will be the case with less advanced learners as well. Thus, to make sure that additional information is available as a backup, text data elicitation is supplemented with an experimental task to find out whether or not learners are able to transform active sentences into their passive counterparts. Not only are learners tested on the morphology of the English passive in various tenses (cf. sentences 1 and 2), but the task also involves ditransitive verbs to find out which object is most likely to be moved to the subject position of the passive sentence (cf. sentence 3). Moreover, learners are presented with constructions that have not or only marginally been part of their EFL instruction (e.g. prepositional verbs or complex-transitive verbs, cf. sentences 4 and 5).

- 1) *My sister's friends often invite me to parties.*
- 2) *The teams will play the last match of the season next Friday.*
- 3) *My grandparents have promised me a new computer.*
- 4) *People look upon the construction of the railroad as a fantastic achievement.*
- 5) *Everyone considered Pat a nice person.*

In the experimental task, learners respond to 12 sentences in about 20 minutes. In addition, they are asked to rate the reliability of their own responses on a 5-point Likert scale. These reliability scores are included into the learner corpus as metadata.

5. Learner Corpus: Annotation

5.1. Metadata

As a result of the procedures described in sec. 4, the

learner corpus comprises information on the following aspects, annotated as metadata:

- age and sex;
- mother tongue and languages spoken at home;
- other second and foreign languages, duration of acquisition and self-rated competence;
- duration of the learner's longest stay in an English-speaking country;
- number of school years skipped or doubled;
- attendance of German primary school and participation in immersive-reflective lessons;
- textbooks used in the EFL classroom;
- participation in CLIL programmes and school subjects affected;
- exposure to English during the learner's spare time;
- aspects of cognitive capacities;
- aspects of motivation;
- self-rated reliability of responses in the experimental task;
- essay topic.

5.2. POS-Tagging

The Learner Corpus was POS-tagged by means of TreeTagger, the same way as TMC. In addition, the CLAWS4 tagger was applied, a hybrid tagger that involves both probabilistic and rule-based procedures (Garside & Smith, 1997). For the purpose of the present pilot study, we have used the C7 tagset, which amounts to a number of 146 tags. CLAWS4 provides probability scores for tags assigned to potentially ambiguous word forms. For the 17,000 word pilot learner corpus, CLAWS4 lists 5,255 ambiguities; of these, 88.4 % are assigned a first tag alternative with 80 % probability or more.

TreeTagger assigned an <UNKNOWN> tag to 423 words that were misspelled. Slightly more than half of these nevertheless received a correct POS-tag. When CLAWS4 was used, only two items received the unknown tag, <FU>. These were misspellings identified as truncations. However, 51 misspelled words received an <ERROR> tag in addition to their POS-tag. 16 of these were correctly POS-tagged despite their spelling error. It is remarkable that 19 of the 35 mistagged words involved proper nouns or adjectives denoting nationalities, spelt without a capital letter. In seven cases, the omission of apostrophes to mark either a genitive or a clitic made it

impossible to assign a correct POS-tag. As CLAWS4 operates using the probability of POS-tags for both individual words and tag sequences, this had rather far-reaching consequences for the tagging of the preceding and following units.

5.3. Parsing

As is the case for the TMC, the Learner Corpus was also parsed by means of MATE. As the parser assigns POS-tags to the word forms analysed, a comparison with TreeTagger and CLAWS4 was performed (cf. sec. 6.2. for details). Tested on the misspelled words tagged <UNKNOWN> by TreeTagger, MATE performed slightly better on the assignment of correct POS-tags (245 vs. 219). MATE and CLAWS4 were almost equally successful on partly erroneous occurrences of *be Ved* (cf. Table 3). To retrieve English passive constructions from the Learner Corpus, in principle no parsing would be needed. Correct syntagms can be found by means of patterns formulated in terms of POS and lemmas; most erroneous occurrences are not classifiable for the parser and thus need to be searched with partial patterns (e.g. participle alone).

6. Retrieval of Passive Constructions

6.1. Manual Analysis

Before an automatic analysis was undertaken, instances of English *be Ved* constructions were retrieved manually from the pilot corpus. 151 occurrences were found, 22 of which being erroneous. The following types of error occurred:

- Omission of *be* (6 instances): **Should the death penalty reintroduced in Germany?*
- Morphological and/or orthographic errors in the form of *be* or related clitics (3 instances): **You arent forced to post anything in the internet.*
- Morphological and/or orthographic errors in the past participle (11 instances): **[...] if the alcohol can just be buyed by 21 old people.*
- Lexical errors (1 instance): **[...] so he is already prisoned by the police.*
- A combination of these (1 instance): **[...] because it's forbideden.¹*

¹ The fact that learners frequently use accents on the keyboard instead of apostrophes presents POS-taggers with problems. However, this will be solved by combining automatic

In addition, 9 instances of *get*-passives were retrieved, three of which were ungrammatical.

6.2. Automatic Analysis

An analysis of which POS-tags TreeTagger (TT), CLAWS4 (CL) and the tagger integrated into the MATE parser (MA) assign to the learners' grammatical *be Ved* and *get Ved* constructions has shown that only TreeTagger was able to find all instances² (cf. Table 2).

	TT	CL	MA
<i>be</i> + past participle (n=129)	129	128	123
<i>get</i> + past participle (n=6)	6	4	5

Table 2: Retrieval of *be Ved* and *get Ved*

An analysis of how the three taggers deal with erroneous occurrences of *be Ved* constructions has revealed that both CLAWS4 and MATE seem to have less difficulty in dealing with ungrammatical past participles than TreeTagger (cf. Table 3).

	TT	CL	MA
correct tag for <i>be</i> (n=16)	12	12	11
correct tag for the past participle (n=22)	11	15 ³	15
corrects tags for <i>be</i> and past participle (n=16)	4	8	8

Table 3: Tags in erroneous occurrences of *be Ved*

7. Conclusion

In this paper, work towards a richly annotated corpus of teaching materials (TMC) and of learner text was described. The corpora are particularly rich in metadata (both on the sources of TMC and on learner parameters), and they have been processed with two POS-taggers (TreeTagger and CLAWS4) and a dependency parser

annotation with manual editing (cf. Granger 1997).

² MATE had some difficulty processing *said* as a participle in passive constructions (4 instances).

³ It is interesting to note that in some cases in which learners overgeneralize the -ed suffix for the formation of past participles (e. g. **bued*, **payed*, **splitted*), CLAWS4 will add <@> to the POS-tag of the respective form, indicating that occurrence is deemed unlikely.

(MATE). Metadata and linguistic annotations can be queried together.

As of summer 2011, the corpora are still very small (TMC: 420,000 words, LC: 17,000 words); they will gradually be enlarged. Both TreeTagger and CLAWS4 will continue to be used concurrently, as TreeTagger seems to perform better on correct forms, and CLAWS4 to be more robust towards erroneous ones. All relevant passive constructions will be extracted from the enlarged corpora, with pattern-based search for the correct forms and semi-automatic procedures for erroneous ones. The retrieved data, together with the pertaining metadata, should allow for an interpretation in terms of the impact of educational settings on the interlanguage of learners.

8. Acknowledgements

The authors would like to thank following companies: Alfred Kärcher Vertriebs-GmbH, Cornelsen Verlag GmbH, Ernst Klett Verlag GmbH, SWN Kreissparkasse, Pearson Assessment & Information GmbH.

9. References

- Bohnet, B. (2010): Very High Accuracy and Fast Dependency Parsing is not a Contradiction. In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), Beijing, pp. 89–97.
- Garside, R., Smith, N. (1997): A hybrid grammatical tagger: CLAWS4. In R. Garside, G. Leech & A. McEnery (Eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman, pp. 102-121.
- Granger, S. (2009): More lexis, less grammar? What does the (learner) corpus say? Paper presented at the Grammar & Corpora conference, Mannheim, pp. 22-24 September 2009.
- Granger, S. (1997): Automated Retrieval of Passives from Native and Learner Corpora. Precision and Recall. In *Journal of English Linguistics* 25(4), pp. 365-374.
- Horn, W. (2003): PSB-R 6-13. Prüfungssystem für Schul- und Bildungsberatung für 6. bis 13. Klassen – revidierte Fassung. Göttingen: Hogrefe.
- Petermann, F. & Winkel, S. (2007): FLM 7-13. Fragebogen zur Leistungsmotivation für Schüler der 7. bis 13. Klasse. Frankfurt/Main: Harcourt.

- Schmid, H. (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. In Proceedings of International Conference on New Methods in Language Processing.
- Svartvik, J. (1966): On Voice in the English Verb. The Hague/Paris: Mouton.

Register, Genre, Rhetorical Functions: Variation in English Native-Speaker and Learner Writing

Ekaterina Zaytseva

Johannes Gutenberg-Universität Mainz, Department of English and Linguistics

Jakob-Welder-Weg 18, 55099 Mainz

E-mail: zaytseve@uni-mainz.de

Abstract

The present paper explores patterns and determinants of variation found in the writing of two groups of novice academic writers: advanced learners of English and English native speakers. It focuses on lexico-grammatical means for expressing the rhetorical function of contrast in academic and argumentative writing. The study's aim is to explore and to compare stocks of meaningful ways of expressing the rhetorical function of contrast employed by native and learner novice academic writers in two different written genres: argumentative essays and research papers. The following corpora are used for that purpose: the *Louvain Corpus of Native English Essays* (LOCNESS), the *Michigan Corpus of Upper-level Student Papers* (MICUSP), the *British Academic Written English* corpus (BAWE) and two corpora of learner English, i.e. the *International Corpus of Learner English* (ICLE) and the *Corpus of Academic Learner English* (CALE) – the latter being a corpus of advanced learner academic writing, currently being compiled at Johannes Gutenberg-Universität Mainz, Germany. The study adopts a variationist perspective and a functional-pedagogical perspective on learner writing, aiming at contributing to the field of second language acquisition (SLA), by focusing on advanced stages of acquisition and teaching English for academic purposes.

Keywords: novice academic writing, rhetorical function of contrast, variation, function-oriented annotation

1. Introduction

The branch of the SLA focusing on advanced levels of proficiency puts forward issues that are problematic for researchers, EAP teachers, and foreign language learners alike. Those include the need for an exhaustive description of language performance on an advanced level and a set of defining characteristics which could be further developed into assessment criteria.

One of the factors responsible for the problematic nature of “advancedness” is a somewhat narrow view of this stage of language acquisition as on the one hand, “no more than ‘better than intermediate level’ structural and lexical ability for use”, as pointed out by Ortega and Byrnes (2008:283); and yet, on the other hand, as language performance, not “flawless” enough to be considered native-like.

2. Theoretical Background

Advanced learner writing has recently been the object of a number of corpus-based studies (cf. e.g. Callies, 2008;

Gilquin & Paquot, 2008; Paquot, 2010). It has generally been analysed from a pedagogical perspective, i.e. against the yardstick of English native-speakers' writing, where features of learner writing have often been characterized as non-native-like. Among the areas identified as problematic for advanced learners are most notably accurate and appropriate use of lexis, register awareness, and information structure management. Yet, studies adopting a variationist perspective on advanced learners' output and considering a possible influence of different kinds of variables are still scarce (cf., however, Ädel, 2008; Paquot, 2010; Wulff & Römer, 2009). One of the reasons for this could be the lack of corpora representing advanced academic learner writing (Granger & Paquot, forthcoming), which makes it difficult, for example, to analyse the importance of genre and writer's genre (un)awareness as possible determinants of variation. The existing corpora include the following projects in progress: the ‘*Varieties of English for Specific Purposes*’ database (VESPA) (cf. Granger, 2009), the

Corpus of Academic Learner English (CALE)¹, and the *Cologne-Hanover Advanced Learner Corpus* (CHALC) (Römer, 2007).

The pedagogical approach to learners' language production has brought forward particular kinds and methods of learner data analysis. One of them is annotating a learner corpus for errors (cf. Granger, 2004). Valuable as it is, this kind of corpus annotation, however, does not allow for a truly usage-based perspective on learner language production, where learners' experience with language in particular social settings is the focus of attention.

Corpus-based analyses of native English academic writing, meanwhile, have revealed that this register is characterised by a specific kind of vocabulary on the one hand (Biber et al., 1999; Coxhead, 2000; Paquot, 2010) and by certain kinds of grammatical structures on the other hand (e.g. Biber, 2006; Kertz & Haas, 2009). In addition, it has been pointed out that the register of native English academic writing displays a certain degree of variation as well, e.g. there is discipline- and genre-based variation in the form and use of lexico-grammatical structures used in written discourse (Hyland, 2008). However, there is little information on possible variation in different genres produced by novice native English academic writers (cf., however, Wulff & Römer, 2009).

3. Project Aims and Objectives

The present paper reports on work in progress exploring patterns and determinants of variation found in the writing of two groups of novice academic writers: advanced learners of English and English native speakers. It focuses on lexico-grammatical ways for expressing the rhetorical function of contrast in academic and argumentative writing. The study's aim is to explore and subsequently to compare stocks of meaningful ways of expressing contrast employed by native and learner novice academic writers in two different written genres: argumentative essays and research papers. For that purpose the following corpora are used: three corpora of native English corpora: the *Louvain Corpus of Native English Essays* (LOCNESS) (Granger, 1996), the *Michigan Corpus of Upper-level Student Papers*

(MICUSP)², the *British Academic Written English* corpus (BAWE) (Nesi, 2008) as well as two corpora of learner English, i.e. the *International Corpus of Learner English* (ICLE) (Granger, 2003) and the *Corpus of Academic Learner English* (CALE)³ - a corpus of advanced learner academic writing, currently being compiled at Johannes-Gutenberg-Universität Mainz, Germany.

