Contrastive Linguistics, Translation Studies, Machine Translation – what can we learn from each other?

Pre-conference Workshop
GSCL 2011, Hamburg
27th September, 2011

Book of Abstracts

Oliver Čulo, Silvia Hansen-Schirra
Johannes Gutenberg-Universität Mainz

Program

9:30-10:00	Introduction Methodological cross-fertilisation: empirical methodologies in (computational) linguistics and translation studies	Erich Steiner, Silvia Hansen-Schirra, Oliver Čulo
10:00-10:50	Can linguistics and translation studies benefit from MT quality barriers? Needs an opportunities from a technology perspective (invited talk)	Hans Uszkoreit
11:15-11:40	What Can Contrastive Linguistics Tell Us about Translating Discourse Structure?	Iørn Korzen, Morten Gylling
11:40-12:05	An analysis of translational complexity in two text types.	Martha Thunes
12:05-12:30	Web-based contrastive comparison of Age-Related Temporal Phrases in Spanish and French	Sofía N. Galicia-Haro, Alexander F. Gelbukh
14:00-14:25	A Contrastive Study on Abstract Anaphors in German and English	Stefanie Dipper, Christine Rieger, Melanie Seiss, Heike Zinsmeister
14:25-14:50	A Corpus-based Contrastive Analysis to Define Minimal Semantics of Inter-sentential Dependencies for Machine Translation	Thomas Meyer, Andrei Popescu-Belis, Jeevanthi Liyanapathirana, Bruno Cartoni
14:50-15:15	Formalising translation behavior with parallel treebanks	Oliver Čulo, Silvia Hansen-Schirra
15:40-16:05	Using annotated corpora for rapid development of new language pairs in MT	Susanne Preuss, Hajo Keffer, Paul Schmidt
16:05-16:30	Phrase Table Support for Human Translation	Gerhard Kremer, Matthias Hartung, Stefan Riezler, Sebastian Padó
16:30-16:55	Inside the monitor model: processes of default and challenged translation production	Michael Carl, Barbara Dragsted
17:00	Discussion	

Methodological cross-fertilization: empirical methodologies in (computational) linguistics and translation studies

Erich Steiner
Universität des Saarlandes, Saarbrücken

Recent years have seen a few, although still limited, attempts at improving empirical methodologies in contrastive linguistics and in translation studies through interdisciplinary collaboration with multi-layer corpus architectures as developed and refined in computational linguistics. These corpus architectures provide data enriched by a variety of techniques ranging from shallow to deep processing (Vela et al 2007, Čulo et al 2008). They allow the posing of linguistic questions as empirical questions even in areas which until recently were considered the province of hermeneutic debates supported by – hopefully representative – examples.

At the same time, explanatory background for empirical results is increasingly sought in more sophisticated models of language contact in typologically based comparative linguistics (e.g. Thomason 2001, Teich 2003, Doherty 2006, Fabricius-Hansen and Ramm eds. 2008, Siemund and Kintana. eds. 2008, Steiner 2008, Miestamo et al. eds. 2008, Dunn et al 2011) on the one hand, and in language processing in situations of multilinguality, including translation, on the other (Alves et al 2010, Carl et al 2008). There remains a significant challenge, though, in closing the gap between the often necessarily high level of abstraction of models, and the data provided through shallow (and cheap), or else deeper (and more expensive), analysis and annotation of electronic corpora. This gap has to be narrowed through concerted efforts involving methodologies from computational linguistics, including machine translation, (contrastive) linguistics and translation studies.

We shall discuss two test cases from DFG-projects for such interdisciplinary work: one of them investigates a key notion of translation (*explicitation*) and the other an under-researched area of language contact (*contact through cohesion*). The gap to be closed consists between the notions of *explicitness/ explicitation* and *contact through cohesion* on the one hand, and the level of the available data (annotation layers, statistics on these, alignment phenomena such as *crossing lines,* and *empty links*) on the other. Seen relative to existing approaches, we are attempting to synthesize individual parameters of language comparisons and language contact into more general dependent variables (*explicitness, cohesion*) on the one hand, and we suggest operationalizations in such a way as to enable empirical corpus-based (and ultimately also experimental) investigations. An attempt is made to identify achievements as well as persistent methodological gaps, and implications are identified for research methodologies.

The first attempt subjected the hypothesis that translations as texts are characterized by the property of *explicitness* relative to original texts, and that this explicitness is due to the translation process, rather than to the factors of register and language (both of which play their independent roles) to elaborated tests on a corpus of originals and translations, partitioned into registers, between English and German. The corpora were compiled using sampling techniques (Biber et al 1998) and annotated for PoS, morphology, chunks, syntactic functions, clauses and sentences. A second, and important, source of data were alignments between originals and their translations on all of the levels annotated. The notions of *explicitness* and *explicitation* were given a careful operationalization in terms of the types of information contained in our data. We shall present a sub-set of the results and argue that it was possible to show *whether* and *to what extent* explicitness and *explicitation* can be traced in the available data. The independent variables *language system, register* and *translation* can be reasonably isolated and related to the observed effects in the data, but the third one of these, if interpreted as *translation process*, is inherently complex and at present still insufficiently-understood (cf. also Becher 2010). This shortcoming can be systematically addressed by subjecting the notion of *translation process* to a more detailed analysis and by independently testing its effect in processing studies involving the cumulation and intersecting of data from key-stroke logging, eye-tracking and post-hoc protocols. As a first evaluation of this line of research, it will be argued that the general corpus-architecture and the processing employed can be trusted to yield more and also methodologically refined results of the type indicated here, but that we need improvements in the areas of *modeling* (internally over-complex variables, representativeness of data), *operationalization* of the models in terms of linguistics features, and in *processing techniques* for corpus data (processing pipelines, evaluation and significance of findings) and for experimental data (amount and naturalness of data and experimental design).

The second attempt sets out from the diagnosis that our current knowledge about English-German *contrasts in cohesion* is weak. We do have reasonably comprehensive system-based accounts for contrastive grammar, yet even these are not yet backed-up by empirical validation. For cohesion, not even a system-based comparison is available, much less an empirical foundation for such a comparison. The tracing of contact phenomena on the level of cohesion is therefore necessarily still in its infancy (but cf. Hansen-Schirra et al 2007 for an early attempt). We shall argue that substantial advances in technologies using multi-layer annotated electronic corpora for text-based investigations of phenomena of cohesion hold the promise of placing constrastive accounts on an empirical basis, and beyond this comparison also allow us to trace contact phenomena in suitably configured corpora. A multi-layer representation will again be used, approaching tree-bank functionality and including aligned data for English and German translations in both directions as a crucial empirical base. Extensive frequency information about cohesive configurations will be incorporated, tied to varieties or registers of the language concerned.

One of the interesting questions is that of whether contrastive properties of cohesion in the two languages point into the same direction as some assumed generalizations in contrastive grammar (directness of mapping from semantics to grammar, different tolerance of various forms of "ellipsis", more explicit encoding in one of the languages in the clause, possibly the opposite tendency in the verb phrase, etc.), or whether cohesion serves as a dialectic counterpart, distributing constraints not in the same direction as in grammar, but possibly in the opposite one. A further interesting object of investigation is the nature of cohesive chains (frequency, length, distance between elements, etc.).
Our corpus-linguistic analysis includes identifying various types of cohesive devices (*reference, substitution, ellipsis, coherence relations, lexical cohesion*), the linguistic expressions to which they connect (the antecedents), as well as the nature of the semantic ties established and properties of the cohesive chains where appropriate. Including translations in the analysis should provide evidence for analogies between cohesive devices in the two languages, but also show areas where one-to-one equivalents are not preferred, or even non-existent.

The currently existing annotation requires an expansion in terms of additional layers of annotation. For instance, particular cohesive devices establishing reference or substitution can be investigated on the part-of-speech level. Other types such as cohesive conjunctions can be identified when examining the part-of-speech as well as the chunk level. For the investigation of ellipsis combined queries into different layers of annotation can be employed. However, for the analysis of nominal, verbal or clausal ellipsis the current annotation is too shallow and does not permit a fine-grained differentiation of types of linguistic devices. Thus, more specific cohesive categories have to be developed and annotated.

In order to narrow the gap between the concept of *contact through* cohesion and the level of our data, a structured grid of hypotheses is specified for empirical analysis as a testing ground for

- contrasts in the uses of *similar* systemic resources

- contrasts in the use of *different* systemic resources for similar cohesive functions/ purposes

- traces of language contact due to different usages in contact vs. non-contact varieties (categorical and/ or in terms of frequency).

Examples of such hypotheses are:

Hypothesis 1: 3rd person singular neuter pronouns vs. masculine and feminine pronouns (frequency E(nglish)>G(erman) for originals (contrast)), in terms of PoS overall and proportionally within their word class.

Hypothesis 2: ETrans(lations)>EO(riginals) in non-ambiguous 3rd person reference and ETrans-T(arget)T(ext)>GO(riginals)-S(ource)T(exts) in explicitated 3rd person reference through use of fully-lexical TTequivalent of pronominal source

Hypothesis 3: E>G in cohesive usage of *it* (because of alternative usage in German of demonstratives of various sorts and pronominal adverbs), measured both in terms of PoS overall and as proportion of cohesive vs. non-cohesive usage of *it*.

Hypothesis 4: EO > GTrans > GO in cohesive usage of *it* because of interference in GTrans

Hypothesis 5: In terms of the phenomena tested in H1 – H4, we predict that in a comparison of originals and translations (always within one and the same language and register), the translations will diverge from the originals in the direction of their source language.

Further hypotheses are developed for *comparisons of vagueness/ ambiguity of reference and scope*. Differences can be expected here deriving from usage of different lexicogrammatical realizations of some constant cohesive relationship, or even from different cohesive relationships altogether. An example would be the contrastive use of a generic full lexical phrase vs. a definite phrase vs. a phrase pre-modified through a determiner (possessive vs. deixis vs. demonstrative) vs. a phrase headed by a pro-form (demonstrative vs. pronoun) as tested on aligned ST-TT pairs. The interest would not be in the phenomenon as such, but in the different kinds of *ambiguity* and/ or *vagueness* associated with each case. In general, we would predict that a) translations are less ambiguous and vague than their originals in SL-TL configurations (explicitation through translation), but also b) that they diverge from their original register-identical counterparts in the direction of the respective source language (interference, shining-through).

A final type of hypothesis will make reference to contrastive register-specificity of cohesive configurations, and again their behaviour under contrast vs. contact conditions. These configurations will be operationalized as length of lexical or referential chains, density of chains, number of chains per text sample, etc. At this stage, we would hypothesize shining-through effects for ST-TT configurations, and for density of chains only a possibly increasing effect of the translation process as such. Our main argument will be that the frequency data that can be obtained through work of the type described here is valid and interesting in itself, and is furthermore only possible through the joining of efforts from (contrastive) linguistics, translation studies, and computational linguistics.

What remains a task for the immediate future in research attempts of the type discussed is an improved understanding of the cut-off point between very costly "deep" (and less reliable) annotation, and more "shallow" (and more reliable) annotation, the latter of which leaves a substantial gap between data and interpretation. We shall also raise the questions of how research architectures can be made more standardized than hitherto, allowing independent repetition and (dis-)confirmation of findings, and of how corpora, their processing pipelines and evaluated results can be related to experimental (processing) studies to pave the way towards more principled explanations of the results obtained.

References:

Alves, F., Pagano A., Neumann S., Steiner E. & Hansen-Schirra S. 2010. "Translation Units and Grammatical Shifts: Towards an Integration of Product- and Process-based Translation Research". In: Shreve, G.M. & Anglone E. (eds.) Translation and Cognition. Benjamins.109-142

Becher, Viktor. 2010. “Abandoning the notion of translation-inherent explicitation: against a dogma of translation studies. In: *Across Languages and Cultures* 11 (1), pp. 1–28 (2010)

Biber, Douglas, Susan Conrad, and Randi Reppen, 1998 *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Carl, Michael, Arnt Lykke Jakobsen, and Kristian T.H. Jensen 2008 Studying human translation behavior with user-activity data. In *Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science, NLPCS 2008, Barcelona, Spain, June 2008,* Bernadette Sharp and Michael Zock (eds.), 114–123.Setúbal, Portugal: INSTICC Press.

Čulo, Oliver, Silvia Hansen-Schirra, Stella Neumann, and Mihaela Vela 2008 Empirical studies on language contrast using the English-German comparable and parallel CroCo Corpus. In *Proceedings of the LREC 2008 Workshop "Building and Using Comparable Corpora"*,Marrakesh, Morrocco, 31 May 2008, 47–51.

Doherty, M. 2006. *Structural Propensities. Translating Nominal Groups from English into German*. Benjamins, Amsterdam:

Dunn, Michael, Simon J. Greenhill, Stephen C. Levinson & Russell D. Gray. 2011. “Evolved structure of language shows lineage-specific trends in word-order universals” in: *Nature 473. 2011: 79-82*

Fabricius-Hansen and Ramm eds. 2008 “Subordination” vs. “Coordination” in sentence and text: a cross-linguistic perspective. Amsterdam, Philadelphia: John Benjamins, Studies in Language Companion Series,

Hansen-Schirra, S., Neumann, S. & Steiner, E. 2007. “Cohesion and Explicitation in an English-German Translation Corpus”. In: Languages in Contrast 7(2): 241-265.

Miestamo, Matti; Kaius, Sinnemäki; and Karlsson, Fred. eds 2008. *Language Complexity. Typology, contact, change*. Amsterdam/ Philadelphia: John Benjamins

Siemund, P. & N. Kintana (eds.). 2008. *Language contact and contact languages*. Amsterdam: Benjamins (Hamburg Studies in Multilingualism Vol. 7).

Steiner, Erich. 2008. “Empirical studies of translations as a mode of language contact - “explicitness” of lexicogrammatical encoding as a relevant dimension.” in: Siemund, Peter and Kintana, Noemi. eds. 2008. *Language contact and contact languages*. Amsterdam: John Benjamins (Hamburg Studies in Multilingualism Vol. 7). pp. 317-346

Teich, E. 2003. *Cross-linguistic variation in system and text. A methodology for the investigation of translations and comparable texts*. Berlin, New York: de Gruyter.

Thomason, S. G. 2001. *Language Contact. An Introduction*. Washington D.C.: Georgetown University Press.

Vela, Mihaela, Silvia Hansen-Schirra, and Stella Neumann 2007 Querying multi-layer annotation and alignment in translation corpora. In *Proceedings of the Corpus Linguistics Conference CL 2007*, Birmingham, UK, 27-30 July 2007, Matthew Davies, Paul Rayson, Susan Hunston, and Pernilla Danielsson (eds.). http://ucrel.lancs.ac.uk/publications/CL2007/paper/97_Paper.pdf.

What Can Contrastive Linguistics Tell Us about Translating Discourse Structure?

Iørn Korzen, Morten Gylling
Copenhagen Business School
Dalgas Have 15, DK-2000 Frederiksberg, Denmark
E-mail: ik.ikk@cbs.dk, mgj.isv@cbs.dk

Abstract

This paper argues that translators can greatly benefit from contrastive studies of discourse structure. Cross-linguistic studies of Italian and Danish point to significant typological differences in information packaging in the two languages, especially in their use of deverbalisation. Italian sentences tend to include a larger number of Elementary Discourse Units (EDUs), especially propositions, than Danish. A higher percentage of these is rhetorically backgrounded by means of non-finite and nominalised predicates. Danish text structure, on the other hand, is more informationally linear and characterised by a higher number of finite verbs and topic shifts. These typological differences are transferred into three simple translation rules concerning 1) the number of EDUs, 2) the rhetorical structure, and 3) the textualisation of rhetorical satellites.

Keywords: discourse structure, information packaging, textualisation, deverbalisation, translation strategies.

1. Introduction

Over the last decades, Contrastive Linguistics and Translation Studies have experienced a veritable explosion of interest and attention from scholars in different fields, but the linguistic focus of attention has typically been confined to lexical and syntactic levels. Contrastive studies on discourse structure and intersentential relations, on the other hand, are much less frequent. For instance, there are extremely few cross-linguistic textual resources annotated for discourse. According to Webber, Egg and Kordoni (2010), they are limited to the ones found in the Copenhagen Dependency Treebanks (CDT), which cover five different Germanic and Romance languages: Danish, English, German, Italian, and Spanish. All CDT texts are annotated for four different linguistic layers (apart from part-of-speech): syntax, discourse, anaphora and morphology, see Buch-Kromann et al. (2010).

The research we shall present in this paper is based partly on our work with the CDT and partly on other resources, and we shall focus on two phenomena related to the information and discourse structures of texts, namely on informational density, i.e. the amount of information per sentence, and on text complexity, here defined as the degree of subordination of the text segments that the Rhetorical Structure Theory labels as “rhetorical satellites” (Mann & Thompson, 1987; Mann, Matthiessen & Thompson, 1992; Matthiessen & Thompson, 1988 and later work). Like other scholars, such as Asher and Vieu (2005), we consider these phenomena part of the “information packaging” of a text, a term suggested by Chafe (1976) and later used, especially in connection with given vs. new entities and definiteness, e.g. by Clark and Haviland (1977), Prince (1984) and Vallduvi and Engdahl (1996).

