Variationism and Underuse Statistics in the Analysis of the Development of Relative Clauses in German[1]

Anke LÜDELING, Hagen HIRSCHMANN, Amir ZELDES

ABSTRACT

In this paper we introduce a corpus based variationist approach to the study of language change, which hinges on the definition and explicit coding of variables and variants, or competing ‘ways of saying the same thing’, within their usage in corpus data. We use multiple extensible annotation levels to examine variants in the development of relative clauses from Old High German to Modern German, using four comparable deeply annotated corpora of different German language stages. We compare the frequencies of different grammatical categories such as word forms, parts of speech and syntactic constructions to diagnose the most significant changes that are evident in our corpus, and show the advantages of dynamically re-examining quantitative results and categorization systems. Finally we discuss in how far our approach can support theories on language change and lead to insights which enrich previous theoretical accounts.

1. Introduction

Many language change theories are quantitative, describing the gradual change from one form to another form. Any quantitative theory is, of course, built on a qualitative (categorical) analysis – one has to decide which forms to compare. It has often been noted (see among many others the discussion in Labov 2004) that here lies a crucial difficulty in diachronic analysis because categorization is difficult within and across language stages. Different categorizations lead to different analyses and often it is not the conclusions that differ but the basis on which they are built. In this paper we want to explore how a multi-layer corpus architecture, where different layers of analysis can be coded simultaneously helps in understanding change phenomena. The main focus of this paper is methodological. In order to illustrate our point we investigate the development of the German relative clause from Old High German to Modern German. We choose this phenomenon because it has facets in several layers of analysis: syntactic, morphological, and semantic making it a challenging testing ground for a methodology analyzing language change. The complexity of the phenomenon in question also makes it necessary to have a corpus architecture capable of expressing different annotation

[1] We want to thank Eva Schlachter and Jürg Fleischer for interesting discussions and valuable comments.

formats. We must say at the outset that we will not find any qualitative conclusions that are radically new about relative clauses (which are well-researched and well-understood), but even though the diachronic corpus we use is small our results fit with, and enrich, previous work on this subject, and show how such analyses can be performed.

The paper is organized as follows: Section 2 introduces the general theoretical framework behind the study of quantitative variation, charting the competition between different variants realized in each language stage through the use of diachronic corpora. Section 3 introduces and illustrates multi-layer architectures and Section 4 shows the use of overuse/underuse statistics as a corpus-based diagnostic. Section 5 then presents the corpus and the case study of German relative clauses, while Section 6 draws the final conclusions.

2. Variation and variationism

For a long time theoretical linguistics has argued that linguistic systems are rather homogeneous and that variation is accidental and therefore not interesting for theory building.[2] Contrary to that view, many studies in sociolinguistics, historical linguistics and synchronic corpus-based linguistics have shown that variation is not random and that speakers of a language have very fine-grained and consistent knowledge of usage. Starting with Labov's famous 1966 study of phonological variation in New York it has been shown again and again that variation happens on all linguistic levels and most of it is quantitative rather than qualitative.

Variation is only possible if there are several ways of doing 'the same thing' from which the speaker can choose. If a speaker of German for example wants to express the fact that something is acceptable, she/he can say: *X ist akzeptabel* or *X ist annehmbar* or *man kann X akzeptieren* etc. This is only interesting if – as it is argued – the choice between the variants is not random but triggered by grammatical and functional factors. The different variants of 'the same thing' are correlated with other linguistic and extralinguistic factors. Labov (2001, 2004), among many others, has linked variation to social variables. Other studies, such as e.g. those of Biber (1988, 2009) show that there is a lot of variation within a speaker and this can be attributed to different functional needs – it is said that each speaker is able to vary his/her linguistic behavior according to the situation/purpose etc. of the utterance. The obvious and very difficult problem is, of course, to decide what counts as 'the same thing'. Here we want to use the terms **variable** for 'the same thing' and **variant** for a possible realization of a variable. A variable is always an abstraction over several variants.

In addition to such functionally triggered synchronous variation there is diachronic variation – the idea that one variant may increasingly come to take on

[2] We will not go into the long-standing debate between competence-based (generative) models and usage-based models for language change. See Wasow (2007) and Sag & Wasow (to appear) for a discussion.

functions or contexts previously associated with another variant, over time. It is probably impossible to tease these two types of variation apart; most of the time a diachronically 'new' variant occurs first in a given register and then becomes 'fashionable'. Those historical linguists who accept (quantitative) variation as a trigger for language change are called variationists (Labov 1994 & 2001, Rissanen 2008). Figure 1 illustrates the idea that language change cannot be described in terms such as 'in period X people used A and in period Y people used B' – rather there is a gradual change where one variant is becoming stronger while another variant is slowly fading.