Another aim of the study is to investigate to what extent the influence of the variable 'genre' is a possible determinant of variation in the written production of various groups of academic writers. In this respect, it is important to address the issue of novice writers' genre awareness and to discuss the question of native-speaker norm. In addition, the paper explores the existence of interlanguage (IL)-specific strategies used by advanced learners to express rhetorical functions in writing.

The latter will be achieved by annotating both corpora of advanced learner writing for the rhetorical function of contrast. This kind of function-oriented annotation, though still rare in English learner corpus research, presents researchers with a valuable opportunity to view learners as active language users, rather than learners demonstrating deficient knowledge of the target language. In addition, the potential of multidimensional corpus analysis (Biber & Conrad, 2001) is currently being considered as a highly useful method of distinguishing between different registers and genres.

The study, thus, adopts a variationist perspective to novice academic writing, considering advanced interlanguage as a variety in its own right. At the same time, a functional-pedagogical perspective allows for a further analysis of those areas of language use that are still problematic for advanced learners, and reveals meaningful ways in which learners cope with writing-related tasks.

4. Function-oriented annotation

The advantage of adding a function-driven annotation is that it makes it possible to generally identify contrast in learner writing and to pin down an extensive stock of language means, treated as writers' lexico-grammatical preferences for signaling this rhetorical function in written discourse.

¹ <http://www.advanced-learner-varieties.info>

² <http://micusp.elicorpora.info/www.micusp.org>

³ <http://www.advanced-learner-varieties.info>

Further on, the encoded information allows for function-driven, together with form-driven searches in learner writing, resulting in a comprehensive and accurate picture of the variety of lexico-grammatical means for expressing contrast used by two groups of (advanced) German learners in their writing. In addition, a subsequent quantitative analysis can provide valuable insights into general and individual preferences of learners in terms of which items are particularly favoured in the context of a specific writing-related task set in a specific situation of language use. Moreover, its combination with a qualitative analysis of patterns and determinants of variation in the ways of expressing contrast in writing promises to shed more light on general written argumentation strategies employed by (advanced) German learners.

In order for this kind of annotation to be reliable, several conditions have to be met, which when applied to the present project, imply clarification of the concept of a rhetorical function and a clear definition of the rhetorical function of contrast in terms of its aim and distinctive characteristics, complemented by a list of possible language items for its realization in writing.

The next step involves annotating each instance of contrast being expressed in written discourse in both corpora of (advanced) German learner writing (i.e. CALE-GE and ICLE-GE). This stage is followed by a detailed description and categorization of the lexico-grammatical means for expressing contrast in learner writing. Subsequently, comparative analyses, quantitative as well as qualitative, are carried out, in order to reveal possible patterns and determinants of variation that exist in the novice academic writing.

Preliminary findings reveal a slight degree of genre-induced variation in German learners' writing in terms of sentence placement of the contrastive item *however*, see Table 1 below.

Corpus	Corpus size, N of tokens	Initial	Non-initial	Total
ICLE-GE	234.423	103	125	228
%		45	55	
CALE-GE	55.000	49	27	76
%		64	36	

Table 1: Position of the contrastive item *however*

As the table shows, German learners seem to prefer the initial sentence positioning of *however* in academic (CALE-GE), rather than in argumentative (ICLE-GE) writing. Thus, the item *however* found in the sentence initial position is almost 1,5 times more frequent in term papers than in argumentative essays. This seems to tie in well with one of the findings recently reported by Wagner (2011). In her empirical study, she points out a tendency for *however* to take up the initial sentence position in literature and cultural studies texts, rather than in linguistic texts and general corpora (2011:43). Due to a modest number of words contained in the version of the CALE corpus used at the time of analysis (see Table 1), the preliminary finding reported in the current paper should be treated with caution. A further analysis of a greater number of occurrences in a bigger corpus is needed in order to provide more empirical evidence for supporting and accounting for this finding.

5. Conclusion

The project presented in the present paper sets out to explore advanced IL-specific strategies for coping with a writing-related task in the context of English academic and argumentative writing. This is achieved by combining a functional-pedagogical view with a variationist perspective on learner writing and annotating the rhetorical function of contrast in the two corpora of learner writing. At the same time, the findings of the project will contribute to the area of variation in novice native English academic writing and will further a definition of the native speaker norm, which advanced learners are generally expected to aim at.

6. References

- Ädel, A. (2008): Involvement features in writing: do time and interaction trump register awareness? In G. Gilquin, S. Papp. & M. B. Díez-Bedmar (Eds.), *Linking up Contrastive and Learner Corpus Research*. Amsterdam, Atlanta: Rodopi, pp. 35-53.
- Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. (1999): *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.
- Biber, D., Conrad, S. (2001): Introduction: Multidimensional analysis and the study of register variation. In S. Conrad & D. Biber (Eds.), *Variation in English: Multidimensional Studies*. London: Longman,

- pp. 3-13.
- Biber, D. (2006): *University Language: A Corpus-Based Study of Spoken and Written Registers*. Amsterdam: John Benjamins.
- Callies, M. (2008): Easy to understand but difficult to use? Raising constructions and information packaging in the advanced learner variety. In G. Gilquin, S. Papp. & M. B. Díez-Bedmar (Eds.), *Linking up Contrastive and Learner Corpus Research*. Amsterdam, Atlanta: Rodopi, pp. 201-226.
- Coxhead, A. (2000): A new academic word list. *TESOL Quarterly*, 34(2), pp. 213-238.
- Gilquin, G., Paquot, M. (2008): Too chatty: Learner academic writing and register variation. *English Text Construction*, 1(1), pp. 41-61.
- Granger, S. (1996): From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg & M. Johansson (Eds.), *Languages in Contrast. Text-Based Cross-Linguistic Studies*. Lund Studies in English 88. Lund: Lund University Press, pp. 37-51.
- Granger, S. (2003): The international corpus of learner English: A new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly*, 37(3), pp. 538-546.
- Granger, S. (2004): Computer learner corpus research: Current status and future prospects. In U. Connor & T. Upton (Eds.), *Applied Corpus Linguistics: A Multidimensional Perspective*. Amsterdam, Atlanta: Rodopi, pp. 123-145.
- Granger, S. (2009): In search of a general academic vocabulary: A corpus-driven study. Paper Presented at the International Conference 'Options and Practices of L.S.A.P Practitioners', 7-8 February 2009. University of Crete, Heraklion, Crete.
- Granger, S., Paquot, M. (Forthcoming): Language for Specific Purposes. Retrieved from http://sites.uclouvain.be/cecl/archives/GRANGER_P_AQUOT_Forthcoming_Language_for_Specific_Purposes_Learner_Corpora.pdf, 17.12.2010.
- Hyland, K. (2008): As can be seen: lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(1), pp. 4-21.
- Kerz, E., Haas, F. (2009): The aim is to analyse NP: the function of prefabricated chunks in academic texts. In R. Corrigan, E. Moravcsik, H. Ouali & K. Wheatley (Eds.), *Formulaic Language: Volume 1. Distribution and historical change*. Amsterdam, Philadelphia: John Benjamins, pp. 97-117.
- Nesi, H. (2008): BAWE: An introduction to a new resource. In A. Frankenberg-Garcia, T. Rkibi, M. Braga da Cruz, R. Carvalho, C. Direito & D. Santos-Rosa (Eds.), *Proceedings of the 8th Teaching and Language Corpora Conference*. Held 4-6 July 2008 at the Instituto Superior de Línguas e Administração. Lisbon, Portugal: ISLA, pp. 239-246.
- Ortega, L., Byrnes, H. (2008): Theorizing advancedness, setting up the longitudinal research agenda. In L. Ortega & H. Byrnes (Eds.), *The Longitudinal Study of Advanced L2 Capacities*. New York: Routledge/Taylor & Francis, pp. 3-20.
- Paquot, M. (2010): *Academic Vocabulary in Learner Writing: From Extraction to Analysis*. United States: Continuum Publishing Corporation.
- Römer, U. (2007): Learner language and the norms in native corpora and EFL teaching materials: a case study of English conditionals. In: S. Volk-Birke & J. Lippert (Eds.), *Anglistentag 2006 Halle. Proceedings*. Trier: Wissenschaftlicher Verlag, pp. 355-63.
- Wagner, S. (2011): Concessives and contrastives in student writing: L1, L2 and genre differences. In J. Schmied (Ed.), *Academic Writing in Europe: Empirical Perspectives*. Göttingen: Cuvillier, pp. 23-49.
- Wulff, S. & Römer, U. (2009): Becoming a proficient academic writer: Shifting lexical preferences in the use of the progressive. *Corpora*, 4(2), pp. 115-133.

Tools to Analyse German-English Contrasts in Cohesion

Kerstin Kunz, Ekaterina Lapshinova-Koltunski

Universität des Saarlandes

Universität Campus, 66123 Saarbrücken

E-mail: k.kunz@mx.uni-saarland.de, e.lapshinova@mx.uni-saarland.de

Abstract

In the present study, we elaborate resources to semi-automatically analyse German-English contrasts in the area of cohesion. This work is an example of applications for corpus data extraction that is designed for the analysis of cohesion from both a system-based and a text-based contrastive perspective

Keywords: cohesion, contrastive analysis, corpus linguistics, extraction of linguistic knowledge, German-English contrasts

1. Introduction

To obtain empirical evidence of cohesion in English and German texts we carry out a corpuslinguistic analysis, which includes investigating a broad range of cohesive phenomena. We particularly focus on the analysis of various types of cohesive devices, the linguistic expressions to which they connect (the antecedents), the nature of the semantic ties established as well as the properties of cohesive chains. Our main research questions are 1) Which cohesive resources provided by the language systems of English and German are instantiated in different registers? 2) How frequent are they? 3) Which cohesive meanings do they express?

Thus, both system-based and text-based contrastive methods to compare English and German in terms of textuality have to our knowledge not received much attention so far, cf. table 1.

With our research, we intend to focus on cohesion as one particular aspect of textuality. As a starting point for our empirical analysis, we take the classification by (Halliday & Hasan, 1976), according to which cohesion mainly includes five categories: **reference, substitution, ellipsis, conjunctive relations and lexical cohesion.**

	system-based approaches	text-based approaches	
discourse/text	information packaging,	e.g. (Fabricius-Hansen 1999), (Doherty 2006)	✓
	cohesive devices	⊗ e.g. (Bosch <i>et al.</i> 2007), (Gundel <i>et al.</i> 2004)	✓
	comprehensive account of cohesion	⊗	⊗

Table 1: Contrastive system- and text-based studies available for English and German

Substantial research gaps in these areas justify such an enterprise: On the one hand, comprehensive accounts of cohesion are only existent from a monolingual perspective, e.g. in (Halliday & Hasan, 1976), (Schubert, 2008), (Linke *et al.*, 2001), (Brinker, 2005). On the other hand, empirical monolingual or contrastive analyses on the level of text and discourse mainly deal with individual phenomena, cf. (Fabricius-Hansen, 1999) and (Doherty, 2006) for certain aspects of information packaging and (Bosch *et al.*, 2007), (Gundel *et al.*, 2004) for the investigation of particular cohesive devices.

In this contribution, we describe our tools to extract evidence for these categories from the English- German corpus GECCo, cf. (Amoia *et al.*, submitted). Currently there are no comprehensive resources known to us that offer a repository of the coherence building systems of one or more language(s)¹. Our analysis design permits

¹ We can only name some resources providing annotations of individual cohesive phenomena, e.g. pronoun coreference in the BBN Pronoun Coreference and Entity Type Corpus, cf. (Weischedel and Brunstein 2005), verbal phrase ellipsis in (Bos and Spenader 2011) or conjunctive relations in PDTB, cf. (Prasad *et al.* 2008) for English, or annotation of anaphora in (Dipper and Zinsmeister 2009) for German.

new insights into cohesive phenomena across languages, contexts and registers. The elaboration of the procedures to extract such phenomena includes compilation, annotation and exploitation of GECCo, which consists of 10 registers of both written and spoken texts, as shown in table 2. The written part of GECCo includes 8 registers² which are based on the CroCo corpus, cf. (Neumann, 2005).

languages	registers
Written (imported from CROCo)	
EO, GO, Etrans, Gtrans	FICTION, ESSAY, INSTR, POPSCI, SHARE, SPPECH, TOU, WEB
spoken	
EO, GO	INTERVIEW ACADEMIC

Table 2: Registers in GECCo

The spoken part contains interviews (INTERVIEW) and academic speeches (ACADEMIC) produced by native speakers of the two languages³. We have chosen such a corpus constellation as we expect considerable differences in frequency and function of cohesive devices between written and spoken registers. Moreover, we depart from the assumption that there is a continuum from written to spoken mode rather than a clear dividing line.

registers and metadata on the text level as shown in figure 1. It additionally contains clause-based alignment of originals and translation⁴. We intend to semi-automatically annotate spoken registers with the information available for the written part, developing a set of automatic procedures for this task. The annotation layer on text level will be also enhanced with metadata information on language variation, speaker age, etc. Further annotations such as coreference, lexical chaining and cohesion disambiguation based on the analyses in (Kunz & Steiner, in progress)'s and (Kunz 2010) will be integrated into both parts of GECCo.

3. Procedures to Analyse Cohesion

The annotated corpus is encoded to be queried with CQP (Corpus Query Processor)⁵. We also plan to encode it for further existing query engines, e.g. ANNIS2 described in (Zeldes et al., 2009). The extracted information on cohesion will be imported into semiautomatic annotation tools in order to refine the corpus annotations on different levels, cf. figure 2.