Other cross-linguistic surveys on information packaging have been conducted e.g. by Fabricius-Hansen (1996; 1999), Ramm and Fabricius-Hansen (2005) and Behrens, Solfjeld and Fabricius-Hansen (2010), who investigate English, German and Norwegian, i.e. three Germanic languages. On information density and explicitness in English-German translations, see Hansen-Schirra, Neumann and Steiner (2007). Alves et al. (2010) examine particularly grammatical shifts, e.g. between finite verbs and nominalisations, in the translation process between English and German.

In this paper, we compare two languages of different language families, viz. Danish and Italian, a Scandinavian (Germanic subgroup) and Romance language respectively. Our results regarding Danish confirm the ones obtained by the first mentioned scholars for Norwegian, whereas their findings on English and German are closer to our results concerning Italian. On the other hand, the Italian features presented in the following, are found also in other Romance languages, for which reason we consider it justified to talk about general typological differences between Scandinavian and Romance languages, *ceteris paribus,* with English and German somewhere "in between".

The paper is structured as follows: In section 2, we examine an Italian and Danish corpus of argumentative texts with regard to informational density, measured as the number of words and Elementary Discourse Units (EDUs, cf. Carlson and Marcu, 2001) per sentence. In section 3, we look at text complexity and the textualisation of rhetorical satellites, and in section 4, we formulate our findings as a few relatively simple rules for (human as well as machine) translators that work with Scandinavian and Romance languages.

2. Information density

2.1. Sentence length

Differences in discourse structure show themselves in many ways, one of which is the simple sentence length, measured as words per sentence[1]. In this context, we used the parallel Europarl corpus, an open source corpus compiled by Koehn (2005). Europarl is a very large multilingual corpus (55 million words) with source and target texts covering all the official languages of the European Union. In fact, the corpus was designed to train and evaluate statistical machine translation, but it can, as we shall see, also be used for other types of cross-linguistic studies. The Europarl texts, which are mainly argumentative (see van Halteren (2008) for a discussion of this), consist of speeches made by the members of the European Parliament from 1996 to 2010, and most of the speeches (88 %) have been tagged with a language attribute indicating the native language (L1) of the speaker. We created a Perl script[2] that extracted all Danish and Italian L1 text from the entire corpus and calculated the average sentence length of all texts. In this context, a sentence is defined as a text segment marked by a full stop, a question mark, or an exclamation mark. We then compared the results with those of the texts translated from one of the two languages into the other (L2). Thus, in Table 1, "Italian L2" texts are translated from Danish into Italian and "Danish L2" texts from Italian into Danish.

Language	Words	Sentences	Words /sentence
Italian L1	1,657,592	47,405	34.97
Danish L1	546,425	22,668	24.10
Italian L2	571,115	22,154	25.78
Danish L2	1,845,951	57,574	32.06

Table 1: Sentence length in L1 and L2 Europarl texts.

We chose Europarl as the empirical basis for a statistical count because it contains both parallel (L1 – L2) texts and comparable texts, i.e. L1 texts created in different languages but dealing with similar topics and produced in similar situations and genres for similar targets. Whereas parallel texts are clearly best suited for projects aimed e.g. at improving machine translation (such as the previously mentioned CDT) because they permit L1–L2 text alignment and evaluation, comparable texts are generally best suited as the empirical basis for descriptive, typological comparisons like the present one. In such cases, parallel texts are inappropriate because the "filter" of the translator and his/her translation strategies "get in the way", and L2 texts risk ending up with a text structure too similar to that of the L1. See McEnery and Wilson (2001) and Baroni and Bernardini (2006) for discussions in this regard.

As the upper part of Table 1 shows, there is a considerable difference in average sentence length between the Italian L1 and Danish L1 Europarl texts, a difference amounting to 10.86 words per sentence or 31.06 %. However, the lower part of Table 1 confirms the problem just mentioned regarding translated L2 texts. As far as sentence length goes, EU translators seem to stick very much to the structure of the L1 text: the Danish L2 texts (translated from Italian) are 24.82 % longer than the Danish L1 texts, while the Italian L2 texts (translated from Danish) are 35.64 % shorter compared to the Italian

[1] We are aware of the many reservations to be made when conducting linguistic measurements in this way, but subject to space limitations we cannot go into detail here. However, we feel that the statistical results cited in this section are convincing enough to be taken into account and used as a first indication of profound typological differences between the two languages analysed.

[2] We thank our colleague Daniel Hardt for his help in this matter.

L1 texts. When it comes to sentence length, these L2 texts are clearly influenced by the L1 structure.

2.2 Elementary Discourse Units

At this point we shall return to the concept of "informational density" and define a little more precisely its application in our project. In order to determine the purpose that the more numerous words in the Italian sentences serve, we then counted the number of Elementary Discourse Units (EDUs) textualised in each sentence, using Carlson and Marcu's (2001) classification. This can be a very time-consuming task, since no parser has been trained to do this convincingly, and we therefore randomly selected a limited part of the Europarl corpus consisting of 7,500 words in each language. We confined ourselves to texts of 200-600 words, and we ended up with a subcorpus in each language consisting of 25 texts of an average length of 300 words each. All texts were manually checked with regard to text type (argumentative), speaker (a certain number of different speakers were required), and date (so that not all text were speeches from the same period).

We discovered a very clear tendency towards a higher number of EDUs in the Italian sentences than in the Danish ones. A statistical count showed that 27.3 % of the Italian sentences contained five or more EDUs. By comparison, only 9.8 % of the Danish sentences contained five or more EDUs.

We also discovered considerable differences in the number of coordinate vs. subordinate clauses. Finite coordinate clauses amounted to 27.2 % of all clauses in the Danish texts, but only to 17.9 % in the Italian texts. Thus, 82.1 % of the Italian clauses were subordinate as opposed to 72.8 % of the Danish clauses. This may not seem a huge discrepancy, but if we examine in detail the distribution of the subordinate clauses, we encounter considerable differences, cf. Table 2:

	With connectives	Relative clauses	Attribution	Subordinate non-finite clauses
IT	22.4 %	40.3 %	13.1 %	24.2 %
DA	25.8 %	40.3 %	22.5 %	11.4 %

Table 2: Distribution of EDUs in subordinate clauses in a Europarl subcorpus

The use of connectives (or "discourse cues" in the RST terminology) and the frequency of relative clauses are more or less equal in the two languages, whereas Danish seems to use attribution more often. In our opinion, this difference should be seen not just as a particular linguistic tendency among Danish parliamentarians, but also as a stylistic feature used to add particular pragmatic values to the argument put forward, a point we shall elaborate in the full version of this paper.

However, the most interesting difference lies in the distribution of non-finite clauses. As Table 2 shows, these occur more than twice as often in the Italian texts as in the Danish ones. Furthermore (not shown in Table 2), Italian uses the whole range of non-finite verb forms (gerund, participles, infinitives and normalisations) much more regularly, whereas Danish mostly confines itself to the use of infinitives (the gerund does not exist in Danish).

3. Text complexity

The differences in sentence length seen in Table 1 also have an impact on the distribution of EDUs. Many EDUs correspond to propositions, and what may be textualised as one multi-propositional sentence in a Romance language may very well correspond to two or more sentences in Scandinavian. In a sequence of propositional EDUs, P1 + P2, such as the following:

P1: *arrive* (John, in town); P2: *go* (John, home)

P1 can be textualised in different ways (possibly with added adjuncts or other linguistic material), as shown in the "Deverbalisation Scale" in Table 3[3]:

	P1 textualised as	Textualisation P1 + P2
↓	a. an independent sentence	*John arrived late in town.* He went straight home.
	b. a main clause, part of sentence	*John arrived late in town* and he went straight home.
	c. a subordinate finite clause	*Since John arrived late in town,* he went straight home.
	d. a subordinate non-finite clause	*Having arrived late in town,* John went straight home.
	e. a nominalisation	*Upon his arrival in town,* John went straight home

Table 3: Examples of textualisation of EDUs.

[3] The scale is based on Hopper and Thompson (1984), Lehmann (1988), and Korzen (1998; 2007; 2009).

The deverbalisation of P1 increases from (a/b) to (e) together with its integration and absorption into the matrix clause. Whereas the finite verb in a main clause, such as (a/b), has its full (language specific) range of grammatico-semantic values and the clause its full range of pragmatic-illocutionary possibilities, these values are gradually reduced or lost in the textualisations further down the scale. The verb in the subordinate finite clause (c) loses its independent tense, mood and illocution; these values will be determined and/or expressed by the matrix clause. The non-finite verb in (d) loses all temporal, modal, and aspectual values and cannot render explicit its subject (see however note 4), and the nominalisation (e) is completely integrated in the matrix clause as a second order entity; its valency complements (here *his*) are syntactically reduced to secondary positions or simply left out.

The further down the scale a proposition is textualised, the fewer grammatico-semantic and pragmatic features are expressed by the verb, i.e. the more the proposition is "deverbalised", and the more it is semantically and rhetorically subordinated and incorporated into the matrix clause. In the case of non-finite and nominalised verbs, (d/e), features such as subject, tense, mood, aspect, and illocution are entirely interpreted on the basis of the matrix clause[4]. Therefore, a non-finite or nominalised structure is entirely pragmatically and semantically dependent on the matrix clause, and such structures express a particularly strong rhetorical backgrounding (or explicit satellite status) of the proposition in question. Furthermore, the lack of subject generally entails an inherent topic continuity (a topic shift typically requires a finite verb with an explicit subject), which means that the situation or event in question is evaluated and interpreted as related and less important to the on-going topic than the situation or event of the matrix clause, textualised with a finite predicate.

Cross-linguistic surveys show that textualisation at the levels (d/e) is much more frequent in the Romance languages than in the Scandinavian ones which show a very clear predilection for finite verbs and textualisation at the levels (a/b/c). These tendencies are not limited to particular text types or genres, such as the (generally argumentative) Europarl texts. Table 4 indicates the percentage of propositions textualised with finite, non-finite, and nominalised verb forms in a number of comparable texts belonging to five different text types and genres. The numbers clearly indicate statistically significant differences between Italian and Danish text structure regarding finite and non-finite verb frequency, independently of text type or genre.

		Verb forms (%)		
		Finite	**Non-finite**	**Nominalised**
a. Legal texts	**IT**	43.9	24.2	31.9
	DA	56.4	10.2	33.4
b. Technical texts	**IT**	47.5	26.8	25.9
	DA	80.7	9.5	9.9
c. News-groups	**IT**	61.1	23.1	15.8
	DA	75.8	11.5	12.7
d. Websites	**IT**	54	27	19
	DA	84	8	8
e. Written narratives	**IT**	52.8	44.2	3.0
	DA	88.0	12.0	0.01
f. Oral narratives	**IT**	72.8	27.1	0.1
	DA	93.6	6.4	0

Table 4: Verb forms in different text types[5]

As stated above, non-finite and nominalised structures explicitly express the satellite status of the proposition in question. Generally – but not necessarily – this is true also of subordinate adverbial clauses, such as (c) in Table 3. On the other hand, the structures in (a/b) of Table 3 are in themselves ambiguous as to mono- or multinuclear interpretation. However, as is well known, the structure in (b), the syndetic coordination with the connective *and* (and cross-linguistic counterparts), often contains a P1 with satellite status, in Table 3 expressing the *cause* of P2. We shall elaborate also on this issue in the full version of our paper[6].

4. Perspectives for translation

The differences described above entail a generally higher

[4] We here ignore the subject of the so-called "absolute constructions" consisting of a participle or gerund + a subject different from the subject of the main verb, e.g. ***Morto il padre***, *Luca partì per Roma* – ***The father [having] died***, *Luca left for Rome,* as well as the "accusative with infinitive" constructions (*Ho visto* ***Luca arrivare*** – *I saw* ***Luca arrive***). In nominalised verb forms the subject may appear as a secondary valency complement, e.g. *L'arrivo di* ***Luca*** – ***Luca's*** *arrival*.

[5] Precise references will appear in the full version of our paper.

[6] Important cross-linguistic studies on *and* and counterparts are found e.g. in Ramm and Fabricius-Hansen (2005), Behrens and Fabricius-Hansen (2010) and Skytte (2000: 652-660).

structural complexity in Italian (and Romance in general) than in Danish (and Scandinavian in general). Romance sentences tend to be longer and to include more propositions, of which a higher number is backgrounded by means of non-finite and nominalised predicates. This results in a multi-layered and hierarchical information structure, characterised by a high degree of topic continuity, in which the various events are evaluated with respect to their importance to the on-going topic.

On the other hand, Scandinavian text structure tends to be more informationally linear and characterised by a higher degree of topic shifts. Each sentence holds fewer EDUs, and different events tend to be textualised more chronologically one after the other and with finite verb forms that permit subject/topic changes.

The results of our study can be transferred into three main rules concerning translations from a Romance to a Scandinavian language or vice versa. The rules regard:

- the number of EDUs per sentence: *ceteris paribus,* there are more EDUs and a higher informational density in Romance than in Scandinavian sentences;
- the textualisation of rhetorical structure: there is a higher tendency in Romance than in Scandinavian to distinguish morpho-syntactically between rhetorical nuclei and satellites;
- the textualisation of rhetorical satellites: there is a tendency to textualise satellites at lower levels of the deverbalisation scale (cf. Table 3) in Romance than in Scandinavian.

Naturally, also phenomena such as e.g. the linguistic register and diamesic dimension (e.g. written vs. spoken text) come into play. The higher the register, the more distinct the mentioned cross-linguistic differences. Oral Italian textualisation and some web variants (such as newgroups, see Table 4) are characterised by a certain structural levelling and are therefore closer to typical Danish textualisation.

5. Conclusion

It is well known that a good translation does not (generally, at least) follow the source text word for word. But especially between language families, a good translation does not often follow the source text sentence for sentence, either. Profound typological differences such as those regarding informational density and text complexity must be taken into account, and contrastive studies on discourse structure provide necessary and highly useful linguistic insights for human as well as machine translators.

The results of our study – presented above and in the full version of our paper – will hopefully provide us with more precise and detailed knowledge of typological differences between Romance and Scandinavian discourse structure, differences which are of importance also for syntax (e.g. in the choice of subject type and voice) and for anaphora (e.g. null-forms vs. pronominal forms), phenomena that we will develop in future work.

6. Acknowledgements

This work was supported by a grant from The Danish Council for Independent Research │ Humanities (FKK). We thank the two anonymous reviewers for their useful comments and references.

7. References

F. Alves, A. Pagano, S. Neumann, E. Steiner, and S. Hansen-Schirra (2010): Translation Units and Grammatical Shifts. Towards an Integration of Product- and Process-based Translation Research. In G.M. Shreve and E. Angelone (Eds). Translation and Cognition. John Benjamins, Amsterdam/Philadelphia, pp. 109–142.

N. Asher and L. Vieu (2005): Subordinating and Coordinating Discourse Relations. Lingua 115, pp. 591–610.

M. Baroni and S. Bernardini (2006): A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text. In Literary and Linguistic Computing, 21,3, pp. 259–274.

B. Behrens and C. Fabricius-Hansen (2010). The Relation Accompanying Circumstance Across Languages: Conflict between Linguistic Expression and Discourse Subordination? In D. Shu and K. Turner (eds.). Contrasting Meaning in Languages of the East and West. Contemporary Studies in Descriptive Linguistics, 14. Oxford et al.: Peter Lang, pp. 531–552.

B. Behrens, K. Solfjeld and C. Fabricius-Hansen (2010): The Relation Accompanying Circumstance Across Languages: Conflict between Linguistic Expression and Discourse Subordination? In D. Shu and K. Turner (Eds.). Contrasting Meaning in Languages of the East and West. Contemporary Studies in Descriptive Linguistics, 14. Oxford et al.: Peter Lang, pp. 531–552.

M. Buch-Kromann et al. (2010): The Inventory of Linguistic Relations used in the Copenhagen Dependency Treebanks. Copenhagen Business School, http://copenhagen-dependency-treebank.googlecode.com/svn/trunk/manual/cdt-manual.pdf.

L. Carlson and D. Marcu (2001): Discourse Tagging Reference Manual. ISI Technical Report, ISI-TR-545.

W.L. Chafe (1976): Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Point of View. In: Li, Charles N. (Ed.). Subject and Topic. Academic Press, New York/San Francisco/London, pp. 25–55.

H.H. Clark and S.E. Haviland (1977): Comprehension and the Given-new Contract. In Discourse Production and Comprehension, R.O. Freedle (Ed.), Hillsdale, NJ: Erlbau, pp. 1–40.

C. Fabricius-Hansen (1996): Informational Density: a Problem for Translation and Translation Theory. Linguistics 34, pp. 521–565.

C. Fabricius-Hansen (1999): Information Packaging and Translation. Aspects of Translational Sentence Splitting (German - English/Norwegian). In *Sprach-spezifische Aspekte der Informationsverteilung*, M. Doherty (Ed.). Akademie-Verlag, Berlin, pp. 175–213.

H. van Halteren (2008): Source Language Markers in EUROPARL Translations. In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008). Manchester, August 2008, pp. 937–944.

S. Hansen-Schirra, S. Neumann and E. Steiner (2007): Cohesive Explicitness and Explicitation in an English-German Translation Corpus. Languages in Contrast 7(2), pp. 241–265.

P. J. Hopper and S. A. Thompson (1984): The Discourse Basis for Lexical Categories in Universal Grammar. Language, 60(4), pp. 703–752.

