Spoken language corpora in German (L1) and Swedish (L1) and the acquisition of foreign language grammar

Christiane Andersen
University of Gothenburg
Department of Languages and Literatures
christiane.andersen@tyska.gu.se

Spoken language vs. written language and corpus analysis

- „(‚written language bias'; Per Linell 1982).
- Die Vorstellungen darüber, was Sprache ist, leiten sich primär aus dem Umgang mit der Reflexion von geschriebener Sprache her.“ (Duden. Die Grammatik 2005:1176)
- „Now what the spoken corpus does for the spoken language is, in the first instance, the same as what it does for the written: it amasses large quantities of text and processes it to make it accessible for study.“ (Halliday 2004:14)
- „So transcribing spoken discourse – especially spontaneous conversation – into written form in order to observe it, and to use the observations as a basis for theorizing language, is a little bit problematic. Transcribing is translating, and translating is transforming; I think to compile and interpret an extensive spoken corpus inevitably raises questions about the real nature of this transformation.“ (Halliday 2004:16)

Spoken language corpora as a basic design concept for a foreign language grammar (German for Swedish students)

- Which grammatical categories are more frequent than others?
- Are there grammatical categories which are more typical in spoken language than in written language?
- Do spoken language corpora detect preferential grammatical structures and categories for different practices in spoken communication?
- In what way could the contrastive perspective on spoken language corpora (for instance the contrast of German and Swedish) be helpful for designing a foreign language grammar?

German and Swedish spoken language corpora in contrast

- The Mannheim Spoken Language Corpus (2006) and transcripts of German-Swedish trade fair discourses (270 transcripts, 717 025 tokens) http://dsav-wiss.ids-mannheim.de/DSAv

- GSLC: Göteborg Spoken Language Corpus (2006) (1 422 830 tokens) http://www.ling.gu.se/projekt/tal/index.cgi?PAGE=3

- Steps of quantitative investigation: 1. Comparison of word forms of the German and Swedish corpus and their frequencies in contrast. 2. Comparison of word forms of subcorpora in accordance with different types of communication.

Types of communication in the investigated German and Swedish spoken language corpora

Kommunikationstyp (Korpora des gesprochenen Deutsch)	Transkripte	Tokens	Kommunikationstyp (Korpora des gesprochenen Schwedisch)	Transkripte	Tokens
Predigt	2	31 04	Kirche (Church)	2	1 027 4
Diskussion	15	67516	Diskussion (Formal Meeting)	49	44513 8
Private Unterhaltung	18	3834 3	Unterhaltung (Informal Conversation)	23	10545 3
Vortrag	12	2848 3	Vorlesung (Lecture)	2	1 468 3
TV-Diskussion	14	9423 1	Diskussion (Political Debate)	4	3 680 5
Verkaufsgespräch	4	45 03	Verkaufsgespräch (Shop Games)	42	2 278 0
Insgesamt	65	23618 0	Insgesamt	122	63513 3

The most frequent word forms in contrast

We compiled the data in accordance with their frequency:

1. All tokens (word forms) in the German and Swedish total corpus were sorted according to their frequency. Only the positions 1-200 were investigated more carefully.