As mentioned above, the annotated corpus can already be queried with CQP, which allows two types of attributes: positional (e.g. for part-of-speech and morphological features) and structural (e.g. for clauses or metadata).

With the help of CQP-based queries that include string, part-of-speech, text and register constraints we are able to extract linguistic items expressing the cohesion categories introduced in section 1. above and classify

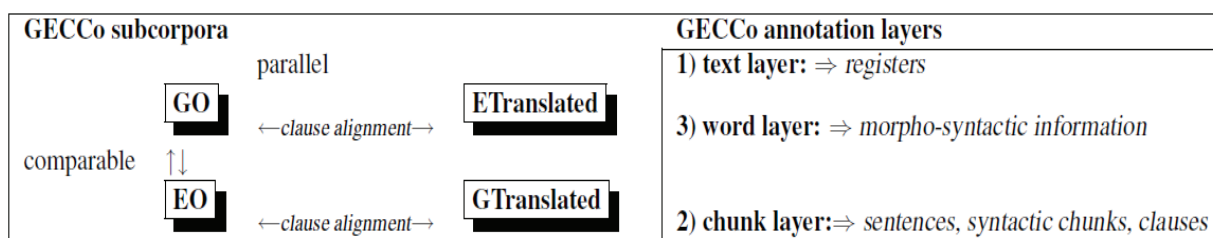


Figure 1: Annotation layers in GECCo

The written part of the multilingual corpus is already annotated with information on lemma, morphology, pos on the word level; sentences, grammatical functions, predicate-argument structures on the chunk level;

them according to their specific textual functions. We use our linguistic knowledge on cohesive devices to develop sets of complex queries with CQP that enable the extraction of cohesion from GECCo. The obtained data are subject to statistical validation (e.g. significance tests

² popular-scientific texts (POPSCI), tourism leaflets (TOU), prepared speeches (SPEECH), political essays (ESSAY), fictional texts (FICTION), corporal communication (SHARE), instruction manuals (INSTR) and websites (WEB).

³ This corpus part will be public and available on the web.

⁴ EO=English originals, GO=German originals, ETrans=English translations, Gtrans=German translations in table 2.

⁵ cf. in (Christ 1994).

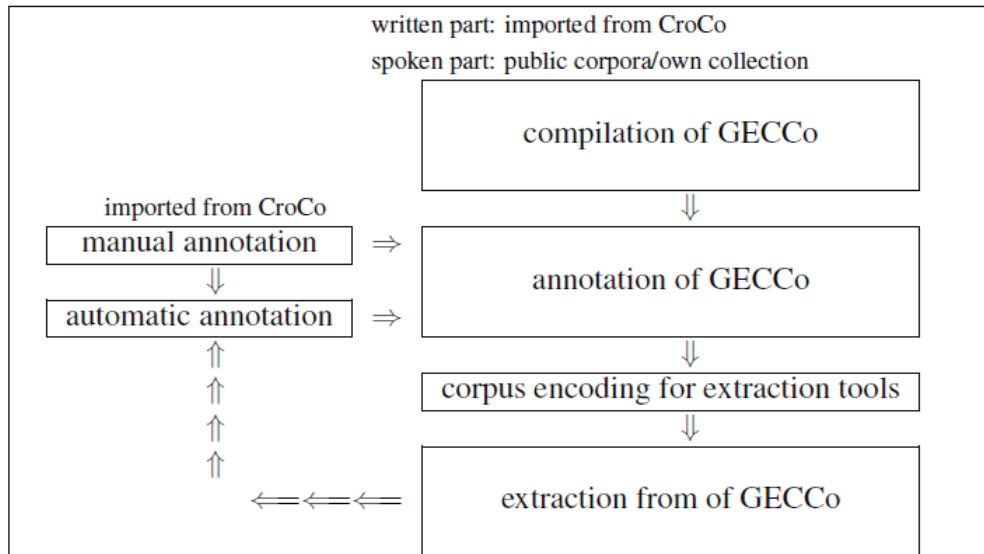


Figure 2: Procedures to analyse Cohesion in GECCo

or variation and cluster analysis) with R, with the help of which we can disambiguate and classify cohesive devices.

Moreover, CQP can also be employed to incrementally improve the corpus annotations, which allows us to semi-automatically enrich the corpus with the annotations on the information extracted as shown in figure 3. However, our observations show that representing nested structures or constituents containing gaps (necessary for annotation of coreference or ellipsis) within CQP is rather problematic, cf. (Amoia et al., submitted). As mentioned above, we therefore attempt to exploit GECCo with further query engines available, e.g. ANNIS2.

4. Preliminary Results

Our preliminary extraction results already show that there exist systematic regularities of language- and register-dependent contrasts in frequency with respect to personal reference. As an example, consider our findings for the distribution of neuter forms of third person pronouns at sentence-initial position in figure 4 (EO = English Original, GO = German Original, ETrans = English Translation, GTrans = German Translation, cf. figure 2). The left side shows the distribution in percentage of sentence initial occurrences of cohesive *it/es*. The right side displays the total numbers for all instances and cohesive instances of sentence initial *it/es*. In addition, we could already show in the analysis of the German demonstrative pronouns *der*, *die*, *das* that there

is a heterogeneity in frequency and function across registers which goes beyond assumptions drawn in the frame of earlier systemic and also textual accounts. For instance, the findings displayed in table 3 suggest a written-spoken continuum, with the register INSTR at one end and INTERVIEW at the other end of the continuum, rather than a clear-cut distinction between written and spoken registers (as already postulated above). Moreover, the differences in numbers between *das* and *der*, *die* call for an in-depth analysis with respect to distinct functions.

	der	die	das
GO_SPEECH	4	4	173
Gtrans_SPEECH	3	-	38
GO_FICTION	15	12	113
Gtrans_FICTION	10	7	100
GO_POPSCI	4	1	110
Gtrans_POPSCI	3	1	44
GO_TOU	9	2	31
Gtrans_TOU	2	1	14
GO_SHARE	3	1	44
Gtrans_SHARE	3	-	46
GO_ESSAY	1	3	90
Gtrans_ESSAY	-	-	49
GO_INSTR	-	-	20
Gtrans_INSTR	-	-	18
GO_WEB	1	2	31
Gtrans_WEB	1	-	27
GO_INTERVIEW	19	47	506

Table 3: Occurrences of *der*, *die*, *das* in German subcorpora

5. Conclusion

The described resources to extract comprehensive linguistic knowledge on cohesion will find application in various linguistic areas. First, they should provide us with evidence for our hypotheses on English-German contrasts in cohesion described in (Kunz & Steiner, in progress). Second, they should yield an initial understanding of how contrast and contact phenomena on the level of cohesion affect language understanding and language production. Furthermore, the obtained information on cohesive mechanisms of English and German will provide valuable insights for language teaching, particularly for translator/ interpreter training. Our tools will also offer new incentives for the automatic exploitation of cohesion, e.g. in machine translation, as they permit extraction from parallel corpora.

6. Acknowledgements

The authors thank the DFG (Deutsche Forschungsgemeinschaft) and the whole GECCo team for supporting this project.

7. References

- Brinker, K. (2005): *Linguistische Textanalyse: Eine Einführung in Grundbegriffe und Methoden*. 6 edition. Berlin: Erich Schmidt.
- Christ, O. (1994): A modular and flexible architecture for an integrated corpus query system. In *Proceedings of the 3rd Conference on Computational Lexicography and Text Research*. Budapest, Hungary.
- Dipper, S., Zinsmeister, H. (2009): Annotation discourse anaphora. In *Proceedings of the Workshop "Third Linguistic Annotation Workshop", LAW III, ACL-IJCNLP 2009*. Suntec, Singapore, pp. 166169.
- Doherty, M. (2006): *Structural Propensities. Translating nominal word groups from English into German*. Amsterdam/ Philadelphia: Benjamins.
- Fabricius-Hansen, C. (1999): Information packaging and translation: Aspects of translational sentence splitting (German - English/ Norwegian). In *Studia Grammatica*, 47, pp. 175-214.
- Gundel, J. K., Hedberg, N., Zacharski, R. (2004): Demonstrative pronouns in natural discourse. In *Proceedings of the Fifth Discourse Anaphora and Anaphora Resolution Colloquium*. Sao Miguel, Portugal. pp. 81-86.
- Halliday, M.A.K., Hasan, R. (1976): *Cohesion in English*. London, New York: Longman.
- Kunz, K., Steiner, E. (in progress): Towards a comparison of cohesion in English and German - contrasts and contact. Submitted for *Functional Linguistics*. London: Equinox Publishing Ltd.
- Kunz, K. (2010): *Variation in English and German Nominal Coreference. A Study of Political Essays*. Frankfurt am Main: Peter Lang.
- Linke, A., Nussbaumer, M., Portmann, P.R. (2001): *Studienbuch Linguistik*. 4 edition. Tübingen: Niemeyer.
- Neumann, S. (2005): *Corpus Design*. Deliverable No. 1 of the CroCo Project.
- Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., Webber, N. (2008): Penn Discourse Treebank Version 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*. Marrakech.
- Schubert, C. (2008): *Englische Textlinguistik. Eine Einführung*. Berlin: Schmidt.
- Weischedel, R., Brunstein, A. (2005): *BBN Pronoun Coreference and Entity Type Corpus*. Linguistic Data Consortium, Philadelphia.
- Zeldes, A., Ritz, J., Ludeling, A., Chiarcos, C. (2009): Annis: A search tool for multi-layer annotated corpora. In *Proceedings of Corpus Linguistics 2009*, Liverpool, July 20-23, 2009.

Comparison and Evaluation of ontology extraction systems

Stefanie Reimers

University of Hamburg

E-mail: 4reimers@informatik.uni-hamburg.de

Abstract

This paper presents the results of an evaluation and comparison of the two semi-automatic, corpus based ontology extraction systems OntoLT and Text2Onto. Both systems were applied to a German corpus and their outputs were evaluated in two steps. First, the Text2Onto-Ontology was evaluated against a Gold Standard ontology, represented by a manually created ontology. Second, both automatically extracted ontologies of the systems were compared to each other. Additional to this, the usability of the tools has been discussed, in order to provide some hints to improve the design of future ontology extracting systems.

Keywords: ontology, ontology learning, ontology extraction, ontology evaluation

1. Introduction

During the last years the application area of ontologies has massively been enlarged. They are not longer only a part of the vision of the semantic web but also used in intelligent search engines, information systems and in the field of model based system engineering. Therefore the need of ontologies increases similarly. But the creation of ontologies is often accompanied by a huge manual effort so that this process remains very time- and cost-intensive. Existing editors like Protégé¹ support the work of ontology developers and make it more comfortable but they only can reduce a little bit of the needed effort. Hence, techniques which reduce the manual part of the process by employing automatic methods are desirable. Corpus based ontology extraction tools seem to be the solution. They take as input a domain specific text corpus and output a domain ontology. Text is especially nowadays an excellent data source because of its permanently updated availability on the web. Under ideal circumstances the ontology extraction process should be fully automatic and produce a domain ontology of good quality. Up to date this remains unrealizable, because an important part of knowledge can't be inferred from text corpora: the commonsense knowledge. Consequently semi-automatic extraction systems present the maximal degree of support during the ontology engineering process. Several tools have been developed and are partly

free available on the web. But how well do they perform? At last they are only useful, if they heavily reduce the manual effort compared to the traditional ontology engineering process. This aspect includes on the one hand that the tool should be easy to use and on the other hand that the resulting ontology should be of good quality, comparable to a manual created one. Another interesting question is, how the outputs of systems differ, if they are applied to the same corpus.

Several works on the evaluation of ontology extraction systems have been published during the last years. But none of them considered a Gold Standard Evaluation against a manually created ontology. Furthermore, there hasn't been an attempt which used a German text corpus as data source. This work aims on exploring these missing aspects by figuring out, how great the advantage of OntoLT² and Text2Onto³ is compared to a manual creation process of an ontology. Therefore both systems were applied to the German text corpus of the Language Technology for eLearning project (LT4eL⁴), their outputs were compared to each other and finally, the Text2Onto-ontology was evaluated against the manually created LT4eL-ontology.

Section 2 introduces the ontology extraction systems OntoLT and Text2Onto as well as the LT4eL-corpus and the LT4eL-ontology. Section 3 gives a short review about

¹ <http://protege.stanford.edu/>

² <http://olp.dfki.de/OntoLT/OntoLT.htm>

³ <http://code.google.com/p/text2onto>

⁴ <http://www.let.uu.nl/lt4el>

current studies dealing with the evaluation of tools of this kind. Section 4 deals with the actual evaluation of the systems and the produced ontologies.

2. Presentation of the used systems and the data resources

The used ontology extraction systems were, because they are freely available and because they are able to process German texts.

2.1. OntoLT

OntoLT is a java based Protégé-Plugin. The in this work employed version 2.0 is exclusively compatible with Protégé 3.2, which is also freely online available. It takes as input a corpus of linguistically annotated texts in XML⁵ format. There are no requirements for a specific XML format, because the user can customize the tool for various formats. This takes place by changing the implemented XPath⁶-expressions which allow addressing specific linguistic elements (like sentences, noun phrases, head nouns, etc). They are needed for the extraction process which is performed via so called *mapping rules*. Those rules determine which concepts, instances and relations will be automatically extracted. Some rules are already implemented but the user has also the possibility to integrate new ones by using the OntoLT native *precondition language*. Rules consist of two parts: constraints and operators. If certain constraints are satisfied, one or more operators take effect. Operators can create concepts and concept properties as well as attach instances to existing concepts.