P. Koehn (2005): Europarl: A Parallel Corpus for Statistical Machine Translation. MT Summit.

I. Korzen (1998): On the Grammaticalisation of Rhetorical Satellites. A Comparative Study on Italian and Danish. In I. Korzen and M. Herslund (Eds.). Clause Combining and Text Structure. Studies in Language, 22, Copenhagen, pp. 65–86.

I. Korzen (2007): Linguistic Typology, Text Structure and Appositions. In I. Korzen, M. Lambert, and H. Vassiliadou. Langues d'Europe, l'Europe des langues. Croisement Linguistiques. Scolia, 22, pp. 21–42.

I. Korzen (2009): Struttura testuale e anafora evolutiva: tipologia romanza e tipologia germanica. In I. Korzen and C. Lavinio (Eds). Lingue, culture e testi istituzionali. Firenze: Franco Cesati, pp. 33–60.

C. Lehmann (1988): Towards a Typology of Clause Linkage. In J. Haiman and S. A. Thompson (Eds.). Clause Combining in Grammar and Discourse. John Benjamins, Amsterdam/Philadelphia, pp. 181–225.

W.C. Mann, C. Matthiessen and S.A. Thompson (1992): Rhetorical Structure Theory and Text Analysis. In W.C. Mann and S.A. Thompson (Eds). Discourse Description. Diverse Linguistic Analyses of a Fund-raising text. John Benjamins, Amsterdam/Philadelphia, pp. 39–78.

W.C. Mann and S.A. Thompson (1987): Rhetorical Structure Theory. A Theory of Text Organization. ISI, Los Angeles, CA, ISI/RS-87-190, pp. 1–81.

C. Matthiessen and S.A. Thompson (1988): The Structure of Discourse and 'Subordination'. In J. Haiman and S.A. Thompson (Eds). Clause Combining in Grammar and Discourse. John Benjamins, Amsterdam/Philadelphia, pp. 275–329.

T. McEnery and A. Wilson (2001): Corpus Linguistics: an Introduction. 2nd Edition. Edinburgh University Press, Edinburgh.

E.F. Prince (1984): Topicalization and Left-dislocation: a Functional Analysis. In Discourses in Reading and Linguistics, Sheila J. White and Virginia Teller (Eds.). Annals of the New York Academy of Sciences 433, Academy of Sciences, New York, pp. 213–225.

W. Ramm and C. Fabricius-Hansen (2005): Coordination and Discourse-structural Salience from a Cross-linguistic Perspective. SPRIKreports 30.

G.. Skytte (2000): Konnexion og diskursmarkering. In G. Skytte and I. Korzen. Italiensk–dansk sprogbrug i komparativt perspektiv. Reference, konnexion og diskursmarkering, Samfundslitteratur, Copenhagen, pp. 621–793.

E. Vallduví and E. Engdahl (1996): The Linguistic Realization of Information Packaging. Linguistics 34, pp. 459–519.

B. Webber, M. Egg, and V. Kordoni (2010): Discourse Structure and Language Technology. Natural Language Engineering, 1(1), pp. 1–49.

GSCL 2011 Workshop “Contrastive Linguistics – Translation Studies – Machine Translation – what can we learn from each other?”

An analysis of translational complexity in two text types.

Martha Thunes, University of Bergen.
martha.thunes@lle.uib.no

Key words: automatisation of translation, English-Norwegian parallel text, translational complexity, text types.

Abstract

This paper is based on the study presented in Thunes (2011), where a selection of English-Norwegian parallel texts have been analysed in order to discuss two primary research questions: firstly, to what extent is it possible to automatise, or compute, the actual translation relation found in the investigated parallel texts, and, secondly, is there a difference in the degree of translational complexity between the two text types, law and fiction, included in the empirical material?

By *automatisation* I here understand the generation of translations with no human intervention, and I assume an approach to machine translation based on linguistic information. In the analysed texts the translations have been produced manually; this is not a study of output produced by machine translation systems, and the automatisation issue is not discussed with reference to any particular translation algorithm or system architecture. Rather, it is related to the assumption that there is a translational relation between the inventories of simple and complex linguistic signs in two languages which is predictable, and hence computable, from information about source and target language systems, and about how the systems correspond. Thus, computable translations are *linguistically predictable*, i.e. predictable from the linguistic information coded in the source text, together with given, general information about the two languages and their interrelations. Further, non-computable translations are correspondences where it is not possible to predict the target expression from the information encoded in the source expression, together with given, general information about SL and TL and their interrelations. Non-computable translations require access to additional information sources, such as various kinds of general or task-specific extra-linguistic information, or task-specific linguistic information from the context surrounding the source expression.

In order to answer the research questions, a measurement of translational complexity is applied to the analysed texts. The degree of *translational complexity* in a given translation task is understood as a factor determined by the types and amounts of information needed to solve the task, as well as by the accessibility of these information sources, and the effort required when they are processed.

For the purpose of measuring the complexity of the relation between a source text unit and its target correspondent, I apply a set of four correspondence types, organised in a hierarchy reflecting divisions between different linguistic levels, along with a gradual increase in the degree of translational complexity. In type 1, the least complex type, the corresponding strings are pragmatically, semantically, and syntactically equivalent, down to the level of the sequence of word forms. In type 2 correspondences, source and target string are pragmatically and semantically

equivalent, and equivalent with respect to syntactic functions, but there is at least one mismatch in the sequence of constituents or in the use of grammatical form words. Within type 3, source and target string are pragmatically and semantically equivalent, but there is at least one structural difference violating syntactic functional equivalence between the strings. In type 4, there is at least one linguistically non-predictable, semantic discrepancy between source and target string, and pragmatic equivalence may, or may not, hold. Thus, the type hierarchy is characterised by an increase with respect to linguistic divergence between source and target string, and by an increase in the need for information and in the amount of effort required to translate, i.e. an increase in the degree of translational complexity. Correspondences of types 1–3 constitute the domain of linguistically predictable, or computable, translations, whereas type 4 correspondences belong to the non-predictable, or non-computable, domain, where semantic equivalence is not fulfilled.

This study applies a strictly product-oriented approach to complexity in translation. The four types of translational correspondences should not be seen as translation methods or strategies, but as descriptions of correspondence relations between given source text units and their existing translations. The empirical analysis of translational correspondences does not aim to study what kinds of knowledge a translator has actually used in order to produce a chosen target expression. Rather, it focusses on the kinds of information about source text expressions that are needed in order to produce the translations.

The correspondence type hierarchy can be seen as a fairly general classification model for translational correspondences. Its main principles were originally defined by Dyvik (1993), and further articulated in Thunes (1998). The approach chosen for the present study is an adapted version of the classification model defined by Thunes (1998). The model is also used as a framework for contrastive language analysis in the studies presented by Hasselgård (1996), Tucunduva (2007), Silva (2008), and Azevedo (in progress).

In the present contribution, the empirical method involves extracting translationally corresponding strings from parallel texts, and assigning one of the types defined by the correspondence hierarchy to each recorded string pair. The finite clause is chosen as the primary unit of analysis, and the main syntactic types among the recorded data are matrix sentences, finite subclauses, and lexical phrases with finite clause(s) as syntactic complement. Since syntactically dependent constructions like finite subclauses occur as translational units, the data include nested correspondences where a superordinate string pair contains one or more embedded string pairs. The assignment of correspondence type to string pairs is an elimination procedure where we start by testing each correspondence for the lowest type and then move upwards in the hierarchy if the test fails. The analysis is thus an evaluation of the degree to which linguistic matching relations hold in each string pair. In cases of nested string pairs, embedded units are treated as opaque items, and the classification of a superordinate correspondence is done independently of the degree of complexity in embedded string pairs. Otherwise, it is a general principle that a string pair is assigned the correspondence type of its most complex non-opaque subpart.

The analysis is applied to running text, omitting no parts of it. Thus, the distribution of the four types of translational correspondence within a set of data provides a measurement of the degree of translational complexity in the parallel texts that the data are extracted from. The extraction and classification of string pairs is done manually as it requires a bilingually competent human analyst. The recorded data cover about 68 000 words, and are compiled from six different text pairs: two of them are law texts; the remaining four are fiction texts. Comparable amounts of text are included for each text type, and both directions of translation are covered.

Since the scope of the investigation is limited, the results do not provide a sufficient basis for generalisations about the degree of translational complexity in the chosen text types and in the language pair English-Norwegian. Concerning the automatisation issue, the complexity measurement across the entire collection of data shows that, in terms of string lengths, as little as 44,8% of all recorded string pairs are classified as computable translational correspondences, i.e. as type 1, 2, or 3, and non-computable string pairs of type 4 constitute a majority (55,2%) of the compiled data. As regards the text type issue, the proportion of computable correspondences is on average 50,2% in the law data, and 39,6% in fiction.

In order to discuss whether it would be fruitful to apply automatic translation to the selected texts, I have considered the workload potentially involved in correcting assumed machine output, and in this respect the difference in restrictedness between the two text types is relevant: law text is strongly norm-governed in a way that fiction text is not. Among the recorded data, I have analysed a set of phenomena that have been identified as recurrent semantic deviations between translationally corresponding units, and this shows that within the non-computable correspondences, the frequency of cases exhibiting only one minimal semantic deviation between source and target string is considerably higher among the data extracted from the law texts than among those recorded from fiction. Such cases can be regarded as minimally non-computable string pairs. Among the law data, as much as 45,7% of the correspondences classified as type 4 are minimally non-computable string pairs, whereas among the fiction data, only 10,5% of the compiled type 4 correspondences are minimal ones. In minimally non-computable correspondences, I assume that only a small effort would be required in order to revise an automatically generated target expression according to the standard of manual translation.

For this reason I tentatively regard the investigated pairs of law texts as representing a text type where tools for automatic translation may be helpful, if the effort required by post-editing is smaller than that of manual translation. This is possibly the case in one of the law text pairs, where 60,9% of the data involve computable translation tasks. In the other pair of law texts the corresponding figure is merely 38,8%, and the potential helpfulness of automatisation would be even more strongly determined by the edit cost. That text might be a task for computer-aided translation, rather than for MT. As regards the investigated fiction texts, it appears likely that post-editing of automatically generated translations would be laborious and not cost effective, even in the case of one text pair showing a relatively low degree of translational complexity. In the analysed pairs of fiction texts, there is a clear tendency that non-computable correspondences exhibit several semantic deviations between the corresponding strings. Hence, I expect that the workload involved in correcting potential machine output would be heavy, and I agree with the common view that the translation of fiction is not a task for MT.

This study is intended to be of relevance to rule-based MT since the chosen analytical framework relies on assumptions about how translations can be computed on the basis of formal descriptions of source and target language systems and their interrelations. However, I assume that the general issue of computability underlying this approach likewise applies to statistical machine translation, which is also dependent on the accessibility of relevant and sufficient information in order to predict correct target expressions from available translational correspondences.

In my view, the framework applied in this study could be used as a diagnostic tool for the feasibility of machine translation in relation to specific text types. That is, by applying the method to limited selections of parallel texts of the same type, it would be possible to estimate to what extent the target text could be generated automatically. If the proportion of assumed computable correspondences would exceed a chosen threshold, it might be worthwhile to tune an MT system for the given language pair to the text type in question. Moreover, in order to estimate the editing distance between potential machine output and a given target text norm, it would be interesting to identify the proportion of minimal type 4 correspondences in a given body of parallel texts. Thus, it would be fruitful to extend the classification model by integrating a fifth correspondence type to be assigned to minimally non-computable string pairs.

References:

Azevedo, Flávia. In progress. *Investigating the problem of codifying linguistic knowledge in two translations of Shakespeare's sonnets: a corpus-based study.* Doctoral dissertation. Federal University of Santa Catarina, Florianópolis.

Aijmer, Karin, Bengt Altenberg, and Mats Johansson (eds). 1996. *Languages in Contrast. Papers from a Symposium on Text-based Cross-linguistic Studies, Lund 4–5 March 1994. Lund Studies in English* 88. Lund: Lund University Press.

Dyvik, Helge. 1993. Text Pair Mapper. Unpublished manuscript. University of Bergen.

Hasselgård, Hilde. 1996. Some methodological issues in a contrastive study of word order in English and Norwegian. In: Aijmer et al. (eds), 1996, 113–126.

Johansson, Stig and Signe Oksefjell (eds). 1998. *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies*. *Language and Computers: Studies in Practical Linguistics* 24. Amsterdam and Atlanta, GA: Rodopi.

Silva, Norma Andrade da. 2008. *Análise da tradução do item lexical* evidence *para o português com base em um corpus jurídico.* Master's thesis. Federal University of Santa Catarina, Florianópolis.

Thunes, Martha. 1998. Classifying translational correspondences. In: Johansson and Oksefjell (eds), 1998, 25–50.

Thunes, Martha. 2011. *Complexity in Translation. An English-Norwegian Study of Two Text Types.* Doctoral dissertation. University of Bergen.

Tucunduva, Camila de Andrade. 2007. *Translating completeness: a corpus-based approach.* Master's thesis. Federal University of Santa Catarina, Florianópolis.

Web-based contrastive comparison of Age-Related Temporal Phrases in Spanish and French

Sofía N. Galicia-Haro[1] and Alexander Gelbukh[2]
[1]Faculty of Sciences UNAM ,Mexico
Depto. de Matemáticas, Ciudad Universitaria
04510 México, D. F.
[2]Center for Computing Research, National Polytechnic Institute, Mexico
Juan de Dios Bátiz, esq. con Miguel Othón de Mendizábal
07738 México, D. F.
E-mail: [1]sngh@fciencias.unam.mx, [2]gelbukh@cic.ipn.mx; www.Gelbukh.com

Introduction

Some words or whole sequences of words in a text are temporal expressions: for example, *yesterday*, *Monday 12*, *two months*, *about a year and a half*; each refers to a certain period of time. Such words or sequences of words mainly share a noun or an adverb of time: *yesterday*, *month*, *year*. This causes a problem in automatically deciding whether a word or a sequence is a temporal expression. It is an important part of many natural language processing applications, such as question answering, machine translation, information retrieval, information extraction, text mining, etc., where robust handling of temporal expressions is necessary.

Automatic recognition of expressions of time was introduced in the Named Entity Recognition task of the Message Understanding Conferences[1] where temporal entities were tagged as "TIMEX." Since then, researchers have developed temporal annotation schemes; for example, [2] and [6] for English, [1] for French, and [7] for Spanish.

In this work, we analyzed Spanish temporal expressions that were not considered in those annotation guidelines. These phrases are recognized by an initial adverb: for example, *around*, *still*; they end with the noun of time, e.g. *year*, and they describe a person's age. For example: *aún a sus 50 años* "although he is 50 years old," *ahora a mis 23 años* "now I am 23 years old," *alrededor de los 55 años* "around 55 years old." These phrases are very interesting since the adverb reinforce the meaning of time.

We can observe the relation between the groups of words in the following Spanish examples:

1. *A sus 30 años Juan se comporta como niño*
2. *Aún a sus 30 años Juan se comporta como niño*
3. *Hoy a sus 30 años Juan se comporta como niño*

The sentences describe the same main fact: *John, who is 30 years old*, *behaves like a child*, but they tell us something else when we introduce a modifier (*aún* "still," *hoy* "today") in each one: they argue for different conclusions.

- Even at 30 years old, John behaves like a child ⇒ in spite of his age he behaves as if he were a child
- Today, at 30 years old, John behaves like a child ⇒ today he behaves like a child

The adverbs "even" and "today" make such conclusions obligatory and reinforce the meaning of time in different forms. Both adverbs are related to time duration; one strict reading refers to 24 hours and the other to a longer period of time, but they also imply a direct judgment on the perception of the speaker, on the behavior of the subject or both.

Owing to the constructions similarity in Spanish and French we supposed that these phrases of interest would be similar in both languages but we found that the French translation[2] "*encore à ses 38 années*" for the phrase *aún a sus 38 años* is not common in French.

In this work, we develop a web-based analysis carried out to compare such Spanish temporal expressions with age-related temporal phrases in French with the objective of determining appropriate annotations for marking up text and translations. First, we present the characteristics of the Spanish phrases and the method we applied to obtain the materials for the comparison.

[1] http://timexportal.wikidot.com/timexmuc6
[2] http://www.online-translator.com/Default.aspx/Text

Then we describe the French phrases obtained with the same method. Finally we present the comparison of such phrases and the application of the results to annotation and machine translation.

Age-Related Temporal Expressions in Spanish

Usually people's age is described by Spanish temporal expressions including the time noun *años* "years." They can be recognized in the following ways: ***5 años de edad*** (5 years old), *Jorge,* ***de 52 años****, entrenaba al Vitesse Arnhem* (Jorge, a 52-year-old, was training for the Vitesse Arnhem), *la niña* ***de 11 años*** (the 11-year-old girl), *falleció ayer* ***a la edad de 95 años*** (died yesterday at 95 years old)

There are, however, other temporal expressions that describe people's age: for example, *aún a sus 65 años*, lit "still at his 65 years", *de alrededor de 20 años*, lit. "of about 20 years." These temporal phrases denote a point in the timeline of a person; it could be a point in the timeline of the events related in the sentence or a point in a tangential timeline.

Material Acquisition (MA)

Texts and their exact translation, i.e. parallel corpora, are widely used mainly to train and evaluate Machine Translation systems. They are also useful in cross-linguistic analysis by looking for translations of a given construction into another language. We also intended to use parallel corpora to make a better comparison of age-related temporal phrases in Spanish and French and for this purpose we examined free texts collections of the European Commission. Many of them were from the law domain, however, and they had few examples of the expressions we were interested in. We decided then to use newspaper texts since they employ freer construction of sentences.