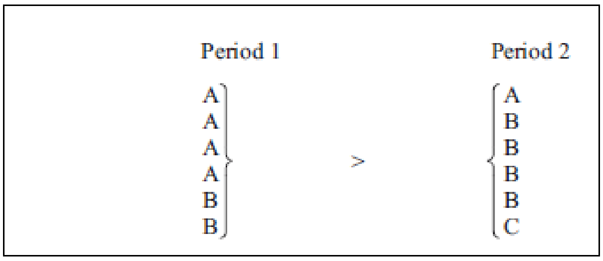


Figure 1: An illustration of how variants of a variable change quantitatively over time. A, B, and C are all variants of a single variable (from Rissanen 2008, 59).

In a variationist approach to language change[3] one therefore needs to define a variable and its variant expressions (see Section 4). Because of the large amount of variation within a language stage (see above) it is crucial to use comparable corpora. Ideally the corpora should contain texts which differ only in one parameter (here: time) so that all differences can be attributed to that one parameter. While it might be possible to build contemporary corpora that fulfil (or come close to fulfilling) that requirement[4], historical texts are, of course, much more diverse and there are many parameters that cannot be controlled for because the information (e.g. about the author or the intended audience) is not known, or a given genre does not exist, or suitable texts (e.g. personal letters) have simply not survived. In this situation some authors use parallel corpora instead of comparable corpora (for European languages this usually means Bible corpora, see Resnik et al. 1999 for a discussion and Zeldes 2007 for an example), but parallel corpora – which necessarily involve translations – come with their own set of problems (see e.g. Baroni & Bernardini 2006 on translationese).

There are many corpus-based studies of language change.[5] While some of them focus on lexical categories that can be researched in unannotated corpora, many involve

[3] Language change does not have to be 'historical'. The same method can be applied to the study of recent or ongoing change, see Mair (2009) for an overview.

[4] This has, for example, been the idea behind the Brown corpus family (see e.g. Leech et al. 2008) or the ICE corpora (Greenbaum & Nelson 2009, http://ice-corpora.net/ice/).

[5] In essence, all historical studies are corpus-based. We use the term corpus here only for electronic corpora.

annotation of some kind. The most well-known annotated historical corpora are probably the treebanks built from the Helsinki corpus (and sometimes additional material, http://www.ling.upenn.edu/hist-corpora/ and Kroch (this volume)) which have been used for many studies. For German there are not (yet) many publically available annotated historical corpora[6].

However while annotated corpora are enormously helpful, they could be even more helpful if some widespread problems are overcome.[7] The annotation is usually done by a group of researchers according to a very specific annotation scheme and research question. The corpus architecture is then not flexible enough to handle annotations with different formats (such as trees, spans or pointing relations) or merge annotations made by different tools. Research questions that do not interest the original annotators or hypotheses that come up during an analysis are not/cannot be included. This means that categorization beyond the provided annotation and quantitative analysis is usually done in separate programs (e.g. spreadsheets) and not coded in the corpus (this is, in essence, the traditional way of working with historical documents, see Meyer 2008). It is therefore not directly available to other researchers and results are not easily reproducible or reusable. In the following we want to show how a flexible corpus architecture that allows various annotation formats, the addition of annotation layers at any point and visualization of quantitative aspects can help in the analysis of linguistic change. Before we go into our case study we will briefly introduce our corpus and the phenomenon we will be looking at.

3. Data and corpus architecture

For our study we use the DeutscheDiachroneBaumbank (DDB, available at http://korpling.german.hu-berlin.de/ddd/search.html), a tiny, but deeply annotated, comparable diachronic corpus of German which consists of the following subcorpora:

- Subcorpus Old High German (OHG), containing the Gospel of Matthew, based on an edition by George Allison Hench (1890). The subcorpus is a part of the Monsee Fragments (written at the end of the 8th century). It consists of 3626 tokens.
- Subcorpus Middle High German (MHG), consisting of a collection of Middle High German sermons, called “Specculum ecclesiae” (written at the

[6] In addition to those described in Kroymann et al. (2004) we are aware of the following annotated historical corpora of German: The Early Modern German Mercurius Treebank (Demske 2007) which is not yet publicly available and the GermanC corpus (http://www.llc.manchester.ac.uk/research/projects/germanc/) which is annotated on several levels (not syntactically).
The situation is changing, however: The projects Referenzkorpus Althochdeutsch (http://www2.hu-berlin.de/sprachgeschichte/forschung/altdeutsch.php) and Mittelhochdeutsche Grammatik (http://www.mittelhochdeutsche-grammatik.de/) will make their material available shortly via ANNIS (Section 3). [Wollen wir hier schon die DDD-URL bekannt machen?]