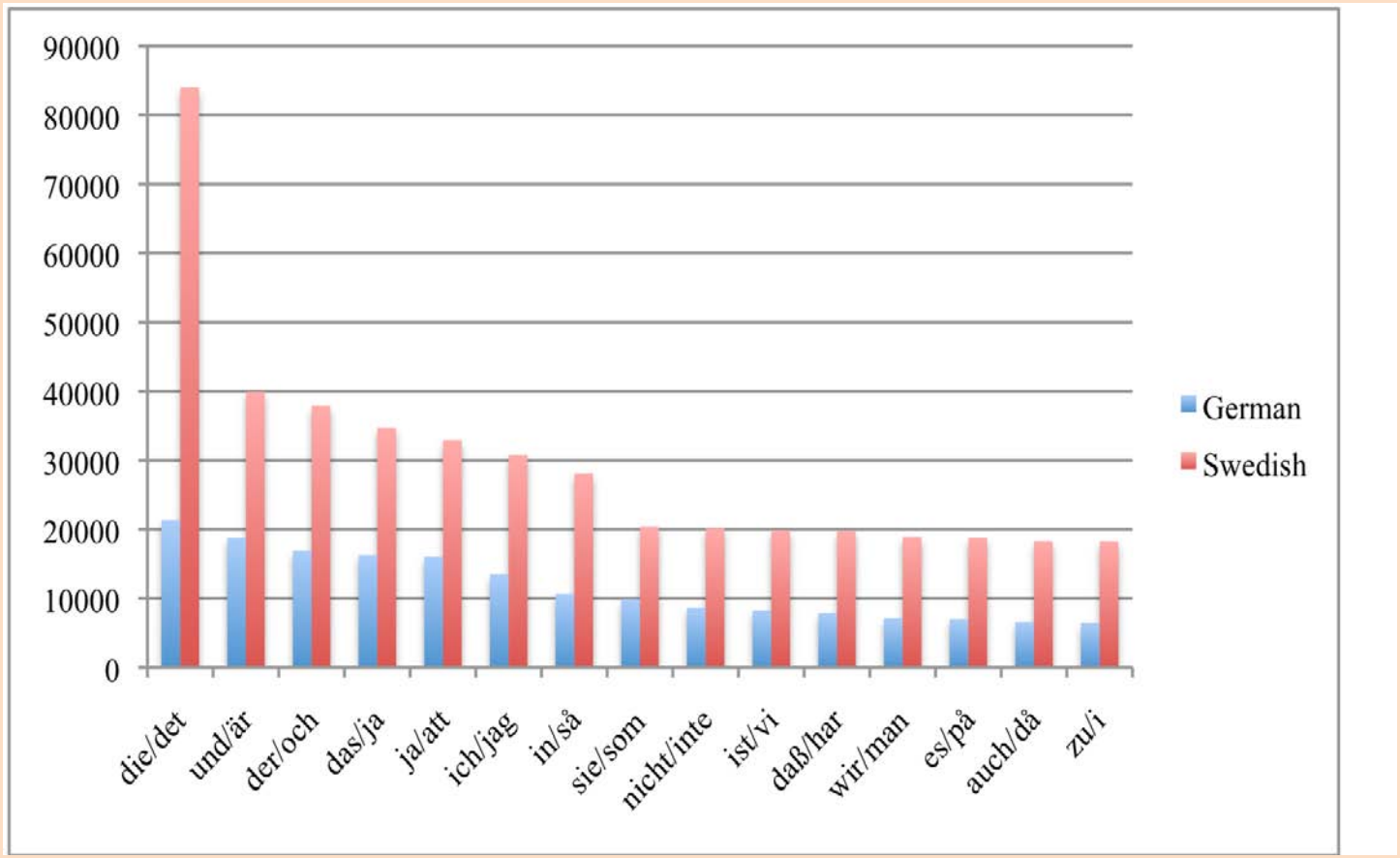
2. We combined six types of communication to subcorpora and calculated their individual frequency lists.
3. We annotated manually the 200 most frequent word forms for every subcorpus.