In previous work, a corpus-based analysis was carried out to determine the context of such Spanish temporal expressions for their automatic determination. Such a method allows the manual selection of examples representing what was considered to be a class: a different combination of an adverb and a preposition before the number of years and then the retrieval of web examples for that class.

The method consists of two steps. The first one is the application of a program to extract the sentences matching the following pattern:

AdvT−something−TimeN

where:

something – corresponds to a sequence of up to six words[3] without punctuation marks, verbs or conjunctions

TimeN – corresponds to *año*, *años* "year, years"

AdvT – adverbs of time, a collection of 51 elements from a dictionary[4]

We applied this step to a text collection compiled from a Mexican newspaper and from 27054 sentences we manually selected one arbitrary example representing a class, the five resulting classes corresponding to *aún a, aún con*, *actualmente de*, *alrededor de*, *ahora de*.

The second step was intended to obtain a more representative group of phrases since the newspaper text collection contained a subset of all possible temporal phrases expressing the age of people. We analyzed diverse methods to obtain a more representative group of phrases and we chose to look for examples on the Internet. This option allowed us to find phrases generated by native speakers more quickly, including the commoner collocations. We realize that searching the Internet has its drawbacks but we decided to do so on the basis that we did not know how the results were classified [4].

Many studies focused in having a corpus that modeled the whole language. To collect information for annotation and translation of the phrases we were interested in, however, we collected only a particular subset of language that corresponded to them. Thus, the research we report here refers to a collection that has been skewed by design.

[3] A larger number of words does not guarantee any relation between the AdvT and the TimeN

[4] DRAE, Real Academia Española. (1995): *Diccionario de la Real Academia Española*, 21 edición (CD-ROM), Espasa, Calpe.

The main idea of obtaining more examples from the Internet was based on obtaining a few examples from the newspaper texts (corresponding to the five classes mentioned above), simplifying them (eliminating determinants, adjectives, etc.) and searching for variants by including Google's asterisk facility [3]. For example: for the phrase *aún con sus jóvenes 48 años* the string when simplified becomes "*aún con año*" and the search is "*aún con * años*" using the Google search engine tool limited to the Spanish language where the asterisk substitutes for the eliminated words. Google returns hits where there is a string of words initiated by "*aún con*" and then a sequence of words, ending with "*años*" for example: *... y el bachillerato en Lleida,* ***aún con dieciséis años*** *entró a trabajar de chico ...*

The process was repeated several times until no new repeated phrases were obtained, determining the sequences of words that appeared with greater frequency. After this compilation of examples, we manually selected 18 cases: for example, *ahora a los* NUM *años*, *actualmente de unos* NUM *años*, where NUM treats numbers represented by digits or letters. We found that some of the 18 classes obtained from the Internet seem to preserve their meaning independently of the context and others require some form of words in context to denote the age of a person. The quantity of pages automatically obtained was limited to 50, i.e. to obtain 500 snippets. For each of the 18 classes we manually analyzed the number of examples that corresponded to a person's age.

Age-Related Temporal Expressions in French

We applied the previous method to 10234 sentences obtained from the Europarl corpus [5] that we called EuropAns. The sentences were retrieved in response to a query on the noun *ans* "years," since people's age is described by French temporal expressions including that noun.

To analyze the temporal phrases expressing age similarly to the Spanish phrases detailed above we applied the first step to the EuropAns. We obtained 257 sentences matching the AdvT−something−TimeN pattern where Adv corresponds to 57 elements. From them we manually selected five classes to process the MA second step by launching the following queries: *autour * ans*, *actuellement * ans*, *encore * ans*, *environ * ans*, *maintenant * ans.*

We applied the second step to access the Google search engine tuned to the French language. The results obtained from the Internet produced 3717 examples that corresponded to 24 cases: for example, *âgé(e,s,és) d'autour de* NUM *ans*, *encore maintenant à* NUM *ans.* For each of the 24 cases we manually analyzed the number of examples that corresponded to a person's age.

Comparison

We considered the results obtained from the Internet comparable since they derived from similarly qualified authors using similar registers. To compare the age phrases, we considered the following elements: (1) adverbs, (2) surface structure of the phrase between adverb and noun *years*, and (3) adjective placement. For this comparison we divided the results into four groups corresponding to the adverbs: *ahora/maintenant* "now," *alrededor/autour-environ* "around," *actualmente/actuellement* "at present," *aún/encore* "still." In general, we classified our compared examples as identical, different in some respects, and having no equivalent.

We matched the above-described groups in Spanish and French by considering first the adverb, then the age meaning and finally the percentage of phrases with age-related meaning. For example:

Type of phrase	# ex/ % age	Type of phrase	# ex/ % age
alrededor de los NUM años	355/84	environ âgé de NUM ans	1/100
		âgé(e,s,és) d'autour de NUM ans	2/100
		âgé(e,s,es) d'environ de NUM ans	37/84
		âgé(e,s,es) d'environ NUM ans	24/80
		âgé(e,s,es) environ de NUM ans	38/74
		autour de NUM ans	430/56
		environ de NUM ans	158/30
		autour de mes NUM ans	3/67
		environ NUM ans	285/19

The group of phrases initiated by the "around" adverb is the only French group including one case where the structure of the phrase between the adverb and the quantity of years is a prepositional phrase with a possessive adjective related to the person whose age is described (*autour de mes* NUM *ans*). This French case has no equivalent in Spanish in this group. The *environ* NUM *ans* case has no Spanish equivalent in this group in addition to being the least productive.

The first case is classified as different in some respects. These phrases have a *de* prepositional phrase in both languages governed by the adjective in four French counterparts. The other three phrases have the prepositional phrase governed by the adverb, being more similar to the Spanish phrase but less productive of age-and-person phrases.

Annotation and Translation

Generally, temporal expressions are annotated as entire constituents, typically noun phrases (e.g. 38 years). For example:

environ 10 secondes 'about 10 seconds' is annotated in [1] as:

<TIMEX3 tid="t1" type="DURATION" value="P10S" mod="APPROX">environ 10 secondes</TIMEX3>

We suggest annotating the phrases we analyzed here with the TLINK tag. TLINK is a temporal link that can represent the relation between two temporal elements: TIMEX3-TIMEX3. The phrases we considered for this annotation were the more productive ones with appropriate contexts extracted from the examples.

The results classified as different in some respects were the cases that we analyzed in detail to obtain alignment templates for their correct translation. These Spanish and French temporal expressions form a contiguous sequence given an appropriate context, and can be translated into the entire sequence as a multi-word unit. In this work we suggest translating them to take into account a different syntactic form and ordering function words' arguments.

We conclude that variety in the structure of temporal expressions necessitates analysis of different combinations of classes of words. Our study provides insights into the cross-lingual behavior of the temporal structure of age expressions particularly in the type AdvT−something−TimeN as realized in Spanish and French.

We made an empirical verification of a substantial degree of parallelism between the realization of such age expressions in Spanish and French but showed the differences in their frequency, structure and variety.

References

1. Bittar, André. ISO-TimeML Annotation Guidelines for French. Version 1.0 http://www.linguist.univ-paris-diderot.fr/~abittar/docs/FR-ISO-TimeML-Guidelines.pdf, (2010)
2. Ferro, Lisa, Laurie Gerber, Inderjeet Mani, Beth Sundheim and George Wilson. TIDES 2005 Standard for the Annotation of Temporal Expressions, MITRE Corporation (2005).
3. Gelbukh, A. and I.A. Bolshakov. Internet, a true friend of translator: the Google wildcard operator. International Journal of Translation 18(1–2), 41–48 Bahri Publications (2006)
4. Kilgarriff, A. Googleology is Bad Science. Computational Linguistics 33, 147–151 MIT Press (2007)
5. Salmon-Alt, Susanne, Eckhard Bick, Laurent Romary and Jean-Marie Pierrel. La FREEBANK: Vers une base libre de corpus annotés. In: Proceedings of TALN 2004. Fes, Morocco. http://corp.hum.sdu.dk/tgrepeye_fr.html
6. Saurí, R., J. Littman, B. Knippen, R. Gaizauskas, A. Setzer and J. Pustejovsky. TimeML Annotation Guidelines Version 1.2.1 (2006) http://www.timeml.org/site/publications/timeMLdocs/annguide_1.2.1.pdf
7. Saurí, Roser, Estela Saquete and James Pustejovsky. Annotating Time Expressions in Spanish. TimeML Annotation Guidelines. Version TempEval-2010.

A Contrastive Study on Abstract Anaphors in German and English

Stefanie Dipper [1], Christine Rieger[2], Melanie Seiss[2] & Heike Zinsmeister[2]

Ruhr University Bochum[1] & Konstanz University[2]

Introduction Abstract anaphors denote anaphoric relations between some anaphoric expression and an antecedent that refers to an abstract object like an event or a fact. In the classical example by Byron (2002), the pronoun *it* (underlined in (1a)) refers to an *event*: the migration of penguins to Fiji. In the alternative sequence, (1b), the demonstrative pronoun *that* refers to the *fact* that penguins migrate to Fiji in the fall.

(1) a. Each Fall, penguins migrate to Fiji. It happens just before the eggs hatch.
b. Each Fall, penguins migrate to Fiji. That's why I'm going there next month.

The automatic resolution of abstract anaphors still poses a problem to language processing systems.

We pursue a contrastive, corpus-based approach to investigate the properties that characterize different instantiations of abstract anaphora in English and German. In the long run, we envisage to derive features from the corpus annotation that will serve us to tackle the automatic resolution of abstract anaphors. In this paper we investigate what kind of anaphoric elements are employed in the two languages to refer to abstract objects. The range of possible realizations includes pronouns, lexical NPs (e.g. *this issue, this situation*, etc.), adverbials (e.g. *likewise*).

We present results of a comparative corpus study on the realization of abstract anaphora in a parallel bi-directional corpus of English and German. Besides comparing the cross-linguistic realizations, we also look into the differences between original text and translated text in both languages.

Most annotation projects that analyze abstract anaphora restrict themselves to pronominal markables (e.g. Byron (2003), Hedberg et al. (2007), Müller (2007), Dipper and Zinsmeister (2011)); some also annotate full NP markables (e.g. Vieira et al. (2002), Pradhan et al. (2007), Poesio and Artstein (2008)). Multilingual corpora have been annotated in Recasens (2008), Navarretta and Olsen (2008), and Vieira et al. (2002). A recent overview of projects annotating abstract anaphora is provided by Dipper and Zinsmeister (2010).

Annotation of parallel texts has been performed in Vieira et al. (2002). The present work deals with the annotation of the full range of abstract anaphors (including full NPs and anaphoric adverbs) in a parallel corpus.

Corpus For our study, we extracted about 100 German and English medium sized turns each (contributions by German and English speakers; average length about 20 sentences), along with their sentence-aligned translations, from the Europarl Corpus (Release v3, 1996–2006, Koehn (2005)). For cross-lingual annotation of the German and English turns, we used two MMAX2 annotation windows, put side by side on the screen.

We started out with a well-defined set of markables in the original language and collected all variants of translations on the side of the "target" language (the translation of the original language). In the first round of annotation, we chose original texts from German, because in German there is—in contrast to English—a pronoun that is unambiguously used as an abstract anaphor: the uninflected singular demonstrative pronoun *dies* ('this'). In addition to this, we defined as markables the (ambiguous) demonstrative pronoun *das* ('that') and the (ambiguous) third person neuter pronoun *es* ('it'). The target language was English.

For the second round of annotation we considered the reversed translation direction: English original texts and their German translations. We extended our set of markables and included the adverbs *as, so* and *likewise*, because these adverbs frequently served as translations of German anaphors in the first round. We will apply this method of bootstrapping back and forth to extend the set of markables iteratively. This approach allows for a fast and efficient way of extracting anaphors in both languages.

Results The German part of the corpus features 223 abstract anaphors—203 of which could be aligned with English text instances. On average, we identified 2.37 abstract anaphors per turn (with the basic set of markables).

The English part of our corpus contains 77 turns. It features 234 abstract anaphors. This corresponds to 3.03 abstract anaphors per turn (with the extended set of markables).

We used our annotations to test the hypothesis that English avoids the use of pronominal abstract anaphors. The results from the German-to-English ('DE-to-EN') annotations seem to support this hypothesis. 35% of German pronominal anaphors (71 out of 203) were not translated as pronouns in English.

We identified the following main strategies to avoid pronominal anaphors in the translation of German to English:

- there is no corresponding material, e.g. a different verb or a different argument frame is employed, see Ex. (2)[1]
- use of full NPs rather than pronouns (*all these things, these measures, this objective, this situation, this thread* ...).
- use of adverbials or conjunctions (*likewise, so, as*)

[1] In the examples, the lines prefixed with "DE" contain the German original text, the "EN" lines the official English translation, and the "DE-LIT" lines a literal translation of (parts of) the German original.

(2) *DE*: Europa ist nie fertig! Aber das Projekt muss entschlossen, gemeinschaftsorientiert und visionär zur politischen Union weiterentwickelt werden. Wenn dies nicht geschieht, verlieren wir das Vertrauen der Bürger.
EN: Europe will never be finished, but we must press on with the project for political union with determination and vision, and on a Community basis. If we do not, the public will lose confidence in us.
DE-LIT: . . . If this does not happen, the public will lose confidence in us.

Following Passonneau (1989) and Navarretta (2008), we hypothesized that English prefers demonstrative pronouns to personal pronouns in abstract anaphora in comparison to other languages. Our findings are that both German demonstrative and personal pronouns tend to be translated as demonstratives in English, as in Ex. (3).

(3) *DE*: Sie selbst haben gesagt: Vertrauen ist herzustellen. Tun Sie es!
EN: You said yourself that trust had to be built up. Do that!
DE-LIT: Do it!

The German pronoun *es* 'it' is often not represented in the English translation. 43% of *es*-anaphors do not receive a pronominal translation vs. only 32% of the demonstrative anaphors are not translated into a pronoun. Furthermore, comparing the frequencies of anaphoric pronouns and selected anaphoric adverbs in the English turns, the annotations show that 73% of the instances are realized by demonstrative pronouns.

So far we discussed results from comparing original German texts ("GO") and their English translations ("ET"). To be able to really interpret the results of this contrastive analysis, it is important to show that there is no significant difference between the English translated texts (ET) and English original texts ("EO"). For the purpose of this abstract, we tested for significant differences at different levels of abstraction. At the coarse-grained level, we found no significant difference between the proportions of pronominal subjects and objects in the EO and ET text. Likewise, on a more fine-grained level, there was no significant difference between the proportions of specially marked constructions in the EO and ET texts, such as topicalization in Ex. (4)-EN. These findings seem to allow us to using the translated texts in comparing German and English usage of abstract anaphors.

(4) *DE*: Wir können es nicht ändern.
EN: That is something we cannot change.
DE-LIT: We cannot change it.

Conclusion Although the main target of our research is to detect cross-linguistic and contrastive features for automated anaphora resolution, we believe that these features are also important for effective, high-standard machine translation. For both applications it is necessary to consider features in detail, such as paying attention to grammatical function and syntactic position.

References

Donna K. Byron. Resolving pronominal reference to abstract entities. In *Proceedings of the ACL-02 conference*, pages 80–87, 2002.

Donna K. Byron. Annotation of pronouns and their antecedents: A comparison of two domains, 2003. Technical Report, University of Rochester.

Stefanie Dipper and Heike Zinsmeister. Towards a standard for annotating abstract anaphora. In *Proceedings of the LREC 2010 workhop on Language Resource and Language Technology Standards*, pages 54–59, Valletta, Malta, 2010.

Stefanie Dipper and Heike Zinsmeister. Annotating abstract anaphora. *Language Resources and Evaluation*, Online First, 2011. doi: DOI 10.1007/s10579-011-9160-1.

Nancy Hedberg, Jeanette K. Gundel, and Ron Zacharski. Directly and indirectly anaphoric demonstrative and personal pronouns in newspaper articles. In *Proceedings of DAARC-2007*, pages 31–36, 2007.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit*, 2005.

Christoph Müller. Resolving *it*, *this*, and *that* in unrestricted multi-party dialog. In *Proceedings of ACL-07 conference*, pages 816–823, 2007. URL http://www.aclweb.org/anthology/P07-1103.

Costanza Navarretta. Pronominal types and abstract reference in the Danish and Italian DAD corpora. In *Proceedings of the Second Workshop on Anaphora Resolution*, pages 63–71, 2008.

Costanza Navarretta and Sussi Olsen. Annotating abstract pronominal anaphora in the DAD project. In *Proceedings of LREC-08*, 2008.

Rebecca J. Passonneau. Getting at discourse referents. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, 1989.

Massimo Poesio and Ron Artstein. Anaphoric annotation in the ARRAU corpus. In *Proceedings of LREC-08*, 2008.

Sameer Pradhan, Lance Ramshaw, Ralph Weischedel, Jessica MacBride, and Linnea Micciulla. Unrestricted coreference: Identifying entities and events in OntoNotes. In *Proceedings of the IEEE-ICSC*, 2007.

Marta Recasens. Discourse deixis and coreference: Evidence from AnCora. In *Proceedings of the Second Workshop on Anaphora Resolution*, pages 73–82, 2008.

Renata Vieira, Susanne Salmon-Alt, and Caroline Gasperin. Coreference and anaphoric relations of demonstrative noun phrases in a multilingual corpus. In *Proceedings of DAARC-2002*, 2002.