In this work, only the implemented rules were used. They specify that concepts will be created according to all heads of noun phrases in the corpus. If there exist adjectives, which belong to the nouns, they will be combined with the concept and result in a subconcept. Another rule effects the extraction of relations, which are inferred from the predicates – together with the subject and its direct objects - of sentences. After the application of the rules, the extracted concepts, relations and instances can be manipulated with the help of Protégé (Buitelaar et al., 2004).

2.2. Text2Onto

Text2Onto is also a java based application and realized as a standalone system. It requires the prior installation of Gate 4.0⁷ and WordNet 2.0⁸, which are both open source. The input consists of a corpus in text, html or pdf format. No linguistic preprocessing is required because the system provides its own preprocessing. Supported languages are English, partially also Spanish and German. The extraction process consists of several steps, which itself consists of different implemented algorithms. The user can chose between the algorithms or employ a combination of them. For example, there are three different methods for identifying concept candidates: rtf⁹, tf-idf¹⁰ and C/NC-value. The results of the algorithms are saved in a so called *probabilistic ontology model (POM)*. It consists of a set of instantiated *modeling primitives*, which are independent of a specific ontology representation language. Each instance gets a numerical value between 0 and 1 (computed by the algorithms), indicating the probability, that it deals with a for the ontology relevant element. The elements, together with its values, are then presented to the user, who shall be supported in the selection process by the assigned values. The instantiation of the primitives takes place by accessing the declarative definition in the *modeling primitive library (MPL)*.

Modeling primitives are: concepts, subconcepts, instances and relations. *Ontology writers* are responsible for the translation of the POMs into a specific ontology language like OWL¹¹ or RDFS¹² (Cimiano & Völker, 2005).

2.3. The LT4eL-corpus

The corpus originates from the LT4eL project and consists of 69 German texts. They were selected by the project participants and belong to the domain *Information Technology for End Users & eLearning*. All texts deal with introductions about how to use programs (like Excel and Word), internet and eLearning. The corpus includes 69 files, on average 5732 words per file and a total of 395547 words. 752 different domain relevant keywords were (manually) identified, which are

⁷ <http://gate.ac.uk/download/index.html>

⁸ <http://wordnet.princeton.edu>

⁹ relative term frequency

¹⁰ Term Frequency Inverse Document Frequency

¹¹ <http://www.w3.org/TR/2004/REC-owl-features-20040210>

¹² <http://www.w3.org/TR/2004/REC-rdf-concepts-20040210>

⁵ <http://www.w3.org/standards/xml>

⁶ <http://www.w3.org/TR/xpath>

all covered by the LT4eL-ontology. Ideally, on the basis of this corpus automatically extracted ontology should also semantically cover all keywords.

The files of the corpus are available in two formats: in text format and in a linguistically annotated xml format, all encoded in utf-8¹³. The text files serve as input for Text2Onto, the xml files for OntoLT. The xml format was determined by the LT4eL members. Sentence structure, noun phrases and tokens as well as corresponding lemmas, parts of speech and some morpho-syntactic information (person, number, gender, case) are annotated. A snippet of an annotated file is presented in figure 1.

```
<tok base="normalerweise" class="word" ctag="ADV" id="t37" msd="0,0,0,0,0,0,modal" rend="div,div,div"> Normalerweise </tok>
<tok base="ignorieren" class="word" ctag="VVFIN" id="t38" msd="sg,0,0,third,0,present,0,0" rend="div,div,div"> ignoriert </tok>
<tok base="Excel" class="word" ctag="NE" id="t39" msd="sg,0,nom,third,0,0,0,0" rend="div,div,div"> Excel </tok>
```

Figure 1 Sample linguistic annotation

The linguistic information is located in the values of the attributes of the token tags. *base* references the lemma, *ctag* the part of speech¹⁴ and *msd* contains morpho-syntactic data. Those are the for OntoLT relevant aspects. The complete corpus of 69 files is used for the Gold Standard evaluation of the Text2Onto-Ontology. Unfortunately, not all files could be processed by OntoLT. The reason for this circumstance could not be detected during this work. Therefore it was not possible to perform a Gold Standard evaluation of the OntoLT-ontology, because the Gold Standard ontology was generated on the basis of the whole corpus, so that a comparison would be unfair. Alternatively, the OntoLT-ontology was compared to a Text2Onto-ontology, extracted on the basis of a reduced form of the corpus. This reduced corpus consists of the files, which could be processed by OntoLT. It contains 43 files, on average 4760 words per file and a total of 204378 words (Mossel, 2007).

2.4. The LT4eL-ontology

The LT4eL-ontology was created on the basis of manually annotated keywords of the corpus. The project members modeled adequate concepts, corresponding to those keywords. They also added further sub- and superconcepts (for example: if *Notepad* was identified as

concept, also *text editor* and *editor* were added as superconcepts). Finally, the ontology was connected to the upper ontology DOLCE Ultralite¹⁵. All in all the ontology contains 1275 concepts – 1002 of them are domain concepts – 1612 subconcept-relations, 116 further relations, including 42 subrelations. Each concept comes with an English definition and a natural language representation. The ontology is available as an owl-file in xml representation (Mossel, 2007).

3. State of the art

During the last two years there were amongst others three publications of studies in the field of evaluation of semi-automatic ontology extraction tools, which used OntoLT and/or Text2Onto.

Hatala et al. (2009) tested the systems OntoGen¹⁶ and Text2Onto mainly according to their usability but also in relation to the quality of the produced ontologies. They used English corpora. 28 participants used the tools and answered questionnaires afterwards. The evaluation showed that the ontology extraction process via Text2Onto was accompanied by two central issues: 1) Due to a missing user guide the participants were not able to preview, what kind of effects the different algorithms or their combination would have on the resulting ontology. 2) The integrated extraction methods identified an enormous amount of concept candidates (several thousand) and the user was supposed to review all items according to their adequacy. Furthermore the quality of the produced ontologies was categorized as very poor, because they were flat and not appropriate to represent the demanded domain knowledge. The OntoGen-Tool was judged as more comfortable and user-friendly than Text2Onto. The participants felt to be more involved into the extraction process and were satisfied with the well structured ontologies, which included several relations (Hatala et al., 2009).

Ahrens published her studies of OntoLT in 2010. She implemented her own extraction rules and applied them to an English corpus. Since the extracted ontology was very flat, additional superconcepts were inserted. Finally the ontology was adequate enough to represent the domain of the corpus. Ahrens concluded that OntoLT – though having some issues – would be a good support

¹³ Universal Character Set Transformation Format-8-bit

¹⁴ STTS (Stuttgart-Tübingen-TagSet)

¹⁵ <http://wiki.loa-cnr.it/index.php/LoaWiki:DOLCE-UltraLite>

¹⁶ <http://ontology.ijis.si/>

during the ontology engineering process (Ahrens, 2010). Also 2010 Park et al. made their work public. They evaluated the systems OntoLT, Text2Onto, OntoBuilder¹⁷ and DODDLE¹⁸ by applying them to an English corpus. They took the usability as well as the quality of the produced ontologies into account. They considered OntoLT to be less user-friendly because the input corpus has to be linguistically preprocessed. After all, Text2Onto was judged as the best tool because of its flexibility on the one hand according to the input format and on the other hand according to the applicability of different extraction algorithms (Park et al., 2010).

All presented studies treat the evaluation of ontology extraction tools. Nevertheless one can't infer predictions or expectations for the evaluation scenario in this work. The results are somehow contradictory: Hatala et al. weren't satisfied with Text2Onto but Park et al. judged it as the best of all tested systems. Ahrens classified OntoLT as helpful, though Park et al. criticized its user-friendliness. Additionally to this, none of the studies includes a comparison between an automatically constructed and a manually created ontology. This fact and the application of a German corpus distinguish this work from all so far published ones.

4. Evaluation

4.1. Gold Standard Evaluation

The Text2Onto-ontology contained 10174 concepts, 13 subconcept relations and 945 instances. But only 981 concepts, 3 subconcept relations and 18 instances made sense. Most of the extracted items were either not domain relevant or consisted of strings, which couldn't be interpreted (due to the partial supported linguistic analysis for German texts). No further relations were identified. The ontology covers ca. 56 % of all domain relevant terms of the corpus. Altogether, its quality is not as high as that of the manually created, well structured LT4eL-ontology. The Text2Onto-ontology includes only few hierarchical relations, so that it is more a list of concepts than a real ontology. Also, the coverage of the domain relevant terms is very low. Most of the concepts are very specific, e.g. *PowerPoint* and *Excel* are included,

but more general concepts like *editor* are missing (although they appear in the texts).

4.2. OntoLT vs. Text2Onto

The OntoLT-ontology consisted of 3939 concepts, 2565 subconcept relations, 105 further relations and 0 instances. 829 concepts, 299 subconcept relations and 87 further relations were considered to be domain relevant. The ontology covers ca. 58 % of all domain relevant terms of the corpus. Many relevant concepts are missing, because the system only extracted terms, which appeared together with a modifier in the text.

The comparison of both semi-automatic extracted ontologies showed, that OntoLT had more problems to detect acronyms whereas Text2Onto often failed to identify compounds. The degree of coverage of domain relevant terms was similar.

It turns out, that both systems need to be improved. Especially Text2Onto extracts an enormous amount of irrelevant concept candidates, so that the user has to spend a lot of time to delete them. In general, the underlying algorithms are not adequate to identify suitable items, because they are based on statistical methods: but the domain relevance of a term mustn't be dependent of the number of its occurrence in a text corpus (Lame, 2004).

5. References

- Ahrens, M. (2010): Semi-autom. Generierung einer OWL-Ontologie aus domänensp. Texten, Dipl. Thesis.
- Buitelaar,P., Olejnik, D. Sintek, M. (2004): A Protégé Plug-in for Ontology Extraction from Text. In: Proc. of the 1st European Semantic Web Symposium.
- Cimiano, P., Völker, J. (2005): Text2Onto – A Framework for Ont. Learning and Data-driven Chance Discovery.
- Hatala, M., Siadaty, M., Gasevic, D., Jovanovic, J., Tomiai, C., (2009): Utility of Ontology Extraction Tools in the Hands of Educators. In: Proc. of the ICSC, USA.
- Lame, G. (2004): Using NLP Techniques to Identify Legal Ontology Components. In: Artificial Intelligence and Law 12, Nr.4, pp. 379-396.
- Mossel, E. (2007): Crosslingual Ontology-Based Document Retrieval. In: Proc. of the RANLP 2007.
- Park, J. Cho, W. Rho, S. (2010): Evaluation ontology extraction tools, In: Data Knowl.Eng. 69, pp. 1043-1061.

¹⁷ <http://ontobuilder.bitbucket.org/>

¹⁸ <http://doddle-owl.sourceforge.net/en/>

System Presentations

New and future developments in EXMARaLDA

Thomas Schmidt, Kai Wörner, Hanna Hedeland, Timm Lehmborg

Hamburger Zentrum für Sprachkorpora (HZSK)

Max Brauer-Allee 60

D-22765 Hamburg

E-mail: thomas.schmidt@uni-hamburg.de, kai.wörner@uni-hamburg.de, hanna.hedeland@uni-hamburg.de,
tim.m.lehmborg@uni-hamburg.de

Abstract

We present some recent and planned future developments in EXMARaLDA, a system for creating, managing, analysing and publishing spoken language corpora. The new functionality concerns the areas of transcription and annotation, corpus management, query mechanisms, interoperability and corpus deployment. Future work is planned in the areas of automatic annotation, standardisation and workflow management.

Keywords: annotation tools, corpora, spoken language, digital infrastructure

1. Introduction

EXMARaLDA¹ (Schmidt & Wörner, 2009) is a system for creating, managing, analysing and publishing spoken language corpora. It was developed at the Research Centre on Multilingualism (SFB 538) between 2000 and 2011. EXMARaLDA is based on a data model for time-aligned multi-layer annotations of audio or video data, following the general idea of the annotation graph framework (Bird & Liberman, 2001). It uses open standards (XML, Unicode) for data storage, is largely compatible with many other widely used media annotation tools (e.g. ELAN, Transcriber, CLAN) and can be used with all major operating systems (Windows, Macintosh, Linux). The principal software components of the system are a transcription editor (Partitur-Editor), a corpus management tool (Corpus Manager) and a KWIC concordancing tool (EXAKT).

EXMARaLDA has been used to construct the corpus collection of the Research Centre on Multilingualism comprising 23 multilingual corpora of spoken language (see Hedeland et al., this volume). It is also used for several larger corpus projects outside Hamburg such as the METU corpus of Spoken Turkish² (Middle Eastern Technical University Ankara, see Ruhi et al., this

volume), the GEWISS corpus of spoken academic discourse³ (Universities of Leipzig, Wrocław and Aston), the Corpus of Northern German Language Variation⁴ (SiN – Universities of Hamburg, Bielefeld, Frankfurt/O., Münster, Kiel and Potsdam) and the Corpus of Spoken Language in the Ruhrgebiet⁵ (KgSR, University of Bochum).

This paper focuses on new functionality added or improved during the last two years and sketches some plans for the future development of the system.

2. New and improved functionality

2.1. Transcription and annotation

The Partitur-Editor now provides additional support for time alignment of transcription and audio and/or video in the form of a time-based visualisation of the media signal. Navigation in this visualization is synchronised with navigation in the transcript, and the visualization can be used to specify the temporal extent of new annotations and to modify the start and end points of existing annotations. This has turned out a way to significantly improve transcription speed and accuracy.