Frequencies of word forms in German and Swedish

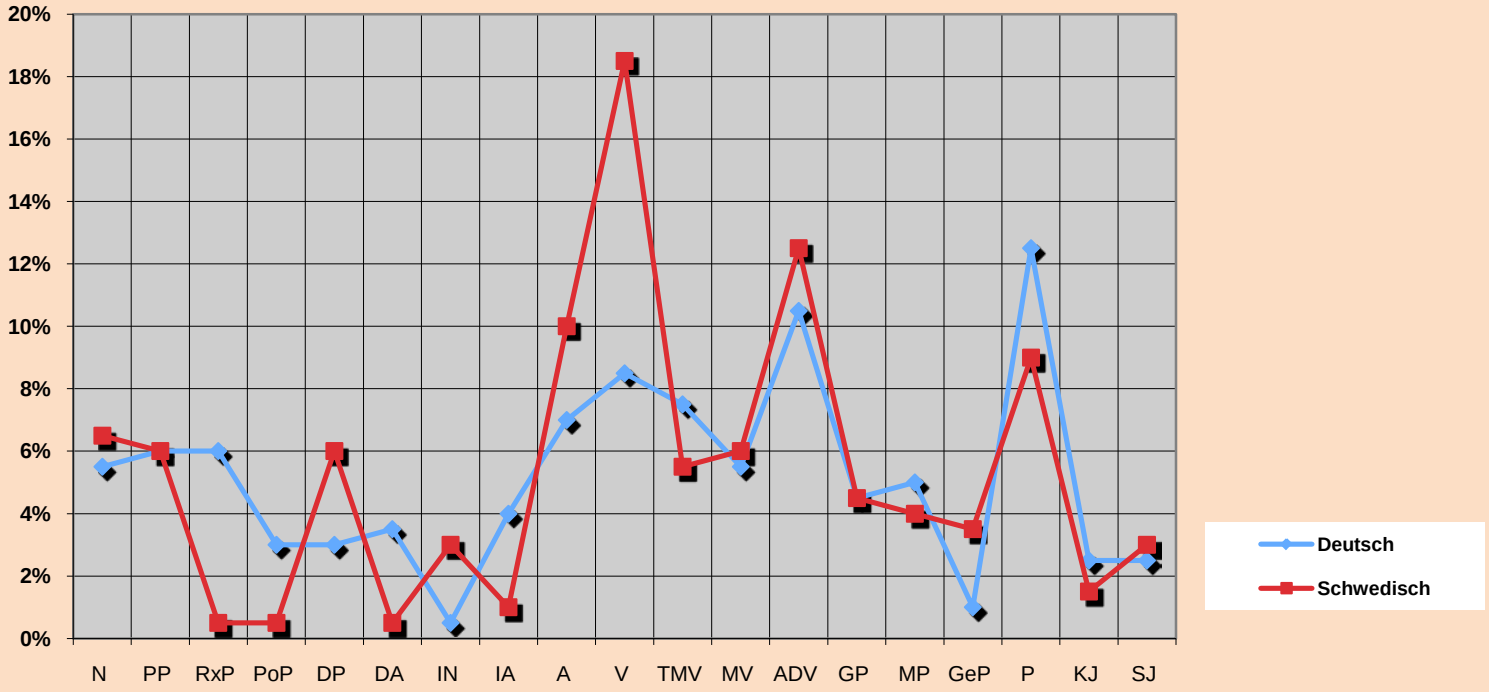
Most frequent word forms in spoken language corpus (Swedish – German, absolute figures)

Wortform	Frequenz	Rang	Wortform	Frequenz
die	21310	1	det	84007
und	18757	2	är	39965
der	16970	3	och	37884
das	16296	4	ja	34731
ja	16033	5	att	32912
ich	13459	6	jag	30745
in	10666	7	så	28079
sie	9898	8	som	20417
nicht	8683	9	inte	20164
ist	8278	10	vi	19880
daß	7911	11	har	19752
wir	7139	12	man	18923
es	6989	13	på	18775
auch	6566	14	då	18265
zu	6474	15	i	18252

Most frequent word forms in spoken language corpus (Swedish – German)



Frequencies of parts of speech in percentage (position 1-200, German, Swedish)



N (noun), PP (personal pronoun), RxP (reflexive pronoun), PoP (possessive pronoun), DP (demonstrative pronoun), DA (definite article), IN (indefinite pronoun), IA (indefinite article), A (adjective), V (verb), TMV (auxiliary verb), MV (modal auxiliary), ADV (adverb), GP (graduate particle), FP (focus particle), MP (modal particle), GeP (discourse particle), P (preposition), K (conjunction), SJ (subjunction)

Frequencies of parts of speech in German and Swedish: the verb forms

- Full verbs are more frequent in the German corpus than auxiliary verbs and adjectives. The most frequent verb forms are *sagen* (2442), *gesagt* (938) and *sagt* (511) followed by *frage, meine, gibt, machen, glaube, kommt, geht, weiß, kommen, sehen, wissen, fragen* and *gemacht*.