A Corpus-based Contrastive Analysis for Defining Minimal Semantics of Inter-sentential Dependencies for Machine Translation

Thomas Meyer, Andrei Popescu-Belis, Jeevanthi Liyanapathirana
Idiap Research Institute, Martigny, Switzerland
Bruno Cartoni
University of Geneva, Switzerland

Abstract

Inter-sentential dependencies such as discourse connectives or pronouns have an impact on the translation of these items. These dependencies have classically been analyzed within complex theoretical frameworks, often monolingual ones, and the resulting fine-grained descriptions, although relevant to translation, are likely beyond reach of statistical machine translation systems. Instead, we propose an approach to search for a minimal, feature-based characterization of translation divergencies due to inter-sentential dependencies, in the case of discourse connectives and pronouns, based on contrastive analyses performed on the Europarl corpus. In addition, we show how to automatically assign labels to connectives and pronouns, and how to use them for statistical machine translation.

1. The Need for Inter-sentential Information in Machine Translation

Long-range dependencies are a well known challenge for machine translation (MT) systems, especially for statistical ones. The correct translation of lexical items such as pronouns often depends on the correct identification of their antecedent. Similarly, the correct translation of multi-functional discourse connectives depends on the correct identification of the rhetorical relation which they convey between two clauses. However, especially when translating between closely related languages, the full disambiguation of such lexical items is sometimes unnecessary for a correct translation. The question that arises is thus how to find the most suitable level of representation for such dependencies, as a trade-off between linguistic accuracy and computational tractability, with the direct aim of improving MT output.

This paper presents a method for finding the minimal semantic/discourse information that must be assigned to two types of lexical items, namely connectives and pronouns, in order to avoid translation mistakes by statistical MT systems. The method starts from contrastive analyses of a frequently used parallel corpus, Europarl (Koehn, 2005), in order to define and annotate the minimal semantic/discourse information necessary for MT. The paper first describes our analyses and manual annotation methods for disambiguating connectives (Section 2.1) and pronouns (Section 2.2), in the context of English/French MT. Section 3 outlines methods for automatically performing these disambiguation tasks, while Section 4 explains how the automatically labeled linguistic items can be integrated into a statistical MT system. Section 5 concludes the paper and outlines future work.

2. Contrastive Analysis of Two Types of Inter-sentential Dependencies

2.1 Discourse Connectives

Discourse connectives are generally considered as indicators of discourse structure, relating two sentences or propositions and making explicit the rhetorical relation between them. Explicit discourse connectives such as *because, but, however, since, while*, etc., are frequent lexical items and are used to mark rhetorical relations such as *Cause* or *Contrast* between units of discourse. Several theoretical frameworks have been proposed for connectives (mainly starting from English ones), such as the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), or the Segmented Discourse Representation Theory (SDRT) (Asher, 1993). In such theories, more than one hundred possible rhetorical relations have been identified, and complex semantic and logical representations have been used to characterize discourse structure. In a more empirically oriented effort, the Penn Discourse Treebank (PDTB) (Prasad et al., 2008) contains manual annotations of discourse connectives with a large set of labels: for example, the connective *while* was annotated with 17 possible senses beyond its for main meanings, which are *Comparison*, *Contrast*, *Concession* and *Opposition* (Miltsakaki et al., 2005).

While a fine-grained characterization provides the necessary theoretical level of linguistic description of discourse structure, it may prove to be intractable to fully automatic processing. Nevertheless, the disambiguation of at least the main senses of discourse connectives is generally required for their translation[1], to avoid the rendering of a wrong sense in translation. For instance, in the following example, the French connective *alors que* in its *contrastive* usage is wrongly translated to the English connective *so*, which signals a *causal* meaning instead[2].

> ***FR:*** *Oui, bien entendu, sauf que le développement ne se négocie pas,* ***alors que*** *le commerce, lui, se négocie.*
>
> ***EN:*** **Yes, of course, but development cannot be negotiated,* ***so*** *that trade can.*

To disambiguate connectives for MT, parallel corpora with sense-labeled connectives are required for training and test. As the PDTB data is in English only, we performed manual annotation on the Europarl corpus. The annotation method, called translation spotting, requires annotators to consider bilingual sentence pairs, and annotate each connective in the source language with its translation in the target language (Meyer et al., 2011). A contrastive analysis showed that these translations can be: a target language connective (in principle signaling the same sense(s) as the source language one), reformulations with different syntactical constructs, or no connective at all. The indications gained with this method are then used in a second step to manually derive and cluster the minimal semantic and theory-independent labels needed to generate correct translations of a connective.

We exemplify this procedure here for the English connective *while*. From the Europarl corpus for English-French, we extracted 499 sentences containing the connective *while*. In 198 cases (43%) the annotators spotted 'no translation' or reformulations of the connective[3]. In the remaining 301 sentences (57%), the annotators identified the corresponding French connectives. As a second step, the French connectives (signaling the same rhetorical relation(s) as *while* itself) were manually clustered under the minimally necessary sense labels to disambiguate the connective *while* in order to translate it correctly from EN to FR. The most frequent French connective clusters and the derived sense labels are the following:

alors que (18%)	Contrast/Temporal
si / même si / bien que / s'il est vrai que (25%)	Concession
tandis que / mais (9 %)	Contrast
tant que (2%)	Temporal/Causal
pendant (1%)	Temporal/Duration
puisque (1%)	Temporal/Causal
lorsque (0.8%)	Temporal/Punctual

[1] The only exception is the case when the ambiguity of a connective is conserved in translation.

[2] Source sentence from Europarl, translated by Moses (Koehn et al., 2007) trained on Europarl.

[3] These are valid translation problems and will be reconsidered for clustering in future work.

Compared to the PDTB sense hierarchy for example, the clustered senses for *while* are as detailed as the PDTB ones on hierarchy level 2, but less detailed than the deepest PDTB level 3. For the temporal meaning of *while,* however, even more differentiation than PDTB level 3 is needed in order to be able to generate the correct translations.

2.2 Pronouns

The resolution of pronouns can be seen as a similar issue to that of resolving connectives in terms of finding a minimal set of features to disambiguate a pronoun for translation. In many cases, depending on the language pair, pronouns can be translated unequivocally, such as the English pronoun *he* generally rendered by *il* in French. However, the French pronouns *il* and *elle* may both be translated into *it* in English if their antecedent, i.e. the noun they refer to, is not human. However, if the antecedent is human, they are in general translated respectively as *he* and *she*. Vice versa, the translation of the English pronoun *it* into French requires knowledge about the gender of its antecedent in the target text. Therefore, whereas the disambiguation of connectives can be done on the source text only, prior to MT, the translation of pronouns requires information about the translation of neighboring fragments.

A close comparison of the English and French pronoun systems shows that the complete list of features characterizing pronoun choice is in theory very large. However, we only aim here to find the minimal set of features which will allow a statistical MT system to avoid generating mistaken pronouns, taking also into consideration the pronoun generated by the system without these features. For instance, in the following example from Europarl, the pronoun generated by Moses is correct in every respect except the gender; therefore, knowledge about the required gender would help correcting *il* into *elle*.

EN: *The **European Commission** must make good these omissions as soon as possible. **It** must also cooperate with the Member States...*

FR: **La **Commission européenne** doit réparer ces omissions dès que possible. **Il** doit également coopérer avec les États membres ...*

3. Automated Disambiguation for Machine Translation

To improve the output of MT, we propose automatic methods that attempt to disambiguate, or at least set additional constraints, on the translation of connectives and pronouns. These methods can either be used as direct input to MT, or to prepare training data for it. For instance, using surface features such as part-of-speech tags or syntactical and dependency parses, we have built classifiers (Meyer et al., 2011) for the senses of the English connectives *since* (Temporal, Causal, or Temporal/Causal) and *while* (Temporal/Causal, Temporal/ Punctual, Temporal/ Durative, Contrast/Temporal, Contrast, or Concession), as well as for the French connective *alors que* (Temporal, Contrast, Temporal/Contrast).

	since	*while*	*alors que*
Baseline (most frequent sense)	51.6%	44.8%	46.9%
SVM classifier	85.7%	60.9%	54.2%

Table 1: Accuracies of sense disambiguation for the connectives *since* (700 sentences), *while* (300) and *alors que* (400). For comparison, the baseline is the majority class in each training set, i.e. respectively *Cause, Concession,* and *Contrast.*

Classifiers were also built for pronoun disambiguation, considering in addition to features from the source text also features from a candidate translation, such as information about the preceding noun phrases, the candidate Moses translation of the pronoun computed from the GIZA++ word alignment, and various ways to determine gender constraints – for the translation of English *it* into French – from the gender of the preceding nouns (e.g., majority, most recent, etc.). Although this method bears similarities with that of LeNagard and Koehn (2010), we do not attempt to identify explicitly the antecedent, in the target language, of the pronoun under consideration, but train classifiers to use the optimal combination of features to infer the correct gender. Of course, this approach cannot pretend to be fully accurate, but compares favorably to state-of-the-art accuracy of automatic pronoun resolution.

The accuracy of the classifier, a decision tree trained using the C4.5 algorithm, is 61% using ten-fold cross-validation on a set of 393 sentences from Europarl annotated with the correct pronoun. The task was to correct the Moses candidate translation of English *it* into French (*il, elle, le, la, l', lui, celui-ci, celle-là, ce, c'*) using automatic alignment and automatically extracted surface features. If the alignment is manually corrected, then the accuracy reaches 64%. This small increase shows that alignment is not the main issue, also because it cannot deal with cases when the MT system omitted the pronoun in translation. However, when the gender prediction is manually corrected, the accuracy reaches 88%, which shows that, as expected, gender is the main feature required for correct translation of *it* into French.

4. Integration into Statistical MT

We experimented on three ways to propagate the above-mentioned discourse information annotated to connectives into the MT processing chain. The integration of annotated pronouns proceeds differently, as a way to post-edit candidate pronouns generated by MT.

The first method to integrate the minimal sets of labels for discourse connectives is to tag their occurrences directly in the phrase table of an already trained statistical MT system. During the training stage, a phrase table is generated with all phrase pairs found by the word alignment, with their lexical probability and frequency scores. We tagged three senses of the connective *while*, namely *Temporal* (1), *Contrast* (2) and *Concession* (3) in the phrase table of a trained Moses MT system for EN-FR. The most frequent French translations were: (1) *pendant que*, *(tout) en + V-ant*, (2) *alors que, tandis que*, (3) *bien que*. Each phrase containing *while* was automatically checked if it is followed by a corresponding translation. If found, the word form *while* was annotated with *while-1*, *while-2* or *while-3*, and, in addition, the lexical probability score was set to one (all other occurrences were left untagged). Translations tests with a set of 20 sentences already led to noticeably better translations (i.e. automatically generated translations closer to the reference translations, especially in terms of the connective) which were also confirmed by a rise in the BLEU score of 0.8 absolute.

A second method that we explored is the opposite of forcing the system to use the tagged connectives. They are instead automatically tagged in a large corpus which is used for SMT training, where all connectives followed by their tags and their corresponding translation in the parallel corpus can be learned by the system. Every occurrence has thereby to be tagged by the disambiguation tool using the classifier model. A third and similar approach to this method is to directly use the manually annotated discourse connectives after the sense clustering. This has the advantage that the hand-annotated resources are correct (gold standard) as opposed to the automated tagging, which is well below 100% accuracy and may therefore propagate a certain error rate in the whole translation process. We built and trained SMT systems able to handle the same manually or automatically tagged data. As a basis for comparison, two other systems were trained on the same two corpora, by discarding all labels (resulting in 4 SMT systems). When comparing the manually tagged system to its untagged counterpart, the tagged system got closer to the reference translations of a test set of 35 sentences in 21 cases versus 14 cases only for the untagged system (the counts were done based on manual checking of the connective translation and the surrounding words and syntax). Even the automatically tagged system, tested on 62 sentences, performed noticeably better in 14 cases compared to its untagged counterpart.

For pronouns, we evaluated the effect on translation of replacing every candidate translation of the English *it*, in the MT output to French, by the translation proposed by our classifier, as a form of post-editing. By definition, this method is only applicable to sentences where a pronoun was indeed generated by MT (about 95% of the sentences). We performed five different runs, training on 353 sentences and testing on 40. In the fully automatic setup, this resulted, on average, in improving pronoun choice from incorrect to correct in 10.8 sentences (27%), but also in turning 6.6 (16%) correct pronouns into incorrect ones. The global result is thus an improvement of about 10% of the overall pronoun accuracy. In these experiments, our classifier did not change the pronoun proposed by MT in 22.6 sentences (56%), of which 27% were correct and 29% were incorrect.

5. Conclusion and Future Work

Integrating discourse information into statistical MT systems remains a challenging task, but one which has the potential to improve over the current sentence-by-sentence MT paradigm. The contrastive corpus analyses and the translation-oriented, multilingual annotation methods have shown to positively affect the output of current statistical MT systems. We will further investigate the automated disambiguation methods for pronouns and connectives as well as for verbal tenses. The performance and error rate of the disambiguation tools is crucial in order to generate annotated resources which are as error-free as possible in order to not negatively influence the SMT training and testing on these resources.

References

Nicholas Asher. 1993. *Reference to Abstract Objects in Discourse*. Kluwer Academic Publisher, Dordrecht, NL.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbs. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL 2007 (45th Annual Meeting of the ACL), Demonstration Session*, pages 177–180, Prague, Czech Republic.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit X*, pp. 79–86, Phuket, Thailand.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: towards a functional theory of text organization. *Text*, 8(3):243–281.

Thomas Meyer, Andrei Popescu-Belis, Sandrine Zufferey, and Bruno Cartoni. 2011. Multilingual Annotation and Disambiguation of Discourse Connectives for Machine Translation. *Proceedings of SIGDIAL 2011 (12th annual SIGdial Meeting on Discourse and Dialogue)*, pp. 194–203, Portland, OR.

Eleni Miltsakaki, Nikhil Dinesh, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2005. Experiments on sense annotations and sense disambiguation of discourse connectives. In *Proceedings of the TLT 2005 (4th Workshop on Treebanks and Linguistic Theories)*, Barcelona, Spain.

Ronan Le Nagard and Philipp Koehn. 2010. Aiding Pronoun Translation with Co-Reference Resolution. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and Metrics MATR*, pp. 258–267, Uppsala, Sweden.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of LREC 2008 (6th International Conference on Language Resources and Evaluation)*, pp. 2961–2968, Marrakech, Morocco.

Formalising translation behaviour with parallel treebanks

Oliver Čulo, Silvia Hansen-Schirra
Johannes Gutenberg-Universität Mainz
culo|hansenss@uni-mainz.de

1 Introduction

Statistical machine translation, in a simplified view, is based on extracting translation probabilities from parallel corpora. However, these corpora are used rather uncritically. Factors like translation direction or text type/register have largely been neglected. The factors mentioned here have a major influence on the production of human translation. As e.g. (Koehn & Schroeder 2007) show, the influence of register is such that training SMT models, both the monolingual language model as well as the translation model, on domain-adapted data has a positive effect on the correctness of translations. But, in order to account for these factors and to derive them automatically, high-quality annotated resources are necessary.

It is probably to a large extent the role of translation studies and contrastive linguistics to study and describe divergences in local structures (cf. Hawkins 1986; König & Gast 2007 for English and German) as well as the role of non-local factors such as text type and their effect on translation (cf. Hansen-Schirra et al. forthcoming for English-German translations). In addition to this, the present contribution introduces the formalisation of factors like register and translation direction as well as local divergences in the dependency structure of a sentence. We try to link these factors to certain translation phenomena, and suggest first steps for the implementation of an algorithm in order to facilitate the adoption of domain, register and translation pair knowledge in MT.

2 Registerial influences

Register has an effect on various dimensions of a text. For instance, register has an effect on word order. Table 1 shows that the number of subjects in sentence-initial position varies depending on the register.[1] German speeches adhere more frequently to the canonical word order SVO than German shareholders' letters and even more frequently compared to German fictional texts. When translating from English, which has a rather fixed SVO word order, into German these register-specific word order patterns should be taken into account otherwise atypical word order frequencies might cause interference effects in the target texts, which might in turn haven in impact on the target language.

GO_FICTION	42,16 %
GO_SHARE	45,87 %
GO_SPEECH	54,54 %

Table 1: Percentage of subjects in sentence-initial position

Figure 1 shows typical shifts in syntactic functions for the language pair English-German and the influence of the register. Shifts from subject to object are, for instance, more common for political essays and instructional texts whereas they are less frequent for fictional texts – and this holds true irrespectively of the translation direction.

[1] All results and examples presented in this article are all taken from the English-German CroCo Corpus (cf. Hansen-Schirra et al. forthcoming).