¹ <http://www.exmaralda.org>

² <http://std.metu.edu.tr/>

³ <https://gewiss.uni-leipzig.de/de/>

⁴ <http://sin.sign-lang.uni-hamburg.de/drupal/>

⁵ <http://www.ruhr-uni-bochum.de/kgSR/>

Similarly, systematic manual annotation with (closed) tag sets is now supported through a configurable annotation panel which allows the user to define one or several hierarchical tag sets, assign tags to keyboard shortcuts and link them to specific labels of annotation layers. It is also possible to specify dependencies between different tag sets so that the user is offered only those tags which are applicable in a certain context. Annotation speed and consistency can thus be improved considerably.

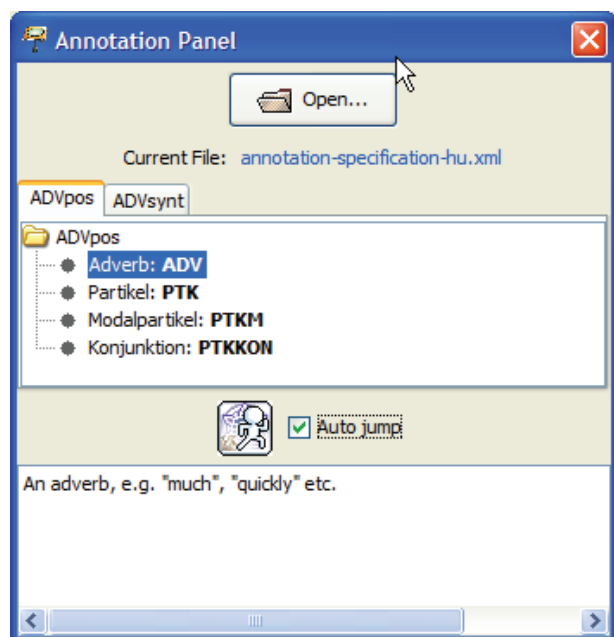


Figure 1: Annotation Panel in the Partitur-Editor

For large scale standoff annotation of corpora, a separate tool – Sextant (Standoff EXMARaLDA Transcription Annotation Tool, Wörner, 2010) – was added to the system’s tool suite. Sextant can be used to efficiently add standoff tags from closed tag sets to a segmented EXMARaLDA transcription. Annotations are stored as TEI conformant feature structures which point into transcriptions via ID references. For further processing, the standoff annotation can also be integrated into the main file.

2.2. Corpus management

The Corpus Manager was supplemented with a set of operations to aid in the maintenance of transcriptions, recordings and metadata. This includes functionality for checking the structural consistency (e.g. temporal integrity of time-alignment, correct assignment of annotations to primary layers etc.), the validity of transcriptions with respect to a given transcription

convention, as well as the completeness and consistency of metadata descriptions. Furthermore, a set of analysis functions operating on a corpus as a whole was added. Users can now generate and manipulate global type/token and frequency lists for a given corpus, perform global search and replace routines or generate corpus statistics according to different parameters. These new features are intended to facilitate both corpus construction and corpus use.

2.3. Query mechanisms

For the query tool EXAKT, several new features were added to support the user in formulating complex queries to a corpus.

A Levenshtein calculation was made available which selects from a given list of words all entries which are sufficiently similar to a form selected by the user. This can help to minimize the risk that (potentially unpredictable) variants – as are common in spoken language corpora – are accidentally overlooked in queries.

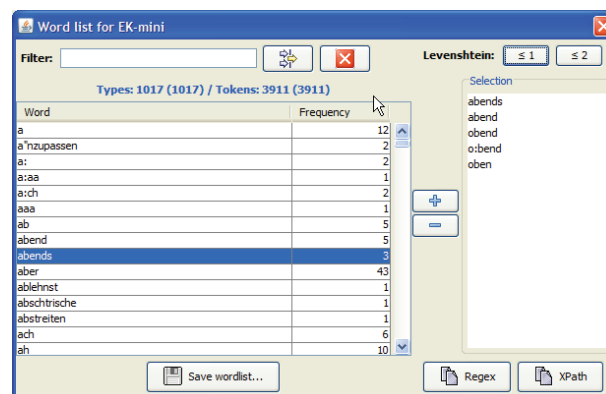


Figure 2: Word list with Levenshtein functionality in EXAKT

A regular expression library can now be used to store and retrieve common queries. This is meant mainly as a help for those users who are not experts in the design of formal queries.

Through an extension of the original KWIC functionality, EXAKT is now also able to carry out queries across two or more annotation layers. This is achieved by adding one or more so called annotation columns in which annotation data from a specified annotation level overlapping with the existing search results are added to the concordance. The type of overlap between annotations can be specified as illustrated in figure 3.

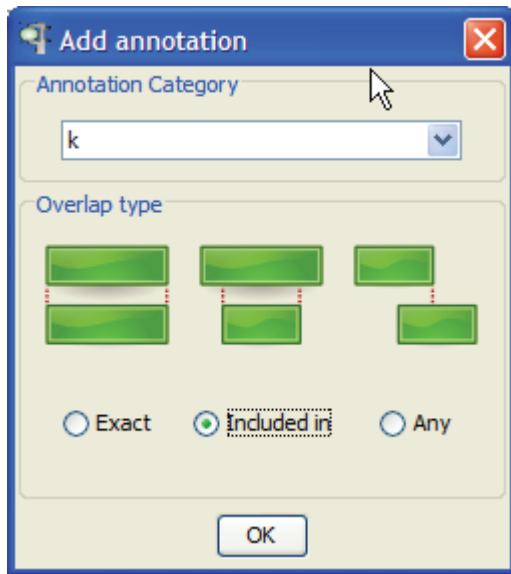


Figure 3: Specifying the overlap type for a multilevel search in EXAKT

2.4. Interoperability

Much work has been invested to further improve and optimise EXMARaLDA's compatibility with other widely used transcription and annotation tools. Wizards for importing entire corpora from Transcriber, FOLKER, CLAN and ELAN were integrated into EXAKT thereby considerably extending the tool's area of application. Moreover, a proposal for a spoken language transcription standard based on the P5 version of the TEI guidelines was formulated (Schmidt, 2011), and a droplet application (TEI-Drop) was added to the EXMARaLDA toolset which enables users to easily transform Transcriber, FOLKER, CLAN, ELAN or EXMARaLDA files into this TEI conformant format.

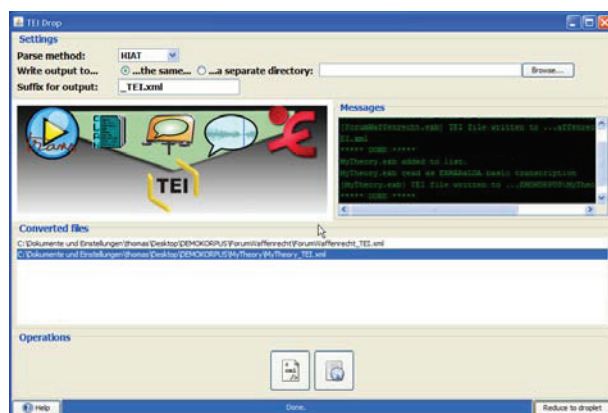


Figure 4: Screenshot of TEI-Drop

2.5. Corpus deployment

Completed EXMARaLDA corpora can now also be made available (i.e. queried) via a relational database with

EXAKT. Compared to the deployment in the form of individual XML files which are then queried either locally or via http with EXAKT, this method not only facilitates data access, but also considerably improves query performance (by a factor of about 10 for smaller corpora, probably more for larger corpora) and allows for a more fine-grained access management. Furthermore, making data available in this way is also a prerequisite for integrating EXMARaLDA data into evolving distributed infrastructures like CLARIN.

With the general availability of HTML5, methods for visualizing corpus data for web presentations could also be simplified and improved considerably. The integration of transcription text and underlying audio or video recording now no longer depends on Flash technology, but can be efficiently realised with standard browser technology.

3. Future work

With the end of the maximum funding period of the Research Centre on Multilingualism in June 2011, EXMARaLDA's original context of development has also ceased to exist. Although the system is now in a stable state and should remain usable for quite some time with some minimal maintenance work, we still see much potential for future development in at least three areas.

3.1. Automatic annotation

Additional manual and automatic annotation methods are required in order to make spoken language corpora more useful for corpus linguistic research. We have consequently started to explore the application of methods developed for written language, such as automatic part-of-speech-tagging or lemmatisation to EXMARaLDA corpora.

First tests were carried out on the Hamburg Map Task Corpus (HAMATAC, Hedeland & Schmidt, 2012) with TreeTagger (Schmid, 1995), which was integrated via the TT4J interface (Eckart de Castilho et al., 2009) into EXMARaLDA. HAMATAC was POS-tagged and lemmatised with the default German parameter file, trained on written newspaper texts. The data were first tokenized using EXMARaLDA's segmentation functionality which segments and distinguishes words, punctuation, pauses and non-phonological segments. Only words and punctuation were fed as input into the

tagger in the sequence in which they occur in the transcription. The tagging results were saved as EXMARaLDA standoff annotation files which can be further processed in the Sextant tool (see above). A student assistant was instructed to manually check and correct all POS tags. An evaluation shows that roughly 80% of POS tags were assigned correctly. The error rate is thus considerably higher than for the best results which can be obtained on written texts (about 97% correct tags). By far the most tagging errors, however, occurred with word forms which are specific to spoken language, such as hesitation markers (“äh”, “ähm”), interjections and incomplete forms (cut-off words). Since especially the former are highly frequent but very limited in form (three forms *äh*, *ähm* and *hm* account for about half of the tagging errors), we expect a retraining of the TreeTagger parameter file on the corrected data to lead to a much lower error rate.

3.2. Standardisation

Further work in standardisation of data models, metadata descriptions, file formats and transcription conventions is needed in order to integrate spoken language data on equal footing with written data into the language resource landscape. EXMARaLDA as one of the most interoperable systems of its kind already provides a solid basis for developing and establishing such standards. Future work in this area should attempt to consolidate this basis with more general approaches like the guidelines of the Text Encoding Initiative, standardisation efforts within the ISO framework and emerging standards for digital infrastructures.

3.3. Workflow management

As we survey, train and support users in constructing and analysing spoken language corpora with EXMARaLDA, we observe how important it is to organise the tools' functionalities into an efficient workflow. Right now, the EXMARaLDA tools operate in a standalone fashion on local file systems, leaving many important aspects of the workflow (e.g. version control, consistency checking etc.) in the users' responsibility. A tight integration of the tools with a repository solution may make it much easier, especially for larger projects, to organise their workflows and construct and publish their corpora in a maximally efficient and effective manner. We plan to explore this

possibility further in the follow-up projects at the Hamburg Centre for Language Corpora (HZSK).⁶

4. Acknowledgements

Work on EXMARaLDA was funded by the University of Hamburg and by grants from the Deutsche Forschungsgemeinschaft (DFG).

5. References

- Bird, S., Liberman, M. (2001): A formal framework for linguistic annotation. In: *Speech Communication* (33), pp. 23-60.
- Eckart de Castilho, R., Holtz, M., Teich, E. (2009): Computational support for corpus analysis work flows: The case of integrating automatic and manual annotations. In: *Linguistic Processing Pipelines Workshop at GSCL 2009 - Book of Abstracts* (electronic proceedings), October 2009.
- Hedeland, H., Schmidt, T. (2012): Technological and methodological challenges in creating, annotating and sharing a learner corpus of spoken German. To appear in: Schmidt, T., Wörner, K.: *Multilingual Corpora and Multilingual Corpus Analysis*. To appear as part of the series 'Hamburg Studies in Multilingualism' (HSM). Amsterdam: Benjamins.
- Schmid, H. (1995): Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. March 1995.
- Schmidt, T., Wörner, K. (2009): EXMARaLDA – Creating, analysing and sharing spoken language corpora for pragmatic research. In: *Pragmatics* (19:4), pp. 565-582.
- Schmidt, T. (2011): A TEI-based approach to standardising spoken language transcription. In: *Journal of the Text Encoding Initiative* (1).
- Wörner, K. (2010): *Werkzeuge zur flachen Annotation von Transkriptionen gesprochener Sprache*. PhD Thesis, Universität Bielefeld, <http://bieron.ub.uni-bielefeld.de/volltexte/2010/1669/>.

⁶ <http://www.corpora.uni-hamburg.de>

Der VLC Language Index

Dirk Schäfer, Jürgen Handke

Institut für Anglistik und Amerikanistik, Philipps-Universität Marburg

Wilhelm-Röpke-Straße 6D

E-mail: {dirk.schaefer,handke}@staff.uni-marburg.de

Abstract

Der Language Index ist eine Sammlung von Audiodaten von Sprachen der Welt. Als Bestandteil der Online-Lernplattform "Virtual Linguistics Campus" repräsentiert der Language Index Sprachaufnahmen in standardisierter Form und typologische Informationen mit Web-Technologien, die zum Zwecke der Analyse, z.B. in der Lehre, verwendet werden können.

Keywords: Audio-Korpus, Typologie, Web

1. Übersicht

Der Language Index als Teil der Online Lernplattform „Virtual Linguistics Campus“ ist eine Sammlung von strukturierten Audiodaten von Sprachen der Welt. Im Rahmen einer Systemvorführung stellen wir vor, wie die Daten präsentiert werden und wie Forscher die vorhandenen Audiodaten nutzen können. Der restliche Artikel beschreibt das Datenformat für die Sprachaufnahmen, und die Benutzerschnittstellen.