[7] The same problem pertains to most contemporary corpora as well.

end of the 12th century), based on an edition by Gert Mellenbourn (1944). The subcorpus consists of 2483 tokens.

- Subcorpus Early New High German (ENHG), consisting of a sermon by the preacher Veit Nuber (written 1544), called “Ein kurtze und einfeltige unterweisung zum sterben nutzlich und heilsam den krancken furzuhalten an irem letzten/aus der heiligen schriften zusamen gelesen,”, extracted from the Bonner Frühneuhochdeutschkorpus (Diel et al. 2002). The subcorpus consists of 2673 tokens.
- Subcorpus New High German (NHG), comprising the first four chapters of the Acts of the Apostles from the *Neue evangelistische Übertragung*, a freely available translation of the entire Bible (New Testament 2003, Old Testament 2009) prepared by Karl-Heinz Vanheiden and available from http://www.kh-vanheiden.de/. The subcorpus consists of 3574 tokens.

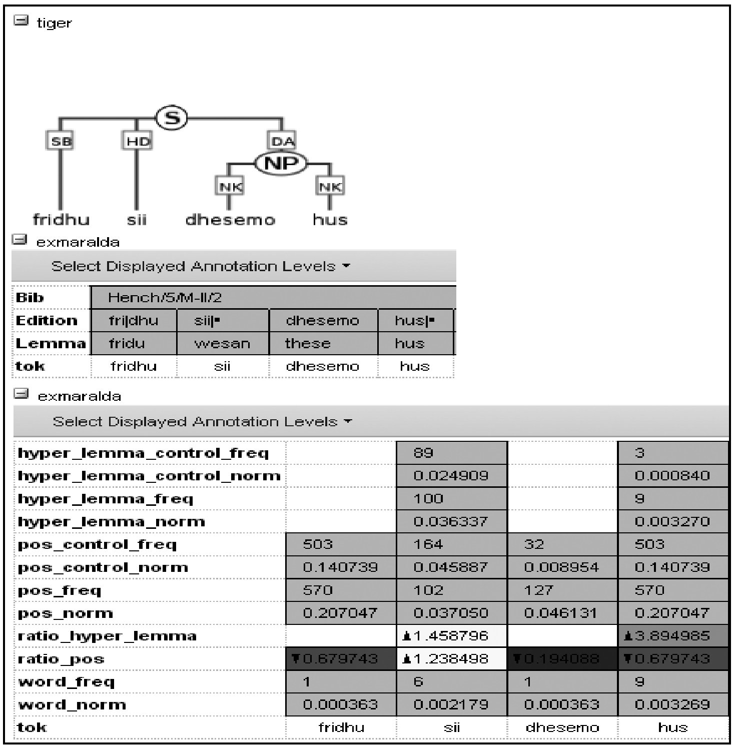
The corpus is annotated as follows:

The NHG corpus contains part of speech tags automatically generated using the TreeTagger (Schmid 1994) and constituency trees generated using the Stanford Parser (Klein & Manning 2003), but no morphological or dependency information. The historical corpora contain the following annotations, which were created manually (see Figure 2).

- part of speech annotation (POS), using the German STTS-tagset (http://www.sfs.uni-tuebingen.de/Elwis/stts/stts.html).
- morphological information based on the TIGER morphological tagset (inflectional morphology).
- syntactic annotation, using the annotation scheme of the Tiger Project (http://www.ims.uni-stuttgart.de/projekte/TIGER/), which is a combination of dependency and constituency annotation[8].
- normalized spelling of the original text based on editions, in order to ensure uniform searchability of word forms.
- hyper-lemmatization to create comparability between language stages, based on the morphologically, or in special cases semantically corresponding New High German lemma.
- absolute and normalized frequencies for word forms, lemmas, POS, and POS-bigrams, as well as Underuse/Overuse ratios and statistical significance for each token as compared to the NHG corpus (see below).[9]

[8] The annotation scheme was developed by Hagen Hirschmann and Sonja Linde. For synchronic corpora of German the TIGER annotation scheme (Brants et al. 2002) has come to be the most influential and widely accepted. In order to make the historical corpora comparable to the modern corpus it was decided to adhere as closely as possible to the original TIGER annotation and propose changes very conservatively

[9] Further annotation levels present in the historical corpora but not used in this study are:


tiger
S
SB
HD
DA
NP
NK
NK
fridhu
sii
dhesemo
hus
exmaralda
Select Displayed Annotation Levels
exmaralda
Select Displayed Annotation Levels