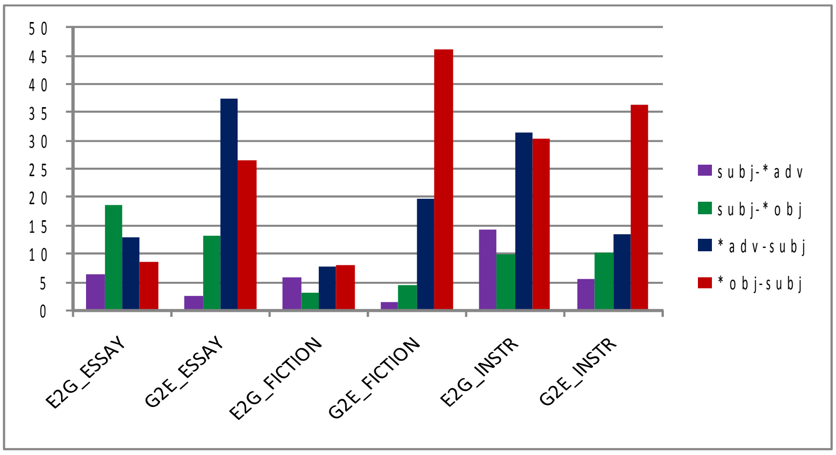


Figure 1: Shifts in syntactic functions

3 Typological influences

Figure 1, however, also shows a typical shift which is triggered by typological differences: the non-canonical subject positions in German often cause objects being placed in sentence-initial position. These sentence-initial German objects are translated with English subjects. This means that the word order is kept in the translation, but as the direct object cannot be kept in sentence-initial position in English an active-passive shift is applied (cf. the translation procedure **modulation** by Vinay & Darbelnet 1958). This phenomenon is illustrated through the following examples taken from the subcorpus of shareholders' letters:

<u>*Wichtige Erfolge*</u> [DIRECT OBJECT] *können wir bereits verzeichnen, weitere werden folgen.*
<u>*Some important*</u> [SUBJECT] *successes have already been achieved, others will follow.*

<u>*Einzelheiten*</u> [DIRECT OBJECT] *können Sie diesem Bericht entnehmen.*
<u>*Additional details*</u> [SUBJECT] *are contained in this report.*

Consequently, for the translation direction from German into English the object-to-subject shift should be more frequent than translating from English into German. This typologically driven pattern can be seen from figure 1 where object-to-subject shifts are more typical of the translation direction German-English – irrespectively from the register.

A similar translation behaviour can be detected when looking at part-of-speech shifts (cf. the translation procedure **transposition** by Vinay & Darbelnet 1958). Figure 2 shows that shifts from verbal to nominal word classes are typical of the translation direction English-German (e.g. adverb-adjective, verb-adjective, verb-noun) while the opposite shifts from nominal to verbal constructions are less frequent (e.g. adjective-adverb, noun-adverb, noun-verb). This conforms to the typological differences between English and German, the latter being more nominal and content-oriented (cf. House 1997). Nevertheless, figure 2 also shows register-specific translation behaviour – e.g. when it comes to high-frequency nominalisation patterns in English-German instructions:

If vertical bars appear on the display after adjusting [VERB] *the focus , press ...*
Falls nach dem Einstellen [NOUN] *des Fokus vertikale Streifen auf dem Display erscheinen , drücken Sie ...*

When replacing [VERB] *the lamp be sure to turn off power and unplug the power cord.*
Schalten Sie vor dem Austausch [NOUN] *der Lampe das Gerät aus und entfernen Sie das Netzkabel.*

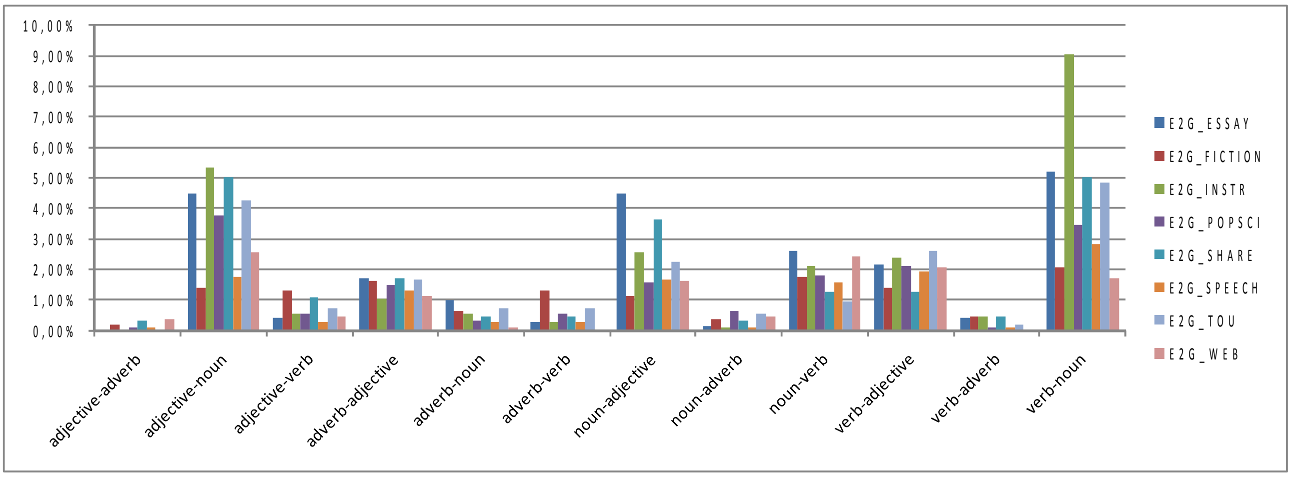


Figure 2: Shifts in part-of-speech

4 “Local” divergencies in dependency structure

Dependency treelet translation (e.g. Ding and Palmer 2004; Quirk, Menezes, and Cherry 2005) is one branch of syntactically informed SMT. (Ding and Palmer 2004) present methods which cope with various changes in the dependency structures, e.g. when head and dependent switches take place or when a dependent is removed and appears as dependent of another head. The account of (Quirk, Menezes, and Cherry 2005) deals with various re-ordering phenomena, e.g. to deal with post- rather than pre-position of modifiers, a typical difference between French and English.

While register as well as typological features may have an influence on the frequency of such phenomena, these phenomena are at first local (effect-wise). We want to add further descriptions of such local phenomena which we have observed while annotating parallel dependency structures and aim to link them to certain translation properties. In the following, we will present a selection of the phenomena we encountered.

Previously, two types of alignment phenomena with regard to translation shifts were defined in the course of the CroCo project: empty links, corresponding to 1:0-alignments, and crossing lines, mostly corresponding to shifts in grammatical function. These alignment phenomena were defined on the flat, top-level only annotation of grammatical functions in CroCo. However, when looking at alignments in dependency annotations, these concepts must be adapted to fit the added dimension of depth in the trees.

We annotated and aligned a sample of the CroCo corpus (around 4,000 sentences from 8 registers) in dependency fashion. For this, we used the tree editor TrEd[2]. The figures 3 to 5 show examples of this annotation and alignment. Due to the display settings in TrEd, the trees from the original sentences are on the right, the trees of the translated sentences on the left.

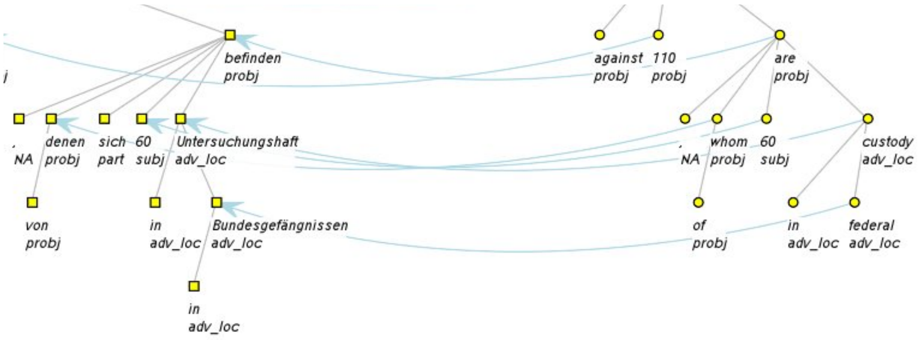


Figure 3: An added leave: sich *is inserted as additional complement in German due to changed valency properties*

[2] http://ufal.mff.cuni.cz/~pajas/tred/ (last visited June 14th 2011)

One typical phenomenon in translation is that of a syntactic complement being added or deleted. In the case of an additional complement, this triggers an **added leaf**. We see such an example in figure 3, where the English verb *be* has been translated with a reflexive variant *sich befinden*. The reflexive pronoun *sich* has no alignment and triggers a new node plus a new incoming edge. In the other translation direction, i.e. in cases where a complement is dropped, we speak of a **dropped leaf**.

Sometimes, the addition or deletion of nodes rather corresponds to the fact that a phrasal expression in one language matches a single word in the other language. We see such a case in figure 4. Here, the expression *in the face of* corresponds to the German adverb *angesichts*. A number of nodes in the English tree is collapsed into a single node in the German tree. This also happens in the very frequent cases of compound nouns which are written as one word in German, but as separate words in English (e.g. *multi word expression – Mehrwortausdruck*). In cases in which several nodes are collapsed into one, we speak of **collapsed nodes**, in the opposite case we speak of **expanded nodes**. Note that in the case of figure 4, for the expressions *in the face of* and *angesichts* we get treelet pairs in which the head is to be filled. In terms of formalisation of the structures, this correspond to type-B trees ("root unlexicalised trees") as defined by (Ding and Palmer 2004).

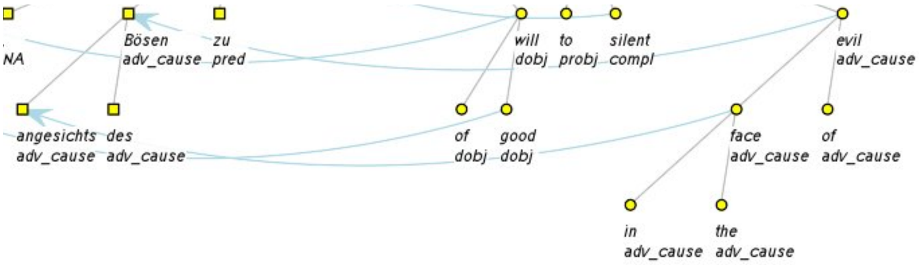


Figure 4: An expanded node: The phrasal expression in the face of *becomes the adverbial* angesichts *in German, a typical divergency for the two languages*

However, changes in the dependency structure not only take place at the terminal nodes of the tree, but sometimes within the tree, changing the path from the root to the affected terminal node(s). In the case of the example in figure 5, the auxiliary verb *did* is present in the English original (*He never gave her gifts the way the old man did*) as verbal substitution, but this verbal branch is cut in the German translation (*Der hat ihr nie was geschenkt, so wie der alte Mann*). This is due to a typical contrast between English and German, the latter rather construing cohesion through elliptical constructions (cf. Hawkins 1986). With respect to dependency annotation, we obtain **cut branches**, or, when translated from German to English, **inserted branches**. Besides enlarging the number of nodes and edges and prolonging the path from the root to the daughter tree of the inserted branch, these inserted branches usually have additional effects. In the example from figure 5, the fact that the verbal branch is cut triggers a shift of the German *Mann* to a modal adverbial, as opposed to its English counterpart which is the subject of the auxiliary.

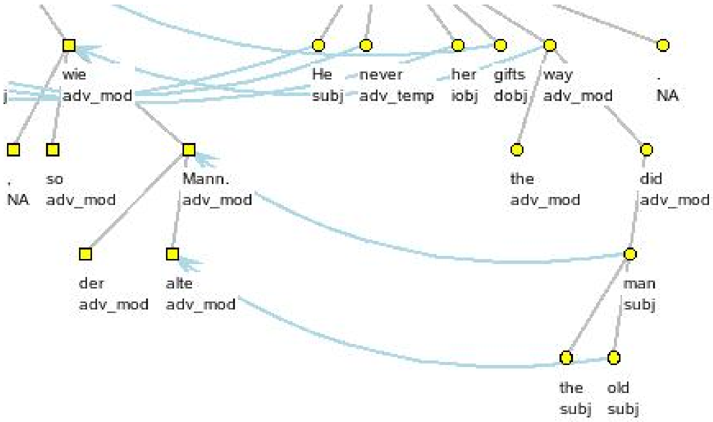


Figure 5: The node did *plus in- and outgoing edges was cut from the tree in the German translation. The former dependent of* did *is shifted from a subject to a modal adverbial.*

5 Discussion

The present contribution has shown the influence of register on translation shifts, as well as of various factors like typological differences between languages or valence divergencies on the dependency structure of translated sentences. The former findings go along with the findings in SMT that training language and translation models on domain-adapted data will improve the performance of the model. The latter findings present a selection of shifts in dependency structures which can be included as tree configurations dealt with in dependency-based SMT

While formal accounts on translation divergencies are available (e.g. Vinay & Darbelnet 1958, Catford 1965), algorithmic formalisations such as (Dorr 1994) are needed in order to facilitate adaption by the MT community. The examples and results presented here have shown that domain and register knowledge as well as patterns typical of the translation direction can be quantified and categorised according to the independent variables involved (language, register, translation direction, etc.). Using parallel treebanks and exploiting them quantitatively and qualitatively has paved the way for the formalisation of human translation behaviour. The implementation of algorithms which represent this translation knowledge for MT will be our next steps in future research.

6 References

Catford, John C. 1965. *A linguistic theory of translation. an essay in applied linguistics*. Oxford: Oxford University Press.

Ding, Yuan, and Martha Palmer. 2004. Automatic learning of parallel dependency treelet pairs. In *The first international joint conference on natural language processing (IJCNLP-04)*.

Hansen-Schirra, Silvia, Stella Neumann, and Erich Steiner. forthcoming. *Cross-linguistic Corpora for the Study of Translations. Insights from the Language Pair English-German*. Berlin: De Gruyter.

Hawkins, John A. 1986. *A comparative typology of English and German. Unifying the contrasts.* London: Croom Helm.

House, Juliane. 1997. Mißverstehen in interkulturellen Begegnungen. In *Wie lernt man Sprachen - wie lehrt man Sprachen*, 154-169.

Koehn, Philipp, and Josh Schroeder. 2007. Experiments in domain adaptation for Statistical Machine Translation. In *ACL Workshop on Machine Translation 2007*.

Quirk, Christopher, Arul Menezes, and Colin Cherry. 2005. Dependency treelet translation: syntactically informed phrasal SMT. In *Proceedings of the 43rd annual meeting of the ACL*, ed. Ann Arbor, 271-79.

Vinay, Jean-Paul, and Jean Darbelnet. 1958. *Stylistique comparée du français et de l'anglais. Méthode de translation*. Paris: Didier.

Using annotated corpora for rapid development of new language pairs in MT

1 Introduction

The present paper introduces the PRESEMT MT system whose most innovative feature is that new language pairs can easily and rapidly be set up. To this end, PRESEMT uses publically available resources and tools as much as possible. These comprise corpora of various types and corpus annotation tools such as statistical taggers and chunkers.

With the advent of statistical machine translation (SMT) corpora have started to play an important role in machine translation (MT). However, whereas large monolingual corpora are mostly available, e.g. through the world wide web, large bilingual corpora are much harder to obtain. The PRESEMT MT system uses a mix of small bilingual corpora and large monolingual corpora to overcome this bottleneck. The small bilingual corpora (several hundred sentences) will not be used to extract statistical parameters (they would be too small), but to automatically (!) extract a bilingual phrase structural mapping, a kind of contrastive (or synchronous) grammar of two languages.

Recent years have seen a rapid increase in the availability of publically available tools for corpus annotation. Statistical taggers and chunkers are available for many languages. These tools will be integrated to do a source language and target language preprocessing which builds the input for the synchronous grammar.

Bilingual dictionaries are available to a large extent for many language pairs, some of them open source, for some publishers are willing to make them available, at least for research purposes. However, they are normally not tagged in appropriate and systematic ways and thus need processing before being integrated into the system.

Another set of tools that build a substantial part of the PRESEMT MT system are statistical modules that make the choices on the target language level. Those tools are mainly language models derived from huge corpora of several billion words. On the basis of these algorithms choices about word readings, placement of articles etc. are made.

2 Problems faced in MT

MT faces two problems, a cost problem and a quality problem. MT systems have high development costs and the translation quality is poor for certain phenomena. Rule-based MT (RBMT) and statistical MT (SMT) face these two problems in different ways:

RBMT:

- Cost problem: manually written linguistic resources are expensive.
- Quality problem: Word translation disambiguation is rather poor.

Strength:

- RBMT has the theoretical means to account for rich morphology and non-local phenomena.

SMT:

- Cost problem: large bilingual corpora are difficult to obtain.
- Quality problem: rich morphology and non-local linguistic phenomena are a problem.

Strength:

- Word translation disambiguation is accounted for rather well.

Hybrid MT systems have been proposed in order to benefit from the strengths of both systems while overcoming their weaknesses. PRESEMT is also a hybrid system. It uses linguistic knowledge in the preprocessing stage and by applying a synchronous grammar (which has been automatically derived from a bilingual corpus). It also uses statistical knowledge to a considerable extent by processing the target language. It also attempts to minimize the cost and the quality problem in the following way:

2.1 Minimising the cost problem

- Rule-based components are not manually written but automatically generated.
- Publically available taggers and chunkers are used as much as possible.
- There is no dependency on large bilingual corpora, instead, a small parallel corpus suffices. In addition, easily available large monolingual corpora are used.
- Available bilingual dictionaries are used. No specific tagging is needed in the lexical entries. The tagging information is taken from other resources such as the tagged monolingual corpora.

2.2 Minimising the quality problem

- Linguistic representations are used to account for rich morphology and non-local phenomena.
- Large monolingual corpora are used to account for word translation disambiguation and TL structure and morphology.