2. Erstellung von Sprachaufnahmen

Die Sprachaufnahmen stellen einen Parallelkorpus dar, da von jedem Sprecher dieselben Wörter, Halbsätze und Sätze gesprochen wurden. Zu diesem Zweck existiert eine Sammlung von standardisierten Datenblättern, die erweitert wird, sobald eine neue Sprache hinzukommt. Für manche Sprachen existieren mehrere leicht voneinander abweichende Datenblätter, da alle Sprecher die Daten entsprechend ihres regionalen Dialekts übersetzt haben. Zurzeit verfügen wir über Datenblätter zu 110 Sprachen und Regionaldialekten, sowie Sprachaufnahmen von 850 Sprechern.

2.1. Verfahren zur Gewährleistung der Qualität von Sprachaufnahmen

Um die Qualität der Sprachaufnahmen zu gewährleisten, hat sich folgendes Verfahren bewährt:

- Der Sprecher überprüft das vorhandene Datenblatt zu seiner Sprache. Ist kein Datenblatt zu seiner Sprache verfügbar, übersetzt er die Keywords und Sätze.
- Verfügt die Sprache über kein Schriftsystem, werden

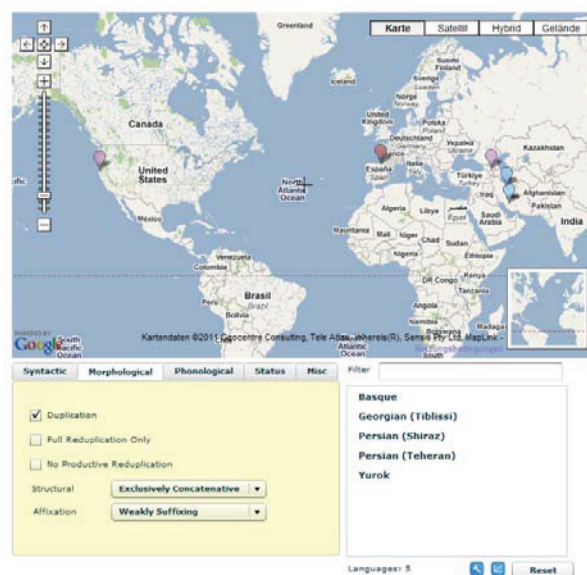


Abbildung 1: Benutzerschnittstelle

- die Datenblätter auf der Basis des IPA-Alphabets durch Interaktion mit dem Sprecher erstellt.
- Der Sprecher liest die Keywords und Sätze im normalen Tempo vor. Die Aufnahme erfolgt vor Ort mit einem digitalen Aufnahmegerät, über das Web mit Hilfe von Skype oder mit einem Headset am heimischen Computer.
- Die aufgenommenen Sprachdaten werden nachbearbeitet und mit Cuepoints versehen.
- Die vollständige Sprachaufnahme mitsamt Transkription und Transliteration wird dem Sprecher zu Kontrolle vorgelegt.
- Die Aufnahme wird über den VLC Language Index verfügbar gemacht.

3. Benutzerschnittstelle

Es gibt besondere Schwierigkeiten bei der Repräsentation solcher audiobasierter Parallelkorpora. Zum Beispiel muss eine einfache Benutzbarkeit gewährleistet sein, die ohne Einarbeitungszeit einen schnellen Zugriff auf alle gewünschten Daten ermöglicht. Außerdem liegt es in der Natur eines Parallelkorpus, dass Vergleichsmöglichkeiten gegeben sein müssen.

Der Language Index ist eine auf Webtechniken basierende Anwendung mit hohen Flash und Flex Anteilen. Seit 2006 wird die Google Maps API zur Darstellung von Sprachdaten auf Karten eingesetzt. Mit dem Anwachsen des Datenbestandes wurde eine Datenbanknutzung notwendig, besondere Verfahren mussten eingesetzt werden, um eine performante Kommunikation zwischen PHP und den auf Flex basierenden Benutzeroberflächen zu gewährleisten.

Der Zugriff auf die Audiodaten im VLC Language Index ist auf verschiedene Weisen möglich:

- Eine Liste von Sprachaufnahmen, nach Sprachen sortiert.
- Eine Google-Map bei der jede Sprachaufnahme als Pin dargestellt wird, beim Daraufklicken öffnet sich ein Popup-Fenster.
- Ein Filterinterface bei dem sich bestimmte syntaktische, morphologische, phonologische und weitere Parameter einstellen lassen.

4. Besondere Features

Es gibt zusätzliche besondere Features, die sich mit dem Datenbestand des Parallelkorpus realisieren lassen. Mit Hilfe des „Cognate Comparison“ Werkzeugs können die Benutzer nach Wahl eines Kognats akustisch miteinander vergleichen, indem der Benutzer Pins auf einer Karte oder Einträge in einer Liste auswählt.

Auf „Acoustic Vowel Charts“ werden die Frequenzen derselben Vokale verschiedener Sprecher visualisiert.

5. Ausblick

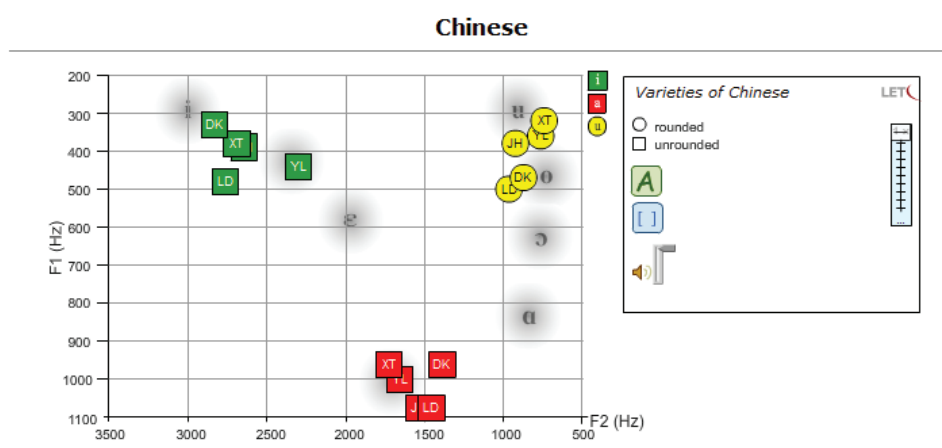
Es gibt verschiedene Einsatzszenarien, für die der VLC Language Index genutzt werden kann, dazu gehören die Lehre, Abschlußarbeiten und Forschung auf Master- und PhD-Niveau.

Jüngstes Feature ist das mp3-Download-Angebot mit einem bibliographischen Referenzierungssystem für alle Sprachdaten, damit diese auf einfache Weise in anderen Werkzeugen genutzt werden können und wissenschaftlichen Arbeiten, die auf diesen Daten basieren, beigelegt werden können.

6. Weblink

Virtual Linguistics Campus (VLC):

<http://www.linguistics-online.de>



The main vowels of Chinese, [i], [a] and [u], spoken by 5 Speakers:

- **Jun Han** (Standard); JH
- **Yuxin Lei** (Suining); YL
- **Xinhue Tian** (Xiangfan); XT
- **Li Li Dong** (Shandong/Changyi); LD
- **Daidao Kun** (Qingdao); DK

Abbildung 2: Acoustic Vowel Chart

2. The New Architecture

Our goal is to make all existing layers of annotation available for simultaneous search, but in a way that allows each one to be searched separately without intervening nodes from other annotation layers. For this purpose, we have converted the latest Version 6 of TüBa-D/Z to the multi-layer XML format PAULA (Dipper, 2005). We then converted and edited the corpus using the SaltNPepper converter framework (Zipser & Romary, 2010), which gives us an in-memory representation of the corpus that can be manipulated more easily. During this step, we disentangled the syntactic, topological and coreference annotations. The resulting corpus was then exported and fed into ANNIS (Zeldes et al., 2009), a corpus search and visualization tool for multi-layer corpora. The resulting annotation layers are visualized in Figure 2, which shows a separate syntax tree (without topological fields), spans representing fields, and a full document view for the coreference annotation in which coreferent expressions are highlighted in the same color.

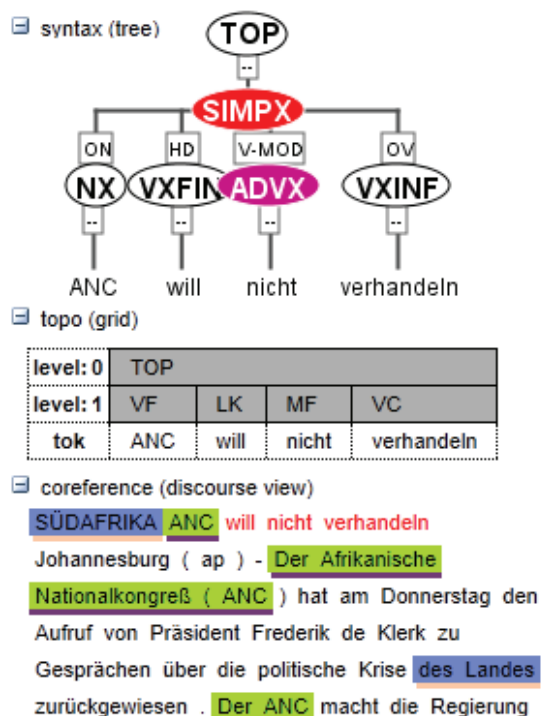


Figure 2: TüBa-D/Z in ANNIS with separate annotation layers.

3. Corpus Search

Using the new architecture and the ANNIS Query Language (AQL)¹ it becomes possible to query syntax, topological fields and coreference more easily and intuitively, both simultaneously and separately. In the following, we will discuss three example applications briefly: one investigating topological fields only, one combining all three annotation layers, and one extracting syntactic frames with the help of the exporter functionality in ANNIS.

3.1. Application 1: Topological Fields

As a simple example of the easily accessible topological field information, we can consider the following query, which retrieves clauses before the left sentence bracket, in the main-clause preverbal domain (Vorfeld, VF), which contain two complementizer fields (C) after one another (the operator '>' represents dominance, and '.*' represents indirect precedence, the numbers '#1' etc. refer to the nodes declared at the beginning of the query, in order):

```
(1) field="VF" & field="C" & field="C"
& #1 > #2 & #1 > #3 & #2 .* #3
```

Figure 3 shows an example result with its separate field grid and syntax tree, for the sentence: *Daß und wie Demokratie funktionieren kann, hat der zähe Kampf der Frauen um den Paragraphen 218 gezeigt* 'The women's tenacious fight for paragraph 218 has shown that, and how, democracy can work.' By directly querying the topological fields we can avoid having to consider possible syntactic nodes intervening between VF and C.

3.2. Application 2: Coreference, Syntax and Fields

Next, let us first search for objects that are cataphors, but not reflexive pronouns. In TüBa-D/Z, cataphors are linked to their subsequents via the 'cataphoric' relation. The AQL expression is given in (2a): there is a node – any node – number 1 and another node, number 2, and node 1 points to node 2 using the 'cataphoric' relation.

```
(2a.) node & node & #1 ->cataphoric #2
```

We now add syntactic constraints: the cataphor, node 1, shall be an object (OA or OD, i.e. accusative or dative). In TüBa-D/Z, the grammatical function of a noun phrase

¹ A tutorial of the query language can be found at <http://www.sfb632.uni-potsdam.de/~d1/annis/>.

topo (grid)

level: 0	TOP																		
level: 1	VF											LK	MF	VC					
level: 2	C		G	MF	VC														
tok	Daß	und	wie	Demokratie	funktionieren	kann	.	hat	der	zähe	Kampf	der	Frauen	um	den	Paragrafen	218	gezeigt	.

Figure 3: Separate fields and syntactic phrases for VF with two C positions

(NX) is specified as a label of the dominance edge connecting this NX and its parent.

```
(2b.) node & node & #1 ->cataphoric
#2 & cat="NX" & #3 == #1 & cat & #4
>[func=/O[AD]/] #3 & pos!="PRF" & #5
== #1
```

(read: there is a node, number 3, of category NX, and node 3 covers the same tokens as node 1, and there is a node of any category, and this node number 4 dominates ‘>’ node number 3, with the edge label ‘func’ (function) = OA or OD. We use regular expressions to specify the label. To exclude reflexive pronouns (part of speech ‘PRF’), we use negation (‘!=’)). The search yields 51 results, with scalable contexts and color-highlighting of the matches (cataphors and their subsequents).

Secondly, let us query for noun phrases in the VF, with a definite determiner and their antecedents in the left-neighbour sentences.

```
(2c.) field="VF" & cat="NX" & #1 ==
#2 & pos="ART" & #2 > #3 &
tok=/[Dd]../ & #3 == #4 & node & #5
== #2 & node & #5 ->coreferential #6
& cat="TOP" & #7 _i_ #1 & cat="TOP" &
#8 _i_ #6 & #8 . #7
```

(‘>’ represents indirect dominance)

This query yields 766 results. Using the match counts of (2c.) and similar queries, we can create a contingency table of definite vs. pronomial VF-constituents and whether their respective antecedents occur in the left-neighbour sentence (‘close’) or more distantly: 43% of the definites and 61% of the pronouns in VF have a ‘close’ antecedent – a difference that is highly significant ($\chi^2=142.72$, $p<.0005$).

3.3. Application 3: Syntactic Dependency

Using the pure syntax trees, it is also much easier to find verbal arguments regardless of their topological environment. For example, the following query finds all direct object head nouns for the verb *schreiben* ‘write’.