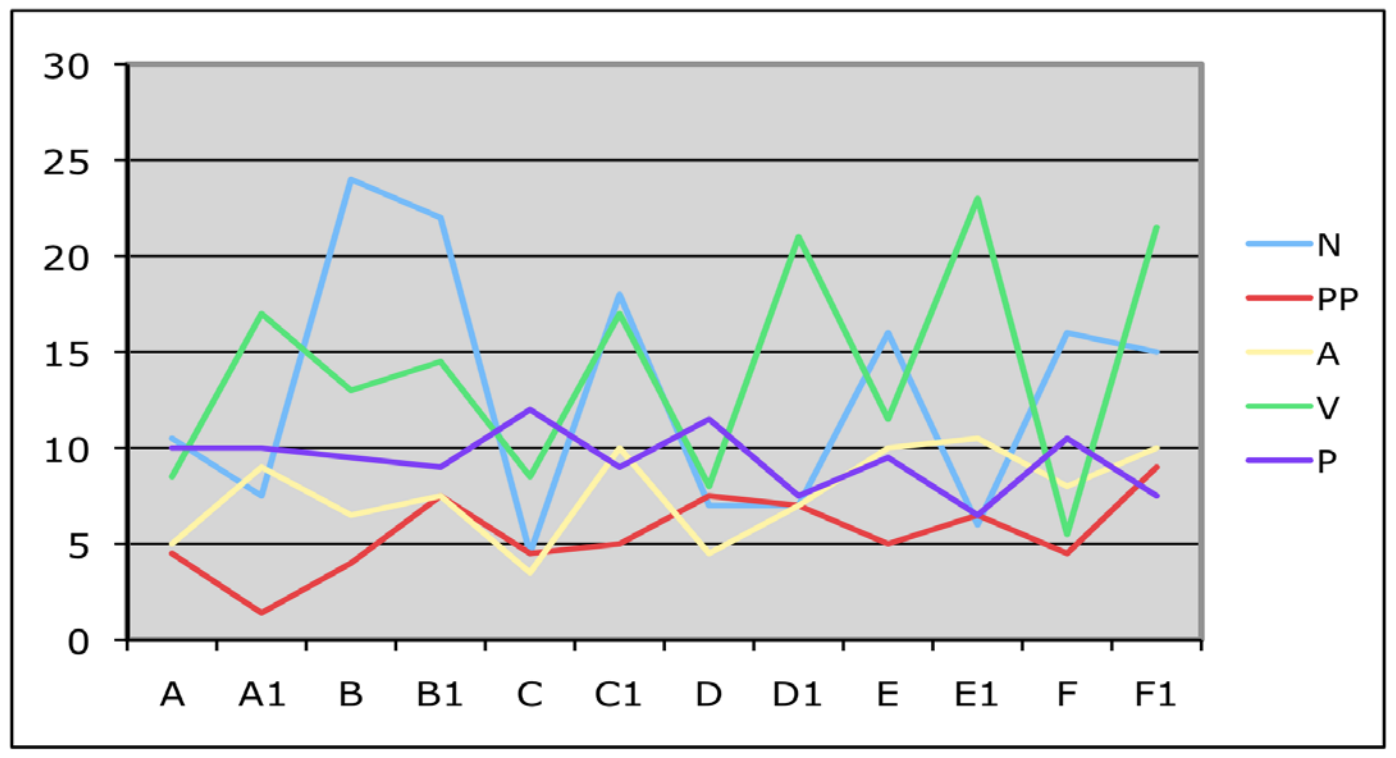
- In the Swedish corpus the most frequent verb forms are *får* [finite form of *dürfen, müssen, bekommen, gestatten* etc.] (4773), *tycker* [finite form of *meinen, finden* etc.] (4074), *säga* [infinitive/finite form of *sagen*] (3977), *finns* [*gibt*] (3770) and *kommer* [finite form of *kommen, werden*] (3632). Frequent are as well verb forms like *tycker, vet, tror, menar, säger, ser, tänker, tänkte* [introducing direct speech]. The preterite forms *sade, kom, fick, tänkte, gjorde* are more frequent than the corresponding German forms.

The Most frequent word forms in the German and Swedish subcorpora

A. Diskussion (Germ.) – A1. Formal Meeting (Sw.); B. Predigt (Germ.) – B1. Church (Sw.); C. TV-Diskussion (Germ.) – C1. Political Debate (Sw.); D. Private Unterhaltung (Germ.) – D1. Informal Conversation (Sw.); E. Verkaufsgespräch (Germ.) – E1. Shop/Games (Sw.); F. Vortrag (Germ.) – F1. Lecture (Sw.)