3 Outline of the translation process

3.1 Preprocessing

Shallow parsers are adopted to annotate a small parallel bilingual corpus, which then is aligned on word, chunk, and clause level, based on information found in a bilingual dictionary. From the bilingual corpus, cross-linguistic information is automatically extracted which serves as rule base for a synchronous grammar parser. The actual translation process consists of two phases:

3.2 Structure Selection module

The incoming source language (SL) sentence is tagged, chunked and clause-chunked by shallow parsers. The resulting flat structure is fed into a synchronous grammar parser that builds up an SL tree structure and a set of corresponding target language (TL) structures.

3.3 Translation equivalent selection module

Information extracted out of large monolingual corpora is used to do lemma disambiguation, to select the best TL structure and to determine morphological features such as case, person, number and gender, and to generate the appropriate TL tokens.

4 Translation pair English – German

For the translation direction English (EN) – German (DE) which is a first prototype the following resources and tools have been used:

Publically available resources and tools:

- A small bilingual corpus comprising 300 parallel sentences
- A large monolingual German corpus comprising 7 million tokens (Again this is only the first prototype. There are German corpora available that comprise several billion words).
- TreeTagger for English and German, RFTagger for German morphology information

Additional resources and tools:

- A bilingual dictionary German – English with about 800 000 entries
- A rule-based clause chunker[1]
- Slightly extended Earley chart parser for parsing synchronous grammars

Derived resources:

- synchronous grammar productions
- language models
- token generation table

[1] Clause chunkers are needed if structural divergences between SL and TL depend on clause type or clause boundaries.

4.1 Bilingual corpus English – German

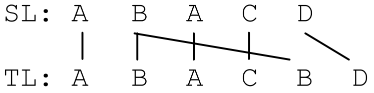
The bilingual corpus is the cross-linguistic heart of the MT system. The bilingual corpus is used to define structural mappings from SL to TL. It should be representative in the sense that it should contain the grammatical constructions that are expected to occur in the sentences to be translated.

The bilingual corpus can be text-type specific. So for example, the bilingual corpus used in DE-EN is taken from a website of the EU (http://europa.eu/abc/12lessons/index_en.htm). It outlines the history of the EU. It does not contain any direct questions or 1. or 2. person pronouns. It is appropriate to take such a corpus if the MT system focuses on descriptive text. In this case it is even advisable not to add direct questions to the bilingual corpus since it would unnecessarily complicate the generation of appropriate TL structures.

4.2 Synchronous grammar generation

The limited size of the bilingual corpus does not allow for statistical methods for deriving synchronous grammars as in Chiang 2007. Also, unlike other tree-to-tree translation approaches (Eisner 2003, Cowan et al. 2006, Zhang et al 2007), the system proposed here does not use deep syntactic processing of the corpus data. Instead, shallow parsers are adopted to annotate the small parallel bilingual corpus. Before deriving synchronous grammar productions, the bilingual corpus is aligned on chunk and tag level using the phrase aligner developed by Tambouratzis et al 2011. The alignments are then converted into productions. In order to extend the sentence patterns covered by the corpus alignments several strategies are employed:

The sentential chunk and tag alignments are broken down into the smallest self-contained alignments. These are then converted into productions. Consider the following schematic example in which the alignment of B can also be taken as an abstract representation of the translation of English simple verbs into German separable prefix verbs which is a non-local phenomenon that poses problems for statistical MT:[2]



The self-contained alignments are converted into productions. Since the alignments of A and D are self-contained one-to-one alignments they are turned into unary productions. Only the complex alignment of B affords a more complex production. Here, some linguistic insight is fed into the production generation. The chunks intervening between the split chunk B are replaced by a clause node CL. Thus the derived production covers more sentence types than the one found in the corpus. The format of the productions is: ‘SL rule’ ⬄ ‘TL rule’. CL is also introduced as mother node for the productions that express chunk alignments. For each production a recursive and a non-recursive variant is generated. In the following only the recursive variants are listed.

$CL\boxed{1} \Rightarrow A\boxed{2}\ CL\boxed{3} \Leftrightarrow CL\boxed{1} \Rightarrow A\boxed{2}\ CL\boxed{3}$
$CL\boxed{1} \Rightarrow C\boxed{2}\ CL\boxed{3} \Leftrightarrow CL\boxed{1} \Rightarrow C\boxed{2}\ CL\boxed{3}$
$CL\boxed{1} \Rightarrow D\boxed{2} \Leftrightarrow CL\boxed{1} \Rightarrow D\boxed{2}$
$CL\boxed{1} \Rightarrow B\boxed{2}\ CL\boxed{3} \Leftrightarrow CL\boxed{1} \Rightarrow B\boxed{2}\ CL\boxed{3}\ B\boxed{2}$

Another way to extend the coverage of the productions beyond the patterns found in the corpus is to define equivalence classes of tags and to multiply templates according to those equivalence classes. For example all finite tags form an equivalence class. Thus if a production has been generated for the finite tag 3.Pl.Pres, the corresponding productions for all other person, number and tense specifications are automatically generated. Another equivalence class consists of different noun tags for names and regular nouns in singular and plural form. And the current system also treats NP and PP chunks as mother nodes in productions as equivalence class. [3]

[2] A corresponding natural language sentences would be:
They **accepted** it immediately.
Sie **nahmen** es sofort **an**.
‘they accepted it immediately SEPPREF’

[3] The TL tags and chunks corresponding to the SL tags and chunks are also specified in the equivalence classes.

In order to cut down the number of TL structures produced by the productions, ambiguities in tag alignments are not spelled out in different productions but represented as a local TL tag disjunction. Tag alignment ambiguities arise if SL and TL taggers assign tags with different granularity. E.g. the English TreeTagger assigns IN to both prepositions and subordinate conjunctions whereas the German TreeTagger assigns APPR and APPRART to prepositions and KOUS to subordinate conjunctions. The following recursive production accounts for three tag alignments, namely SLT1 – TLT1, SLT1 – TLT2 and SLT1 – TLT3. The disjunctive TL tags are separated by the pipe symbol.

$$\mathtt{A}\boxed{1} \Rightarrow \mathtt{SLT1}\boxed{2}\ \mathtt{A}\boxed{3} \Leftrightarrow \mathtt{A}\boxed{1} \Rightarrow \mathtt{TLT1|TLT2|TLT3}\boxed{2}\ \mathtt{A}\boxed{3}$$

Which TL tag is suitable in a given translation is determined by a lookup in the token generation table which contains for each TL lemma also all possible tags. If this lookup is not decisive then the language models will further disambiguate.

5 Evaluation

The preliminary version of the PRESEMT system has been evaluated using a test set of 50 sentences. In a first run, the PRESEMT prototype has achieved 65% of the NIST scores of Google translate (http://translate.google.de/#) and 40% of the BLEU scores of Google translate. The Moses (http://www.statmt.org/moses/?n=Public.Demos) scores are in between the PRESEMT scores and the Google scores.

A manual evaluation of the translations has shown that the lemma disambiguation often produces suboptimal results. Therefore, it is expected that replacing the simple 3-gram models with more sophisticated language models will considerably improve the test scores. This is planned for the future.

6 Extending to new language pairs

Since even the modules that employ non-statistical methods such as the synchronous grammar parser use only automatically derived resources it is relatively easy to extend the PRESEMT system to new language pairs. Work on DE-EN has already started. Other language pairs that are planned to be included in the near future are: Greek, Norwegian and Czech as SL and English and German as TL.

References

Koehn, Philipp, Abhishek Arun, and Hieu Hoang. 2008. Towards better Machine Translation quality for the German – English Language pairs. *Proceedings of the Third Workshop on Statistical Machine Translation, Association for Computational Linguistics*. 139-142.

Chiang, David. 2007. Hierarchical phrase-based translation. *Computational Linguistics* 33(2):201–228.

Cowan, Brooke, Ivona Kucerova, and Michael Collins. 2006. A discriminative model for tree-to-tree translation. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. 232-241

Earley, Jay 1968. *An efficient context-free parsing algorithm*. Ph.D. thesis, Carnegie Mellon University, Pittsburg, PA.

Eisner, Jason 2003. Learning non-isomorphic tree mappings for machine translation. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of the International Conference on New Methods in Language Processing*. 44-49.

Schmid, Helmut and Florian Laws. 2008. Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging. *COLING 2008*, Manchester, Great Britain.

Tambouratzis, George; Sofianopoulos, Sokratis; Vassiliou, Marina; Simistira, Fotini; Tsimboukakis, Nikos 2011. A resource-light phrase scheme for language-portable MT. *Proceedings of the 15th International Conferecne of the European Association for Machine Translation*. 185-192.

Zhang, Min, Hongfei Jiang, Ai Ti Aw, Jun Sun, Sheng Li, and Chew Lim Tan. 2007. A Tree-to-Tree Alignment-based Model for Statistical Machine Translation. *MT-Summit-07*. 535-542.

Phrase Table Support for Human Translation

Gerhard Kremer Matthias Hartung Sebastian Padó Stefan Riezler

Institute of Computational Linguistics (ICL), University of Heidelberg

Selecting an appropriate translation for a word in context is a difficult task for humans, and the support offered by bilingual dictionaries is unprincipled and spotty. This paper reports on an ongoing study that aims at testing how non-professional translators can profit directly from the data in *phrase tables* generated from large parallel corpora for the purpose of machine translation. We present the design of an experiment in which human translators were asked to translate adjective–noun pairs in context while being supported with different types of information extracted from phrase tables. Our hypothesis is that bigram information from phrase tables will lead to faster and more accurate translation.

1 Introduction

Translating a sentence adequately from one language into another is a difficult task for humans. One of its most demanding subtasks is to select, for each source word, the best out of many possible alternative translations. This subtask is known, in particular in computational contexts, as *lexical choice* or *lexical selection* (Wu and Palmer, 1994). Bilingual lexicons which are commonly used by human translators contain by no means all information that is necessary for adequate lexical choice, which is often determined to a large degree by *context*. Often, dictionaries merely list a small number of translation alternatives, or a small set of particularly prototypical contexts is provided. The provided translations are neither exhaustive, nor do they provide distinguishing information on which contexts they require.

In this study, we ask whether the shortcomings of traditional dictionaries can be evaded by using a data structure used in most current machine translation (MT) systems, namely *phrase tables* (cf. Koehn, 2010b). Phrase tables are merely bilingual lists of corresponding word sequences observed in parallel corpora, and thus provide a compact representation of the translation information inherent in a corpus, complemented with statistical information about the correspondences (e. g., frequencies or association measures). Phrase tables can potentially provide both smaller and larger contexts surrounding a particular target word (i. e., context size can be adapted to specific needs of a translator), but they are not prepared for easy interpretation by human translators.

We aim to investigate how phrase tables can be presented to translators for faster and better translation. We approached this question through an experiment in which users had to solve a translation task. They were presented with different types of phrase table information, and we compare the efficacy of different modes of presenting the information.

To keep the experiment manageable, the current study focuses on one particular construction, namely the translation of adjectives in attributive position (preceding a noun). Adjectives are known to be highly context-adaptive in that they express different meanings depending on the noun they modify (Sapir, 1944; Justeson and Katz, 1995). Second, adjectives tend to take on figurative or idiomatic interpretations, again depending on the semantics of the noun in context (Miller, 1998). Lexical choice is therefore nontrivial, and context-dependent translations are seldom given systematically in dictionaries. For example, consider the adjective *heavy*. In noun contexts like *use*, *traffic*, and *investment*, its canonical translation as German *schwer* is inappropriate. It might be translated as *intensiv(e Nutzung)*, *stark(er Verkehr)*, and *groß(e Investition)*.

After presenting related work, section 3 describes in more detail the experimental setup that we have designed, the data, and our hypotheses.

2 Related Work

Interactive MT systems aim to aid human translators by embedding MT systems into the human translation process. Several types of assistance by MT systems have been presented: *Translation memories* provide translations of phrases recurring during a project, but they have to be provided by the translator the first time they appear, and they are typically restricted to a document, a project, or a domain (cf. Zanettin, 2002; Freigang, 1998). In the TransType system of Langlais et al. (2000), the machine translation component makes *sentence completion predictions* based on the decoder's search graph. The interactive tool is able to deal with human translations that diverge from the MT system's suggestions by computing an approximate match in the search graph and using this as trigger for new predictions (Barrachina et al., 2008). Other types of assistance integrate the phrase tables of the MT systems more directly: Koehn and Haddow (2009) and Koehn (2010a) deploy a phrase-based MT system to display word or phrase *translation options* alongside the input words, ranked according to the decoder's cost model. Finally, full-sentence translations can be supplied for *post-editing* by the user.

While the above cited previous work could show a significant increase in productivity and quality for machine-assisted translation, especially for less qualified translators, the presented experiments allow only for a weak correlation between translation times and translation quality. This is due to the varying complexity of test examples and the varying degree of expertise of human translators. In our experiments we tried to control the variable of translation complexity by restricting the task to adjective–noun pairs of roughly the same ambiguity rate and providing machine assistance for these pairs only. Furthermore, the human translators in our experiments were all native speakers of the target language with a similar educational background (regarding experience in this project's source language English). The goal of our pilot experiment is to provide a basis for re-interpretation of results by using a clear and simple experimental design which allows us to analyse the contribution of each variable.

Table 1: Partitions of the set of 30 adjective stimuli presented to each participant for the factors variability and support. Factor context: Each adjective was embedded into 1 out of 4 sentences, where each adjective occurs with a different adjacent noun.

Variability Class	Translation Support Condition			Noun Context
	none	adjective unigrams	adjective–noun bigrams	
high	5	5	5	× 4
low	5	5	5	

3 Experimental Setup

We conducted series of two experiments. In the first one (the main experiment), participants performed a translation task with different kinds of supporting information. In order to test the impact of presenting phrase tables on translation speed, we measured several time points during each of the participants' translation tasks, using time gain/loss[1] as a measure for the usefulness of machine-aided human translation (as discussed in Gow, 2003).

The goal of the second experiment is to complement the time aspect with a measure of the translation's quality.[2] For this purpose, we collected human judgements for all translations from experiment 1 on a simple 3-point scale. We do not believe (at least not a priori) that this step can be automated through the use of MT evaluation strategies like BLEU (see Papineni et al., 2002) or edit distance (discussed in Gow, 2003), given the restricted phenomenon and the semantic nature of the distinctions that we focus on.

In this abstract, we concentrate on describing the setup of the first experiment. Participants were asked to translate an attributive adjective in sentential context, given one of our set of translation support types. With German participants, we investigated translation from English into German, the participants' native language. This is the preferred type of translation direction in professional human translation as the translator's experience of commonly used words in a particular semantic context is more extensive in the native language. In this experiment we assumed three factors to interact with translation speed and accuracy (cf. table 1): variability class (2 levels), translation support (3 conditions), and noun context (4 sentences per adjective, each sentence with a different adjacent noun).

Variability classes. Stimuli for the translation experiment have been collected by examining the most frequent adjectives from the British National Corpus (BNC), many of which are polysemous, i. e., showing high context-dependent variability in translation (cf. section 1). An analysis with 200 high-frequent adjectives in the BNC showed a highly significant correlation (Spearman's ρ=0.5121) between corpus frequency and variability in translation (operationalised as the number of unique translations in the EUROPARL (Koehn, 2005) v6 phrase table). We divided adjectives

[1] The response times might vary a lot depending on differing translation habits of the participants: One person might select the first cognitively available lexical items, while other persons might take their time to consider alternatives. We plan to modify experiment instructions to equalise translation strategies among participants.

[2] Note that there have been ongoing debates on how translation quality can be assessed objectively (cf. House, 1998).

into two classes: one set that shows a particularly high variability in unique translations, and one set with a relatively low translation variability.
Hypothesis. Highly variable adjectives are more difficult to translate, but translators will profit more from the presentation of phrase table information.

Adjectives and Contexts. For each variability class, we selected 15 adjectives according to the phrase table. For each English adjective, we randomly sampled four full sentences from the BNC (with the adjective in attributive position). In order to minimise variation in translation times, we have restricted the length of sentences to a defined range of number of words and characters (so that reading times are comparable). Note that our setup results in a domain difference between the sentences to be translated (sampled from the BNC) and the phrase table (drawn from EUROPARL)—a standard situation for translation.
Hypothesis. We expect speed differences among adjectives and noun contexts, but will treat them as random effects (cf. section 4).

Translation Support. Finally, we provided three kinds of translation support to participants: (a) no support, (b) the list of translations for the adjective unigram from the phrase table, and (c) the list of translations for the adjective–noun bigram from the phrase table. We presented three distinct candidate translations as support, ranked by their order in the n-best list produced by the moses[3] MT system (trained and tuned on EUROPARL v6) that decoded each target sentence.
Hypothesis. Presenting unigram translations leads to faster and more appropriate translations. Bigram phrases will produce the most appropriate translations, even if translating in this condition might be slower due to the need to read through more complex translation suggestions.

4 Procedure and Evaluation

Participants were all required to be native German speakers with at least a working knowledge of English (they were asked to specify their level of proficiency). Each participant was asked to translate all 30 adjectives, but each adjective in only one sentence context, in order to avoid faster translation of previously seen target adjectives. Of the 30 adjectives, each set of 10 was presented in one of the three translation support conditions. In summary, 85 persons participated.

Note that the design of our experiment does not conform to classic psychological or psycholinguistic expectations: it is not balanced and, with the exception of translational variability, we did not control any variables regarding the sentence stimuli. Instead, our materials mirror the distribution in the corpus. This is a conscious decision that we have taken because (a) there is a very large number of potentially influential factors which are very difficult to control; and (b) we are interested in testing our hypothesis under "practical" rather than idealised conditions.