Bib	Hench/5/M-II/2			
Edition	fri|dhu	sii|▪	dhesemo	hus|▪
Lemma	fridu	wesan	these	hus
tok	fridhu	sii	dhesemo	hus

hyper_lemma_control_freq		89		3
hyper_lemma_control_norm		0.024909		0.000840
hyper_lemma_freq		100		9
hyper_lemma_norm		0.036337		0.003270
pos_control_freq	503	164	32	503
pos_control_norm	0.140739	0.045887	0.008954	0.140739
pos_freq	570	102	127	570
pos_norm	0.207047	0.037050	0.046131	0.207047
ratio_hyper_lemma		▲1.458796		▲3.894985
ratio_pos	▼0.679743	▲1.238498	▼0.194088	▼0.679743
word_freq	1	6	1	9
word_norm	0.000363	0.002179	0.000363	0.003269
tok	fridhu	sii	dhesemo	hus

Figure 2: Sample sentence (“peace be this house”) from DDB-OHG with all annotation layers: From top to bottom: syntactic annotation, bibliographic information, text representation in the original text edition, lemmatization, normalized word layer and statistical information for token annotations.

Corpus architecture

The representation of the heterogeneous types of data described above requires a special corpus architecture which is both searchable on all levels simultaneously (i.e. we can find all cases of certain syntax-tree structures overlapping certain spans of orthographic forms with significantly deviating frequencies) and extensible, so that further levels of annotation can be added, modified or removed in the course of the study, easily and independently. The currently most versatile technique for achieving these goals is the use of standoff XML formats, in which primary data and each annotation level are all kept in separate XML files (see Carletta et al. 2003, and

- Normalized lemmatization according to standard dictionary norms for each language stage.
- Bibliographic annotation referring to the editions’ scheme for coding lines in the original manuscripts.

especially Lüdeling/Poschenrieder/Faulstich 2005 in the context of historical corpus architectures). In this case we used PAULA XML (Dipper 2005) to merge annotations from multiple source formats: TigerXML (see http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/doc/html/TigerXML.html) for syntactic annotations and EXMARaLDA XML (see http://www.exmaralda.org/) for other span based annotations, as well as output from automatic tools like the TreeTagger (see above). Through the use of standoff XML it becomes possible for researchers to work concurrently on the same source data (the transcribed manuscripts) without altering it using multiple annotation tools, with the possibility to later revise separate annotations or even apply several versions of the same annotation layer.

To search through the annotated data and visualize our search results we use ANNIS2 (see Zeldes et al. 2009, http://www.sfb632.uni-potsdam.de/d1/annis/). This system grants corpus access to multiple users over a web-browser and provides a query language AQL (ANNIS Query Language) to express arbitrary annotation graphs being searched for. Query results are then visualized in multiple levels according to annotation types, e.g. with syntactic annotations receiving a tree visualizations and span annotations being displayed as grids. For more detailed information on the corpus architecture, the reader is referred to (Zeldes et al., submitted). In the following we show how multiple annotation layers are simultaneously needed to study the development of relative clauses.

4. Comparing quantities: under and overuse of corpus measurements

In order to compare the distribution of variants in different language stages it is necessary to code them in a way that makes them identifiable and extractable for researchers. Since more complex types of variation involve not just surface word forms but also higher-level categories, such as parts-of-speech or syntactic structures, these must be annotated wherever they occur. The idea behind such annotations is that researchers' analyses of language data should be made explicit within the corpus, allowing them to search for and review occurrences of relevant phenomena, no matter how complex (see Leech 1993, Garside et al. 1997). Each annotation category or combination of categories can in a first approximation be seen as a variable in the sense introduced above: different surface forms or lower levels of annotation are the variants of a variable. In this sense, linguistic developments between language stages in a comparable diachronic corpus are already coded in the data itself. The normalized frequency of each phenomenon in each stage can then be extracted and compared.

Once frequencies for a phenomenon have been collected in each comparable subcorpus, standard statistical tests such as the chi-square test or a test of equal proportions can be used to evaluate whether there is a significant deviation between the respective data samples or to compute a statistical model of the development. In the case of language data a particularly high significance is expected, since the assumption of statistical independence between linguistic phenomena in a text is not granted and since the usually large sample size (thousands or even millions of words) makes even small deviations in frequency appear significant (see Kilgarriff 2001; Evert 2005, 2006).