Rang	*A*	*A1*	*B*	*B1*	*C*	*C1*	*D*	*D1*	*E*	*E1*	*F*	*F1*
1	die	det	die	och	die	det	ja	det	ja	det	die	det
2	und	är	und	som	der	att	das	jag	das	ja	und	att
3	der	och	der	det	und	och	und	är	und	är	der	man
4	das	att	das	att	ich	vi	ich	ja	sie	jag	das	och
5	ich	så	ist	i	das	som	die	och	die	så	in	är
6	ja	jag	daß	vi	ja	är	der	så	is	inte	ist	så
7	in	ja	den	är	sie	i	so	du	dann	och	wir	i
8	daß	vi	wir	en	ist	har	da	att	ich	du	sie	som
9	nicht	som	es	till	in	för	nicht	inte	der	har	ich	en
10	wir	har	nicht	den	nicht	en	dann	ju	da	dom	es	på
11	ist	i	dem	för	daß	eh	in	men	so	den	daß	här
12	sie	man	er	på	wir	den	is	på	s	en	zu	då
13	es	då	in	så	zu	jag	also	i	also	på	nicht	för
14	zu	inte	sie	av	es	på	ist	då	man	här	eine	har
15	den	på	dann	inte	auch	om	auch	dom	n	då	von	öh

Percentage of parts of speech (position 1-200) in accordance with types of communication: nouns (N), prepositions (PP), adjectives (A), full verbs (V), pronouns (P)



A. Diskussion (Germ.) – A1. Formal Meeting (Sw.); B. Predigt (Germ.) – B1. Church (Sw.); C. TV-Diskussion (Germ.) – C1. Political Debate (Sw.); D. Private Unterhaltung (Germ.) – D1. Informal Conversation (Sw.); E. Verkaufsgespräch (Germ.) – E1. Shop/Games (Sw.); F. Vortrag (Germ.) – F1. Lecture (sw.)

Frequencies of parts of speech and word forms

The most frequent verb forms in German and Swedish subcorpora *Verkaufsgespräch* vs. *Shop*:

- *kommt, machen, gibt, nehmen, sehen, sagen, lassen, geht, brauchen (braucht), saugen, stimmt, wiedersehen, gefüllt, wissen, verlegen, auswechseln, reicht, spülen*

- *få (får, fick, fått), tror, vet, komma (kommer, kommit), finns, gå, (går), se(ser,sett), sade, ta, göra (gjort, gjorde), köpa (köper, köpt), heter, ser, tycker (tyckte), tar, kolla, ligger, säga (säger), står, spela (spelar), vara, menar, kostar, titta, vänta, tänkte, prata, sälja (säljer)*

Nicht and *inte* in foreign language grammar and in spoken language corpus

- Grammatical functions of *nicht* in corpus samples:

    1. NICHT with a modal verb
    2. NICHT with a degree marker (*Gradpartikel)*
    3. NICHT as focus marker (*Fokuspartikel)*
    4. NICHT as discourse marker (confirmation signal)
    5. NICHT in a rhetorical question
    6. NICHT in an affiliation
    7. NICHT as negation in an opinion

Frequencies of *nicht* and *inte* in accordance with their grammatical/discourse functions

	MIT MODAL-VERB	GRAD-PARTIKEL	FOKUSIE-RUNG	GESPRÄCHS-PARTIKEL/ RÜCK-VERSICHERUNG	RHETO-RISCHE FRAGE	IN DER FEST-STELLUNG/ SATZNEGA-TION	IN DER ANSICHTS-ÄUSSERUNG
MANNHEIMER KORPUS (DS00 2,DS 00 3)	15	2	1	14	1	4	12
GÖTEBORGER KORPUS (A3201011)	15	3	11	6	12	35	32

Occurrences of *nicht* in corpus samples

***1. Nicht* (können, wollen)**

$S1: [das KANN ich **nicht**]
$S2: [kann ich **nicht** / will ich] **nich**
$S1: ja
$S2: [mach ich **nicht**] peng
$S1: genau

2. Nicht* as *Rückversicherungssignal

$S2: [hm n / ja n]
$S1: [**nicht** / und da] sag ich nanu was denn /
$S2: ja
$S1: **nicht** /
$S2: ja

3. *Nicht* with verbs like *sagen, meinen* (*Rückversicherungssignal* and *Fokuspartikel*)

$S1: und eh dann sprach ich mit meinem mann darüber und der war dann sehr zornig und sagte also so etwas gehört sich **nicht** wir müssen unsere tochter mal VORNEHMEN **nicht/** was da [nun LOS ist]
$S2: [ja vornehmen] JA aber **nich**t sagen was sich / was sich **nich**t gehört/$S1: ja n
$S2: **sondern** ANDERS rum /