In order to quantify the influence of the individual factors and test the hypotheses formulated above, we analysed the reading times with a mixed effects model (see, e. g., Baayen et al., 2008). We treated the variability class and the translation support as fixed effects, and the identity of adjective, context, and participant as random effects. On the quality judgement data we (measured the inter-rater correlation and) performed an analysis of variance for our experiment conditions.

[3] URL http://www.statmt.org/moses

References

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59:390–412.

Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Kadivi, S., Lagarda, A., Ney, H., Thomas, J., Vidal, E., and Vilar, J.-M. (2008). Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28.

Freigang, K.-H. (1998). Machine-aided translation. In Baker, M., editor, *Routledge Encyclopedia of Translation Studies*, pages 134–139. Routledge, New York.

Gow, F. (2003). Metrics for evaluating translation memory software. Master's thesis, University of Ottawa.

House, J. (1998). Quality of translation. In Baker, M., editor, *Routledge Encyclopedia of Translation Studies*, pages 197–200. Routledge, New York.

Justeson, J. S. and Katz, S. M. (1995). Principled disambiguation. Discriminating adjective senses with modified nouns. *Computational Linguistics*, 21:1–27.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit*, pages 79–86. Asia-Pacific Association for Machine Translation (AAMT).

Koehn, P. (2010a). Enabling monolingual translators: Post-editing vs. options. In *Proceedings of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, Los Angeles, CA.

Koehn, P. (2010b). *Statistical Machine Translation*. Cambridge University Press.

Koehn, P. and Haddow, B. (2009). Interactive assistance to human translators using statistical machine translation methods. In *Proceedings of Machine Translation Summit XII*, Ottawa, Ontario, Canada.

Langlais, P., Foster, G., and Lapalme, G. (2000). TransType: A computer-aided translation typing system. In *Proceedings of ANLP-NAACL Workshop on Embedded Machine Translation Systems*, Seattle, WA.

Miller, K. J. (1998). Modifiers in WordNet. In Fellbaum, C., editor, *WordNet. An Electronic Lexical Database*, pages 47–67. MIT Press.

Papineni, K., Roukos, S., Ward, T., and 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*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sapir, E. (1944). Grading. A study in semantics. *Philosophy of Sciences*, 11:83–116.

Wu, Z. and Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138.

Zanettin, F. (2002). Corpora in translation practice. In Yuste-Rodrigo, E., editor, *Proceedings of the Workshop Language Resources for Translation Work and Research at the International Conference on Language Resources and Evaluation (LREC)*, pages 10–14, Las Palmas de Gran Canaria. LREC.

Inside the Monitor Model: Processes of Default and Challenged Translation Production

Michael Carl and Barbara Dragsted
ISV, Copenhagen Business School, Dalgas Have 15, DK-2000 Frederiksberg

Abstract

It has been the subject of debate in the translation process literature whether human translation is a sequential and iterative process of comprehension-transfer-production or whether and to what extent comprehension and production activities may occur in parallel. Tirkkonen-Condit (2005) suggests a "monitor model" according to which translators start with a *literal default rendering procedure* where a monitor interrupts the default procedure when problems occurs. This paper suggests an extension of the monitor model in which comprehension and production are processed in parallel by the default procedure. Deviations from this default behaviour are triggered through text production problems and involve conscious decision-making processes, related to text comprehension or to text production problems.

In an experiment we compare text copying with translation activities under the assumption that text copying is a prototypical literal default rendering procedure. Both tasks, translation and text copying, require decoding, retrieval and encoding of textual segments, but translation additionally requires transfer into another language. Comparing user behaviour obtained in copying and translation experiments, we observe surprisingly many similarities between the two different activities. Copyists deviate from the default literal text reproduction into more effortful text understanding, and much of the translators' behaviour resembles that of copyists. We discuss how extended source text (ST) comprehension is triggered through production problems, during translation as well as during text copying.

The stratificational model of translation production

According to the eye-mind hypothesis (Just & Carpenter, 1984)[1] there is a strong correlation between where one is looking and what one is thinking about. The eye-mind hypothesis is controversial; on the one hand it is well known that the observation of a correlation between two events does not imply a causal relation, and hence no conclusion can be made regarding the *existence* or the *direction* of cause and effect only from the observation that two events, e.g. gaze location and mental processes, correlate. On the other hand, the strong correlation assumption between gaze and mind has be questioned, e.g. by (John R. Anderson, Dan Bothell, and Scott Douglass, 2004) who find that longer gaze durations do not correlate with greater problems of memory retrieval.

In Translation Process Research it has often been stated that gaze location reflects the focus of attention of the translator (e.g. Hyrskykari, 2006). That is, when the gaze focusses on the ST the mind is involved in ST comprehension processes, and when the gaze is directed at the TT, the mind is involved in text production processes. Longer gaze durations on the ST or TT reflect bigger comprehension or production problems respectively (Pavolovic and Jensen, 2009). These assumptions fit well with a stratificational process model of translation, which states that at any one time the translator either reads (understands) the ST, transfers it into the target language, or types the translation.

Craciunescu et al. (2004), for instance, claim that "the first stage in human translation is complete comprehension of the source language text". Only after this complete (i.e. *deep*) comprehension is achieved can the translation be produced. Similarly Gile (2005) suggests a stratificational translation process model, in which a translator iteratively reads a piece of the ST and then produces its translation. First the translator creates a "Meaning Hypothesis" for a ST chunk (i.e. a Translation Unit) which is consistent with the "context and the linguistic and extra linguistic knowledge of the translator" (p. 107). Subsequently, a translation is produced.

Also Angelone (2010) supports that translators process in cycles of comprehension-transfer-production and that "uncertainties" of translators can be attributed to any of the comprehension, transfer, or production phases. He claims that "non-articulated indicators, such a pauses and eye-fixations, give us no real clue as to how and where to allocate the uncertainty" [p.23]

[1] "there is no appreciable lag between what is fixated and what is processed"

The monitor model

Some scholars challenge this view, stating that translation processes can also be based on a *shallow* understanding and that ST understanding and TT production can occur in *parallel*. According to Ruiz et al. (2008) "the translator engages in partial reformulation while reading for the purpose of translating the source text". They assume that in parallel processing "code-to-code links between the SL and TL [are involved] at least at the lexical and syntactic level of processing". Similarly, Mossop (2003) claims the existence of "direct linkages in the mind between SL and TL lexicogrammatical material, independent of 'meaning'", and that a translator "automatically produces TL lexical and syntactic material based on the incoming SL forms".

In a study comparing reading behaviour for different purposes, Jakobsen & Jensen (2008) investigate (among other things) the difference between test persons reading a text for comprehension and reading a similar text in preparation for translating. Their study showed that reading purpose has a "clear effect on eye movements and gaze behaviour" and they suggest "that a fair amount of pre-translation probably enters into the reading of a text as soon as it is taken to be the source text for translation" [p.116].

Although it is unclear what is exactly meant by "pre-translation", such findings are obviously in contrast with the eye-mind hypothesis when assuming a stratificational model of translation. Reading with "a fair amount of pre-translation" implies certainly different mental activities than reading for understanding, but the eyes remain in both cases on the ST. Since it may be difficult (if not impossible) to disentangle which parts of the gaze behaviour are to be linked to text understanding and which correspond to pre-translation, either the eye-mind hypothesis has to be weakened or the stratificational model of translation has to be reconsidered.

We assume, with Tirkkonen-Condit, that "literal translation is a default rendering procedure, which goes on until it is interrupted by a monitor that alerts about a problem in the outcome. The monitor's function is to trigger off conscious decision-making to solve the problem" (Tirkkonen-Condit 2005: 407-408). In our interpretation of the model, the literal default rendering procedure implies parallel, tightly interconnected text production and comprehension processes: while the mind is engaged in the production of a piece of text, the eyes search for relevant textual passages to gather the required information needed to continue the text production flow. When this default procedure is interrupted by the monitor, can we observe gaze patterns on the ST or on the TT which indicate comprehension- or production-related translation problems. Note, however, that these decision-making processes are triggered problems related to production activities. Similarly, Gile (2005) reports that deeper understanding of the ST may emerge through problems in TT production, rather than when first reading a ST passage. He points out that the translation practice indicates processing from a production-based perspective:

> Oftentimes, the translator does not test Meaning Hypothesis until after verbalising it in the target language (...) Frequently, he or she only realizes there is a problem when trying to read the first target-language version (...) in other words, when already in the reformulation phase.

A clear-cut allocation of "uncertainties" to one of the stratificational processes then becomes difficult, since such processes do not normally exist independently in the translator's mind. Not only is it infeasible (or impossible) to distinguish between comprehension and pre-translation activities during reading for translation, but also the borders between ST understanding and TT production problems become blurred.

Observations from translation experiments

We investigated patterns of typing behaviour from text copying and translation experiments. Our investigation is based on empirical data obtained in 10 copying sessions and 15 translation sessions. The experiments were recorded using the Translog 2006 software (Jakobsen and Schou, 1999), which logs keystrokes and gaze movements during a reading, translation or text production task.

We take it that copying (i.e. re-typing) a text may be processed in a much more shallow/parallel manner than translation since: 1) apart from a lexical encoding and decoding (John, 1999), text copying does not, in theory, require any deep ST (or TT) understanding; 2) copying can proceed in parallel to a maximal degree, since no revision[2] and no lexical or structural transfer is required. Typing patterns and speed would thus essentially depend on the typing skills of the copyist. Comparing copying behaviour and translation

[2] some revision may be going on, for instance correction of typos, but these activities are of a different kind than most of those in translation revisions.

behaviour would reveal the additional effort of translation.

We observe that most of the text is copied smoothly and straight-forwardly, with only little look-ahead in the ST. But we also observe that the copying activity may trigger extended reading activities in the ST context when a text passage is unclear. That is, word meaning seems to be processed also during text copying. As predicted in the monitor model (Tirkkonen-Condit, 2005), the copying pause occurred when typing the unclear expression, rather than when reading it the first time. Lack of comprehension is discovered (or at least actions are taken) only during (re)production of the text, rather than during first reading. We then compared typing in a copying task with typing in a translation task and observed basically the same patterns. While much of the typing activity for translation resembles text copying in L2, the gaze is on average slightly further ahead in the ST during translation than when copying.

We also looked at passages of conscious, effortful text production, which uncovers more entangled relations between comprehension and production. Similar to text copying, translation examples clearly show that difficulties occur when reformulating (render and address) the translation rather than when reading the ST. The translation of a phrase may already start before the translator knows how to go on with the translation. The initial translation guess is not always appropriate, and sometimes the beginning of the phrase must be revised. In some cases the ST context has to be re-consulted and in other cases the produced TT is reconsidered.

Inter-key time spent during unchallenged production was approximately the same in copying and in translation during periods of coherent typing. In addition we observed parallel and alternating reading and typing behaviour, where reading and writing activities occurred respectively simultaneously or sequentially. There are more pauses during translation than during copying, indicating more alternating processes in translation.

Looking at gaze activities we found that the number of ST word fixations during parallel unchallenged translation activities equalled approximately those of unchallenged copying while there were more ST fixations during alternating translation activity.

Conclusions

Two types of translation behaviour can be distinguished:

1. Much of the translation drafting is unproblematic and approximately within the time limits predicted for text copying by Johns' (1999) TYPIST model[3]. Translators look only a few words ahead into the ST from the position where they are currently translating. In an alternating mode, ST decoding adds to the typing time, while in a parallel mode decoding and encoding run in parallel. Many of the smaller translation problems, such as multi-word translations or local reordering, may be solved by looking only a few words ahead. We suspect that the degree of parallel activity depends on experience and typing skills of the translator. A touch typist would more likely exhibit parallel processing behaviour, while a translator with less developed typing skills would show alternating translation patterns.
2. At some points in the translation, extensive reading behaviour can be observed, signalling more serious translation problems. Depending on the type of problem, it may be necessary for the translator to re-scan the ST or the TT. In both cases, the increased reading activity seems to be triggered by a TT production problem rather than by a ST comprehension problem. That is, we observed that the ST was understood, and meaning hypotheses were generated only to the extent required to keep on producing target text. If, for whatever reason, TT production cannot go on smoothly, and the typing flow is interrupted, the missing information needs to be retrieved. This may lead to the re-reading of a ST passage with a view to verification or reinterpretation, and/or the revision of the produced TT.

In a stratificational comprehension-transfer-production theory of translation, this behaviour is difficult to explain. Ruiz et al. (2008) point out that "comprehension for translation does not differ from normal monolingual comprehension since comprehension and reformulation occur at different stages ". But if the ST would first have to be completely understood before a translator could start translating it, why would the translation purpose have an impact on the ST reading behaviour? Instead, we assume that "Meaning Hypotheses" are constructed to the extent and at the moment they are needed to continue the task at hand. Different meaning hypotheses are required for different kinds of activities, e.g. a technician reading a car

[3] This conclusion is based on our translation material from English into Danish, two relatively close languages with similar word order.

repair manual needs a different kind of understanding than a translator translating the same text into another language. The reading purpose thus determines what kind and depth of meaning representation is required. During translation and text copying, the ST meaning is often not elaborated and tested until the writing process – which leads to the surprising conclusion that comprehension does not precede, but rather follow text production.

An extended version of this paper will appear in the journal “Translation: Computation, Corpora, Cognition”.

References

Anderson, John R.; Dan Bothell, and Scott Douglass, (2004) Eye Movements Do Not Reflect Retrieval Processes Limits of the Eye-Mind Hypothesis, PSYCHOLOGICAL SCIENCE , Volume 15—Number 4, 2004,

Angelone, Erik. 2010. “Uncertainty, uncertainty management and metacognitive problem solving in the translation task”. In **Translation and Cognition**, Shreve, Gregory M. and Erik Angelone (eds.), 17–40.

Carl, Michael and Martin Kay (2011) Gazing and Typing Activities during Translation: A Comparative Study of Translation Units of Professional and Student Translators, Accepted for publication in META

Carl, M. and Jakobsen, A. L. (2010). Towards statistical modelling of translators’ activity data. International Journal of Speech Technology, 12(4), 124-146.

Craciunescu, Olivia; Constanza Gerding-Salas, Susan Stringer-O'Keeffe (2004), Machine Translation and Computer-Assisted Translation: a New Way of Translating? Translation Journal, Volume 8, No. 3 , July 2004 **http://translationjournal.net/journal/29computers.htm**

Gile, Daniel (2005) La Traduction. La comprendre, l'apprendre. Paris: Presses Universitaires de France

Dragsted, B. (2010). “Coordination of reading and writing processes in translation: An eye on uncharted territory”. In G. M. Shreve, and E. Angelone (eds), Translation and Cognition, pp. 41–62, Amsterdam/Philadephia, Benjamins

Hyrskykari, Aulikki 2006, Utilizing eye movements: Overcoming inaccuracy while tracking the focus of attention during reading , Attention aware systems - Special issue: Attention aware systems

Jakobsen, Arnt Lykke (1999): „Logging target text production with Translog“. In Hansen, Gyde, Hrsg.: Probing the Process in Translation: Methods and Results. (Copenhagen Studies in Language 24). Kopenhagen: Samfundslitteratur. 9–20.

Jakobsen, A. L. and Schou, L. 1999. Logging target text production with Translog, In Copenhagen Studies in Language, volume 24, Samfundslitteratur, Copenhagen, 9-20

Jakobsen, A. J. and Jensen, K. T. H. (2008) Eye movement behaviour across four different types of reading task, In Göpferich et al. (eds) Looking at Eyes. Eye Tracking Studies of Reading and Translation Processing, CSL 36, 103-124.

John, Bonnie E.. Typist: a theory of performance in skilled typing. Hum.-Comput. Interact., 11(4):321–355, 1996.

Just, M.A., & Carpenter, P.A. (1984). Using eye fixations to study reading comprehension. In D.E. Kieras & M.A. Just (Eds.), New methods in reading comprehension research (pp. 151–182). Hillsdale, NJ: Erlbaum.

Kromann, M. T. (2003). The Danish Dependency Treebank and the DTAG treebank tool. In Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003), 14-15, November, Växjö. 217–220.

Mossop, Brian. An Alternative to ‘Deverbalization’. Technical report, York University, 2003. **http://www.yorku.ca/brmossop/Deverbalization.htm**

Pavlovic, Nataša, and Kristian T. H. Jensen. (2009). Eye tracking translation directionality. In Translation Research Projects 2, eds. Anthony Pym and Alexander Perekrestenko. Tarragona: Intercultural Studies Group, p. 93-109. Available at: http://isg.urv.es/publicity/isg/publications/trp_2_2009/index.htm.

Perrin, Daniel. 2003. Progression analysis (PA): investigating writing strategies at the workplace. Pragmatics, 35:907–921.

Ruiz, C.; N Paredes, P Macizo, M T Bajo , (2008), Activation of lexical and syntactic target language properties in translation. Acta Psychologica, Volume: 128, Issue: 3, Pages: 490-500

Staub, A. and Rayner, K. 2007. Eye movements and on-line comprehension processes. In Gaskell, G. (ed.), The Oxford Handbook of Psycholinguistics. Oxford: Oxford University Press. 327-342.

Tirkkonen-Condit,Sonja (2005) The Monitor Model Revisited: Evidence from Process Research, META, Volume 50, numéro 2, avril 2005, p. 405-414