Using the new lemma annotation in Version 6 of TüBa-D/Z and the fact that in the pure syntax trees, a verb and its arguments are dominated by the same clause node, query (3) becomes possible:

```
(3) cat="SIMPX" & cat=/VX.* / &
lemma="schreiben" & cat="NX" &
pos="NN" & #1 > #2 & #2 > #3 & #1
>[func="OA"] #4 & #4 >[func="HD"] #5
```

This query searches for a verbal phrase dominated by a clause (SIMPX) and dominating the lemma *schreiben*, where the same clause also dominates a nominal phrase with the function OA, which in turn dominates its head noun (pos="NN", func="HD"). Using the built-in WEKA exporter, we can produce a list of all the nominal object arguments of a verb much like in a dependency treebank, along with the form and part-of-speech of the relevant verb, as shown in Figure 3. Note that both finite and non-finite clauses are found, as well as verb-second and verb-final clauses, which now all have similar tree structures regardless of topological fields.

```
'271192', 'wer immer seine Texte schreibt', 'SIMPX',
'271186', 'Texte', 'apm', 'NN', '271187', 'seine Texte',
'NX', '271189', 'schreibt', '3sis', 'VVFIN', '271190',
'schreibt', 'VXFIN'
'1134826', 'Songs schreiben', 'SIMPX', '1134820',
'Songs', 'apm', 'NN', '1134821', 'Songs', 'NX',
'1134823', 'schreiben', '-', 'VVINF', '1134824',
'schreiben', 'VXINF'
'1526561', 'Ich schreibe Satire', 'SIMPX', '1519602',
'Satire', 'asf', 'NN', '1526559', 'Satire', 'NX',
'1519599', 'schreibe', '1sis', 'VVFIN', '1526557',
'schreibe', 'VXFIN'
```

Figure 3: Excerpt of results from the WEKA Exporter for query (3).

The exporter gives the values of all annotations for the nodes we have searched for, in order, as well as the text covered by those nodes. We can therefore easily get

tabular access to the contents of the clause (e.g. *Songs schreiben* ‘to write songs’), the object (*Songs*), the form and part-of-speech of the verb (*schreiben*, VVINF), morphological annotation (apm for a plural masculine noun in the accusative), etc.

4. Conclusion

We have suggested an advanced, layer-separated representation architecture for TüBa-D/Z. This architecture facilitates corpus querying and exploitation. By means of examples, we have shown that the corpus search tool ANNIS allows for a qualitative and quantitative study of the interplay of syntactic, topological and information structural factors annotated in TüBa-D/Z.

5. References

- Dipper, S. (2005): XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation. *Proceedings of Berliner XML Tage 2005 (BXML 2005)*. Berlin, Germany, pp. 39-50.
- Hinrichs, E. W., Kübler, S., Naumann, K., Telljohann, H., Trushkina, J. (2004): Recent developments in linguistic annotations of the TüBa-D/Z treebank. *Proceedings of the Third Workshop on Treebanks and Linguistic Theories*.
- Lezius, W. (2002): *Ein Suchwerkzeug für syntaktisch annotierte Textkorpora*. PhD Thesis, Institut für maschinelle Sprachverarbeitung Stuttgart.
- Mengel, A., Lezius, W. (2000): An XML-based encoding format for syntactically annotated corpora. *Proceedings of the Second International Conference on Language Resources and Engineering (LREC 2000)*. Athens.
- Müller, C., Strube, M. (2006): Multi-Level Annotation of Linguistic Data with MMAX2. In: Braun, Sabine, Kohn, Kurt & Mukherjee, Joybrato (eds.), *Corpus Technology and Language Pedagogy*. Frankfurt: Peter Lang, pp. 197-214.
- Telljohann, H., Hinrichs, E. W., Kübler, S. (2003): *Stylebook for the Tübingen Treebank of Written German*.
- Zeldes, A., Ritz, J., Lüdeling, A., Chiarcos, C. (2009): ANNIS: A Search Tool for Multi-Layer Annotated Corpora. *Proceedings of Corpus Linguistics 2009, Liverpool, July 20-23, 2009*.
- Zipser, F., Romary, L. (2010): A model oriented approach to the mapping of annotation formats using standards. *Proceedings of the Workshop on Language Resource and Language Technology Standards, LREC 2010*. Malta, pp. 7-18.

MT Server Land Translation Services

Christian Federmann

DFKI – Language Technology Lab

Stuhlsatzenhausweg 3, D-66123 Saarbrücken, GERMANY

E-mail: cfedermann@dfki.de

Abstract

We demonstrate MT Server Land, an open-source architecture for machine translation services that is developed by the MT group at DFKI. The system can flexibly be extended and allows lay users to make use of MT technology within a web browser or by using simple HTTP POST requests from custom applications. A central broker server collects and distributes translation requests to several worker servers that create the actual translations. User access is realized via a fast and easy-to-use web interface or through an XML-RPC-based API that allows integrating translation services into external applications. We have implemented worker servers for several existing translation systems such as the Moses SMT decoder or the Lucy RBMT engine. We also show how other, web-based translation tools such as Google Translate can be integrated into the MT Server Land application. The source code is published under an open BSD-style license and is freely available from GitHub.

Keywords: Machine Translation, Web Service, Translation Framework, Open-Source Tool

1. Introduction

Machine translation (MT) is a field of active research with lots of different MT systems being built for shared tasks and experiments. The step from the research community towards real-world application of available technology requires easy-to-use MT services that are available via the Internet and allow collecting feedback and criticism from real users. Such applications are important means to increase visibility of MT research and to help shaping the multi-lingual web. Applications such as Google Translate¹ allow lay users to quickly and effortlessly create translations of texts or even complete web pages; the continued success of such services shows the potential that lies in usable machine translation, something both developers and researchers should be targeting.

In the context of ongoing MT research projects at DFKI's language technology lab, we have decided to design and implement such a translation application. We have released the source code under a permissive open-source license and hope that it becomes a useful tool for the MT community. A screenshot of the MT Server Land application is shown in Figure 1.

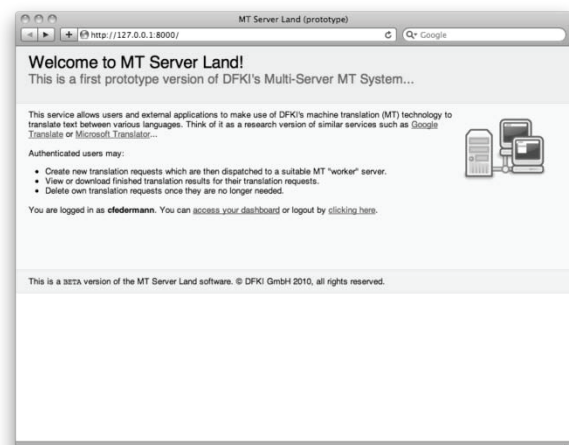


Figure 1: Screenshot of MT Server Land

2. System Architecture

The system consists of two different layers: first, we have the so-called **broker** server that handles all direct requests from end users or via API calls alike. Second, we have a layer of **worker** servers, each implementing some sort of machine translation functionality. All communication between users and workers is channeled through the broker server that acts as a central “proxy” server. An overview of the system architecture is given in Figure 2.

For users, both broker and workers “constitute” the MT

¹ <http://translate.google.com>

Server Land system; the broker server is the “visible” part of the application while the various worker servers perform the “invisible” translation work. The system has been designed to make it easier for lay users to access and use machine translation technology without the need to fully dive into the complexities of current MT research. Within MT Server Land, translation functionality is available by starting up suitable worker server instances for a specific MT engine. The startup process for workers is standardized using some easy-to-understand parameters for, e.g., the hostname/IP address or port number of the worker server process. All “low-level” work (de-/serialization, transfer of requests/results, etc.) is handled by the worker server instances. Of course, it is possible to design and create new worker server instances, e.g., to demonstrate new features in a research translation system or to integrate other MT systems.

Human users connect to the system using any modern web browser; API access can be implemented using HTTP POST and/or XML-RPC requests. It would be relatively easy to extend the current API interface to support other protocols such as SOAP or REST. By design, all internal method calls that connect to the worker layer have to be implemented with XML-RPC. In order to prevent encoding problems with the input text, we send and receive all data encoded as Base64 Strings between broker and workers; the broker server takes care of the necessary conversion steps. Translation requests are converted into serialized, binary Strings using Google protocol buffer compilation.

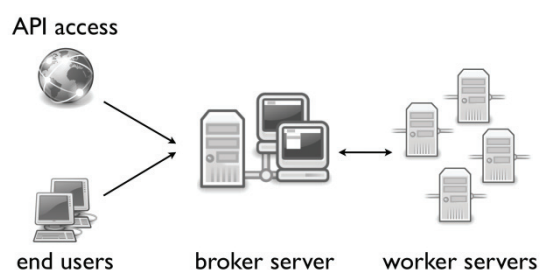


Figure 2: Architecture overview of MT Server Land

2.1. Broker Server

The broker server has been implemented using the **django web framework**², which takes care of low-level tasks and allows for rapid development and clean design of components. We have used the framework for other project work before and think it is well suited to the task. The framework itself is available under an open-source BSD-license.

2.1.1. Object Models

The broker server implements two main **django models** that we describe subsequently. Please note that we have also developed additional object models, e.g. for quota management. See the source code for more information.

- **WorkerServer** stores all information related to a remote worker server. This includes source and target language, the respective hostname and port address as well as a name and a short description. Available worker servers within MT Server Land can be constrained to function for specific user and/or API accounts only.
- **TranslationRequest** models a translation job and related information such as the chosen worker server, the source text and the assigned request id. Furthermore we store additional metadata information. Once the translation result has been obtained from the translation worker server, it is also stored within the instance so that it can be removed from the worker server’s job queue.

2.1.2. User Interface

We developed a browser-based web interface to access and use the MT Server Land application. End users first have to authenticate before they can access their **dashboard** that lists all known translation requests for the current user and also allows creating new requests. When creating a new translation request, the user may choose which translation worker server should be used to generate the translation for the chosen language pair. We use a validation step to ensure that the respective worker server supports the selected language pair and is currently able to receive new translation requests from the broker server; after successful validation, the new translation request is sent to the worker server that starts processing the given source text.

² <http://www.djangoproject.com/>

Once the chosen worker server has completed a translation request, the result is transferred to (and also cached by) the object instance inside the broker server's data storage. The user can view the result within the dashboard or download the file to a local hard disk. Translation requests can be deleted at any time, effectively terminating the corresponding thread within the connected worker server (if the translation is still running). If an error occurs during translation, the system will recognize this and deactivate the respective translation requests.

2.1.3. API Interface

In parallel to the browser interface, we have designed and implemented an API that allows connecting applications to the MT functionality provided by our application using HTTP POST requests. Again, we first require authentication before any machine translation can be used. We provide methods to list all requests for the current “user” (i.e. the application account) and to create, download, or delete translation requests. Extension to REST or SOAP protocols is possible.

2.2. Worker Server Implementations

A layer of so-called worker servers that are connected to the central broker server implements the actual machine translation functionality. For the MT Server Land, we have implemented worker servers for the following MT systems:

- **Moses SMT**: a Moses (Koehn et al., 2007) worker is configured to serve exactly one language pair. We use the Moses Server mode to keep translation and language model in memory, which helps to speed up the translation process. As the limitation to one language pair effectively means that a huge number of Moses worker server instances has to be started in a typical usage scenario, we have also worked on a better implementation which allows to serve any number of language pairs from one worker instance. Future improvements could be achieved by integrating “on-the-fly” configuration switching and remote language models to reduce the amount of resources required by the Moses worker server.
- **Lucy RBMT**: our Lucy (Alonso & Thurmair, 2003) worker is implemented using a Lucy Server mode wrapper. This is a small Python program running on

the Windows machine on which Lucy is installed. We have implemented a simple XML-RPC based API interface to send translation requests to the Lucy engine and later fetch the corresponding results. For integration in MT Server Land, we simply had to “tunnel” our Lucy worker server calls to this Lucy server mode implementation.

- **Joshua SMT**: similar to the Moses worker, we have created a Joshua (Li et al., 2010) worker that works by creating a new Joshua instance for each translation request.

We have also created worker servers for popular online translation engines such as **Google Translate**, **Microsoft Translator**, or **Yahoo! BabelFish**. We will demonstrate the worker servers in our presentation.

3. Acknowledgements

This work was supported by the EuroMatrixPlus project (IST-231720) that is funded by the European Community under the Seventh Framework Programme for Research and Technological Development.

4. References

- Alonso, J. A., Thurmair, G. (2003). The Compendium Translator System. In Proceedings of the Ninth Machine Translation Summit.
- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C. J., Bojar, O., Constantin, A., Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pp. 177–180, Prague, Czech Republic. Association for Computational Linguistics.
- Li, Z., Callison-Burch, C., Dyer, C., Ganitkevitch, J., Irvine, A., Khudanpur, S., Schwartz, L., Thornton, W., Wang, Z., Weese, J., Zaidan, O. (2010). Joshua 2.0: A Toolkit for Parsing-based Machine Translation with Syntax, Semirings, Discriminative Training and other Goodies. In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, pp. 133–137, Uppsala, Sweden. Association for Computational Linguistics.