Occurrences of *inte* in corpus samples

1. With *Fokuspartikel*

$A: men då e0 de{t} ju **inte bara** elevernas fel då / att alla ett visst år ska läsa till en viss sak de{t} e0 ju även lärarna som då avråder samtliga

$B: **INTE lärarna** för de{t} kommer faktiskt upp ännu högre uppifrån tycker ja{g}

2. In rhetorical questions

$B: ja varför **skulle vi inte kunna göra** de{t} du vet de{t} är inte lång tid innan blå+ / +klockorn försvinner dom som < love > sa{de} att ja{g} skulle försöka hitta // < [1 klockorna]1 > finns men bladen hittar ja{g} inte längre för dom har försvunnit sa{de} han

3. With verbs or adjectives in an opinion

$B: men nu **vet man ju inte** i dagens läge

$A: **nä** sen e0 de{t} väl de{t} att / vissa / perioder så // e0 de{t} vissa / utbildningslinjer som e0 populära å0 då blir det för många utav de{t} å0 sen andra perioder så e0 de{t} ingen som

Nicht in foreign language grammars

Helbig/Buscha (2005): Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht. Langenscheidt.

1. NICHT as sentence negation, word order(90)
2. NICHT as *Abtönungspartikel*. „Es bleibt lediglich das nicht-negierende *nicht* übrig [...] das in der Tat Abtönungspartikel ist (aber nicht Negationspartikel).“ (242)
3. NICHT as word negation, (skopos) (547)
4. NICHT as „besonderes Stilmittel zur vorsichtigen Bejahung (nur in der Kopplung *nicht – un-* und in der Kopplung *nicht ohne*)“ (559)

Andersson et al (2002): Tysk syntax för universitetsnivå. Studentlitteratur.

1. NICHT as sentence negation
2. NICHT as stress particle in rhetirical questions (selten: „i vissa Fall“; *Was du nicht alles weißt*. [*Vad du vet mycket*!]): idiomatical meaning.

Word form frequencies in spoken language corpora and foreign language grammar (German – Swedish). Which conclusions can we draw?

- Adverbs, particles and pronouns should be introduced earlier and from a syntactic-semantic perspective.
- In traditional foreign language grammars, noun and verb morphology is in main focus. Syntax and semantics of nouns and verbs (especially tense and mood) should be integrated as well.
- Subordinate clause constructions should definitely come in focus starting with *das*s-sentences in contrast to *att*-constractions in Swedish.
- The semantics and word order of discourse particles should come in focus. Discourse particles (*Modalpartikeln, Gesprächspartikeln, Interjektionen*) and several adverbs should be introduced in grammars and text books.
- Foreign language grammars should incorporate adapted samples from natural language corpora.

References

Andersen (Pankow), Christiane (2007): Korpora der gesprochenen Sprache und Fremdsprachengrammatik. In: Hall, Christopher & Kirsi Pakkanen-Kilpiä (Hrsg.). Deutsche Sprache, deutsche Kultur und finnisch-deutsche Beziehungen. Festschrift für Ahti Jäntti zum 65. Geburtstag. Peter Lang. Frankfurt am Main. 197-210.

Andersson, Sven-Gunnar/Brandt, Gisela/Rosengren, Inger (2002): Tysk syntax. Lund: Stu-dentlitteratur.

Duden. Die Grammatik. (2005). Hrsg. von der Duden-Redaktion. Mannheim, Leipzig, Wien, Zürich: Dudenverlag.

Halliday, Michael A. K. (2004): The spoken language corpus: a foundation for grammatical theory. In: Aijmer, Karin/Altenberg, Bengt (Hrsg.) Advances in Corpus Linguistics. Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23). Göteborg 22-26 May 2002. Amsterdam, New York: Rodopi, 11-29.

Helbig, Gerhard/Buscha, Joachim (2001): Deutsche Grammatik. Ein Handbuch für den Aus-länderunterricht. Berlin, New York: Langenscheidt.

Linell, Per (1982): The Written Language Bias in Linguistics. (= Studies in Communication 2.) Linköping: University of Linköping.

Telemann, Ulf/Hellberg, Staffan/Andersson, Erik (1999): Svenska Akademins grammatik 2. Ord. Stockholm: Norstedts.

Christiane Andersen
University of Gothenburg
Department of Languages and Literatures
christiane.andersen@tyska.gu.se