WORKING PAPERS IN MULTILINGUALISM • Series B
ARBEITEN ZUR MEHRSPRACHIGKEIT • Folge B

Publications to date • Bisher erschienen:

1. Jürgen M. Meisel: *On transfer at the initial state of L2 acquisition: Revisiting Universal Grammar.*
2. Kristin Bührig, Latif Durlanik & Bernd Meyer: *Arzt-Patienten-Kommunikation im Krankenhaus: konstitutive Handlungseinheiten, institutionelle Handlungslinien.*
3. Georg Kaiser: *Dialect contact and language change. A case study on word-order change in French.*
4. Susanne S. Jekat & Lorenzo Tessitore: *End-to-End Evaluation of Machine Interpretation Systems: A Graphical Evaluation Tool.*
5. Thomas Ehlen: *Sprache - Diskurs - Text. Überlegungen zu den kommunikativen Rahmenbedingungen mittelalterlicher Zweisprachigkeit für das Verhältnis von Latein und Deutsch.*
Nikolaus Henkel: *Lateinisch-Deutsch.*
6. Kristin Bührig & Jochen Rehbein: *Reproduzierendes Handeln. Übersetzen, simultanes und konsekutives Dolmetschen im diskursanalytischen Vergleich.*
7. Jürgen M. Meisel: *The Simultaneous Acquisition of Two First Languages: Early Differentiation and Subsequent Development of Grammars.*
8. Bernd Meyer: *Medizinische Aufklärungsgespräche: Struktur und Zwecksetzung aus diskursanalytischer Sicht.*
9. Kristin Bührig, Latif Durlanik & Bernd Meyer (Hrsg.): *Dolmetschen und Übersetzen in medizinischen Institutionen. Beiträge zum Kolloquium 'Dolmetschen in Institutionen' vom 17. - 18.03. 2000 in Hamburg.*
10. Juliane House: *Concepts and Methods of Translation Criticism: A Linguistic Perspective.*
11. Bernd Meyer & Notis Toufexis (Hrsg.): *Text/Diskurs, Oralität/Literalität unter dem Aspekt mehrsprachiger Kommunikation.*
12. Hans Eideneier: *Zur mittelalterlichen Vorgeschichte der neugriechischen Diglossie.*
13. Kristin Bührig, Juliane House, Susanne J. Jekat: *Abstracts of the International Symposium on Linguistics and Translation, University of Hamburg, 20th - 21st November 2000.*
14. Sascha W. Felix: *Theta Parametrization. Predicate-Argument Structure in English and Japanese.*
15. Mary A. Kato: *Aspects of my Bilingualism: Japanese as L1 and Portuguese and English as L2.*
16. Natascha Müller, Katja Cantone, Tanja Kupisch & Katrin Schmitz: *Das mehrsprachige Kind: Italienisch – Deutsch.*
17. Kurt Braunmüller: *Semiconmunication and Accommodation: Observations from the Linguistic Situation in Scandinavia.*
18. Tessa Say: *Feature Acquisition in Bilingual Child Language Development.*
19. Kurt Braunmüller & Ludger Zeevaert: *Semikommunikation, rezepptive Mehrsprachigkeit und verwandte Phänomene. Eine bibliographische Bestandsaufnahme.*
20. Nicole Baumgarten, Juliane House & Julia Probst: *Untersuchungen zum Englischen als 'lingua franca' in verdeckter Übersetzung. Theoretischer Hintergrund, Weiterentwicklung des Analyseverfahrens und erste Ergebnisse.*
21. Per Warter: *Lexical Identification and Decoding Strategies in Interscandinavian Communication.*
22. Susanne J. Jekat & Patricia J. Nüßlein: *Übersetzen und Dolmetschen: Grundlegende Aspekte und Forschungsergebnisse.*
23. Claudia Böttger & Julia Probst: *Adressatenorientierung in englischen und deutschen Texten.*
24. Anja Möhring: *The acquisition of French by German children of pre-school age. An empirical investigation of gender assignment and gender agreement.*
25. Jochen Rehbein: *Turkish in European Societies.*
26. Katja Francesca Cantone & Marc-Olivier Hinzelin: *Proceedings of the Colloquium on Structure, Acquisition, and Change of Grammars: Phonological and Syntactic Aspects. Volume I.*
27. Katja Francesca Cantone & Marc-Olivier Hinzelin: *Proceedings of the Colloquium on Structure, Acquisition, and Change of Grammars: Phonological and Syntactic Aspects. Volume II.*
28. Utta v. Gleich: *Multilingualism and multilingual Literacies in Latin American Educational Systems.*
29. Christine Glanz & Utta v. Gleich: *Mehrsprachige literale Praktiken im religiösen Alltag. Ein Vergleich literaler Ereignisse in Uganda und Bolivien.*
30. Jürgen M. Meisel: *From bilingual language acquisition to theories of diachronic change.*
31. Florian Coulmas & Makoto Watanabe: *Japan's Nascent Multilingualism.*
32. Tanja Kupisch: *The acquisition of the DP in French as the weaker language.*
33. Utta v. Gleich, Mechthild Reh & Christine Glanz: *Mehrsprachige literale Praktiken im Kulturvergleich: Uganda und Bolivien. Die Datenerhebungs- und Auswertungsmethoden.*
34. Thomas Schmidt: *EXMARaLDA - ein System zur Diskurstranskription auf dem Computer.*
35. Esther Rinke: *On the licensing of null subjects in Old French.*
36. Bernd Meyer & Ludger Zeevaert: *Sprachwechselphänomene in gedolmetschten und semikommunikativen Diskursen.*

37. Annette Herkenrath & Birsal Karakoç: *Zum Erwerb von Verfahren der Subordination bei türkisch-deutsch bilingualen Kindern – Transkripte und quantitative Aspekte.*
38. Guntram Haag: *Illokution und Adressatenorientierung in der Zweitler Gesamtübersetzung und der Melker Rumpfbearbeitung der 'Disticha Catonis': funktionale und sprachliche Einflussfaktoren.*
39. Kristin Bührig: *Multimodalität in gedolmetschten Aufklärungsgesprächen. Grafische Abbildungen in der Wissensvermittlung.*
40. Jochen Rehbein: *Pragmatische Aspekte des Kontrastierens von Sprachen – Türkisch und Deutsch im Vergleich.*
41. Christine Glanz & Okot Bengé: *Exploring Multilingual Community Literacies. Workshop at the Ugandan German Cultural Society, Kampala, September 2001.*
42. Christina Janik: *Modalisierungen im Dolmetschprozess.*
43. Hans Eideneier: „Rhetorik und Stil“ – der griechische Beitrag.
44. Annette Herkenrath, Birsal Karakoç & Jochen Rehbein: *Interrogative elements as subordinators in Turkish – aspects of Turkish-German bilingual children's language use.*
45. Marc-Olivier Hinzelin: *The Acquisition of Subjects in Bilingual Children: Pronoun Use in Portuguese-German Children.*
46. Thomas Schmidt: *Visualising Linguistic Annotation as Interlinear Text.*
47. Nicole Baumgarten: *Language-specific Realization of Extralinguistic Concepts in Original and Translation Texts: Social Gender in Popular Film.*
48. Nicole Baumgarten: *Close or distant: Constructions of proximity in translations and parallel texts.*
49. Katrin Monika Schmitz & Natascha Müller: *Strong and clitic pronouns in monolingual and bilingual first language acquisition: Comparing French and Italian.*
50. Bernd Meyer: *Bilingual Risk communication.*
51. Bernd Meyer: *Dolmetschertraining aus diskursanalytischer Sicht: Überlegungen zu einer Fortbildung für zweisprachige Pflegekräfte.*
52. Monika Rothweiler, Solveig Kroffke & Michael Bernreuter: *Grammar Acquisition in Bilingual Children with Specific Language Impairment: Prerequisites and Questions.*
Solveig Kroffke & Monika Rothweiler: *The Bilingual's Language Modes in Early Second Language Acquisition – Contexts of Language Use and Diagnosis of Language Disorders.*
53. Gerard Doetjes: *Auf falsche[r] Fährte in der interkandinavischen Kommunikation.*
54. Angela Beuerle & Kurt Braunmüller: *Early Germanic bilingualism? Evidence from the earliest runic inscriptions and from the defixiones in Roman utility epigraphy.*
Kurt Braunmüller: *Grammatical indicators for bilingualism in the oldest runic inscriptions?*
55. Annette Herkenrath & Birsal Karakoç: *Zur Morphosyntax äußerungsinterner Konnektivität bei mono- und bilingualen türkischen Kindern.*
56. Jochen Rehbein, Thomas Schmidt, Bernd Meyer, Franziska Watzke & Annette Herkenrath: *Handbuch für das computergestützte Transkribieren nach HIAT.*
57. Kristin Bührig & Bernd Meyer: *Ad hoc-interpreting and the achievement of communicative purposes in specific kinds of doctor-patient discourse.*
58. Margaret M. Kehoe & Conxita Lleó: *The emergence of language specific rhythm in German-Spanish bilingual children.*
59. Christiane Hohenstein: *Japanese and German 'I think-constructions'.*
60. Christiane Hohenstein: *Interactional expectations and linguistic knowledge in academic expert discourse (Japanese/German).*
61. Solveig Kroffke & Bernd Meyer: *Verständigungsprobleme in bilingualen Anamnesegesprächen.*
62. Thomas Schmidt: *Time-based data models and the Text Encoding Initiative's guidelines for transcription of speech.*
63. Anja Möhring: *Against full transfer during early phases of L2 acquisition: Evidence from German learners of French.*
64. Bernadette Golinski & Gerard Doetjes: *Sprachverstehensuntersuchungen im semikommunikativen Kontext.*
65. Lukas Pietsch: *Re-inventing the 'perfect' wheel: Grammaticalisation and the Hiberno-English medial-object perfects.*
66. Esther Rinke: *Wortstellungswandel in Infinitivkomplementen kausativer Verben im Portugiesischen.*
67. Imme Kuchenbrandt, Tanja Kupisch & Esther Rinke: *Pronominal Objects in Romance: Comparing French, Italian, Portuguese, Romanian and Spanish.*
68. Javier Arias, Noemi Kintana, Martin Rakow & Susanne Rieckborn: *Sprachdominanz: Konzepte und Kriterien.*
69. Matthias Bonnesen: *The acquisition of questions by two German-French bilingual children*
70. Chrystalla A. Thoma & Ludger Zeevaert: *Klitische Pronomina im Griechischen und Schwedischen: Eine vergleichende Untersuchung zu synchroner Funktion und diachroner Entwicklung klitischer Pronomina in griechischen und schwedischen narrativen Texten des 15. bis 18. Jahrhunderts*
71. Thomas Johnen: *Redewiedergabe zwischen Konnektivität und Modalität: Zur Markierung von Redewiedergabe in Dolmetscheräußerungen in gedolmetschten Arzt-Patientengesprächen*
72. Nicole Baumgarten: *Converging conventions? Macrosyntactic conjunction with English 'and' and German 'und'*
73. Susanne Rieckborn: *Entwicklung der ‚schwachen Sprache‘ im unbalancierten L1-Erwerb*

74. Ludger Zeevaert: *Variation und kontaktinduzierter Wandel im Altschwedischen*
75. Belma Haznedar: *Is there a relationship between inherent aspect of predicates and their finiteness in child L2 English?*
76. Bernd Heine: *Contact-induced word order change without word order change*
77. Matthias Bonnesen: *Is the left periphery a vulnerable domain in unbalanced bilingual first language acquisition?*
78. Tanja Kupisch & Esther Rinke: *Italienische und portugiesische Possessivpronomina im diachronischen Vergleich: Determinanten oder Adjektive?*
79. Imme Kuchenbrandt, Conxita Lleó, Martin Rakow, Javier Arias Navarro: *Große Tests für kleine Datenbasen?*
80. Jürgen M. Meisel: *Exploring the limits of the LAD*
81. Steffen Höder, Kai Wörner, Ludger Zeevaert: *Corpus-based investigations on word order change: The case of Old Nordic*
82. Lukas Pietsch: *The Irish English "After Perfect" in context: Borrowing and syntactic productivity*
83. Matthias Bonnesen & Solveig Kroffke: *The acquisition of questions in L2 German and French by children and adults*
84. Julia Davydova: *Preterite and present perfect in Irish English: Determinants of variation*
85. Ezel Babur, Solveig Chilla & Bernd Meyer: *Aspekte der Kommunikation in der logopädischen Diagnostik mit ein- und mehrsprachigen Kindern*
86. Imme Kuchenbrandt: *Cross-linguistic influences in the acquisition of grammatical gender?*
87. Anne Küppers: *Sprecherdeiktika in deutschen und französischen Aktionärsbriefen*
88. Demet Özçetin: *Die Versprachlichung mentaler Prozesse in englischen und deutschen Wirtschaftstexten*
89. Barbara Miertsch: *The acquisition of gender markings by early second language learners of French*
90. Kurt Braunmüller: *On the relevance of receptive multilingualism in a globalised world: Theory, history and evidence from today's Scandinavia*
91. Jill P. Morford & Martina L. Carlson: *Sign perception and recognition in non-native signers of ASL*
92. Andrea Bicsar: *How the "Traveling Rocks" of Death Valley become "Moving Rocks": The Case of an English-Hungarian Popular Science Text Translation*
93. Anne-Kathrin Preißler: *Subjektpromina im späten Mittelfranzösischen: Das Journal de Jean Héroard*
94. Ingo Feldhausen, Ariadna Benet, Andrea Pešková: *Prosodische Grenzen in der Spontansprache: Eine Untersuchung zum Zentralkatalanischen und porteño-Spanischen*
95. Manuela Schönenberger, Franziska Sterner, Tobias Ruberg: *The realization of indirect objects and case marking in experimental data from child L1 and child L2 German*
96. Hanna Hedeland, Thomas Schmidt, Kai Wörner (eds.): *Multilingual Resources and Multilingual Applications – Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL) 2011*