In historical corpora, the amount of data is often quite small, as is the case here, though as we will show below, even small corpora can yield interesting quantitative results given appropriate annotation layers. In order to detect a change phenomenon we compare the frequencies of a variant of a given variable across the language stages. We choose one language stage (here New High German) as the reference frequency and calculate how frequencies in the other language stages differ from this. Here we use the terms *overuse* and *underuse* to describe the deviations.[10]

Since we do not know in advance which variants of a given variable are more or less widespread in each stage, we can initially test all variables in the corpus as an exploratory diagnostic to find the most extreme cases of overuse or underuse. For example, Table 1 shows normalized frequencies for several part-of-speech categories in the different subcorpora. The older stages' frequencies are coded with ▲ to signify overuse (higher frequency) and ▼ to signify underuse (lower frequency) with respect to the NHG subcorpus; the depth of the shading in each cell signifies the extent of the deviation. The same information is coded for each word and each category in the corpus itself so that it can be searched for in conjunction with other annotations. This information can be used as a *diagnostic* for finding interesting change candidates. Wherever we find a word or an annotation category that displays a uniform change pattern (all underuse or overuse, and a deep shade leading to a light shade across time) we can suspect that there could be a uniform, possibly ongoing, change. In Table 1 the categories VVINF (infinitive verbs) and PRELS (relativizers) show such a pattern. The seemingly gradient change of PRELS is directly related to the object of our case study of relative clauses, and forms the starting point for our study in the following section.

Table 1: Comparison of part-of-speech frequencies in the subcorpora. Underuse and overuse are marked with arrows and progressively deeper shades for stronger deviations with respect to NHG.

[10] Overuse and underuse are defined as statistically significant deviations in frequency as compared to another language stage or stages serving as a control population. This strategy has been employed especially in contrastive inter-language analysis (CIA), a paradigm comparing texts from language learners with different native tongues and native speakers (see Selinker 1972, Granger/Hung/Petch-Tyson 2002). The post-hoc nature of underuse/overuse diagnostics means that their results are not as compelling as pre-hoc hypothesis testing, but ideally results from such studies can then be tested in further data sets (for an underuse study of learner German along these lines see Zeldes/Lüdeling/Hirschmann 2008)

Pos	OHG	MHG	ENHG	NHG
PDAT	▲0.046131	▲0.011679	▼0.007105	0.008954
PPER	▲0.083545	▼0.052759	▲0.075916	0.075825
ART	0	▲0.07934	▲0.065445	0.061835
VVINF	▼0.01126	▼0.015707	▼0.018325	0.022104
PRELS	▼0.009444	▼0.011679	▼0.013837	0.016788
VAFIN	▼0.03705	▼0.035038	▲0.04786836	0.045887
VAINF	▼0.001453	▼0.001208	▲0.00411369	0.003078

However, these observations do not yet supply an interpretation of the data. To explain the development of one variant we must understand the variants with which it competes. Table 1 also shows the frequencies of articles (ART), which are rather frequent in NHG but not present in OHG.[11] Like many of the older Indo-European languages, German developed a definite article from its demonstrative stem *d-* (akin to Eng. *th-* in *the* and *this*) and an indefinite article from the numeral *ein-* 'one' (on the development of the German articles see Oubouzar 1992). In OHG, these forms are only just forming, with many nominal phrases having no article where one would be expected in NHG (1), while other cases have a corresponding demonstrative (with the tag PDAT) which can still be interpreted as such (2):

(1) Hench 1890, ch. I, line 18[12]

mannes	sunu	habet	gauualt	in	herdhu	za	forlazanne	suntea
man's	son	has	power	in	earth	to	forgive	sins

the son of man has the power on earth to forgive sins

(2) Hench 1890, ch. I, line 8

enti	gasah	iesus	iro	galaupin	quhad	dem	lamin
and	saw	Jesus	his	faith	said	this	paralytic

and Jesus saw his faith [and] said to this paralytic

The annotation scheme of the OHG corpus considers all such determiners to be demonstratives when they are present, thus the overuse of the PDAT tag in the OHG column in Table 1 directly expresses researchers' interpretation of the data. In the MHG and ENHG corpora article use is similar to NHG (slightly overused), and PDAT also behaves similarly, meaning article use is quantitatively comparable for all these periods. In other words: The category ART can be interpreted as a variable with several variants

[11] The situation is more complicated. For this article it suffices to say that the (few) forms that look like articles in OHG are often analyzed as demonstratives. The annotation here follows this analysis.

[12] All citations in Hench 1890 refer to the Gospel According to Matthew in this edition

(the specific forms of articles in the different language stages) if one wants to see the development of article forms. If one wants to find out about how the category ‘article’ evolved one has to assume a more abstract variable (something like ‘pre-nominal determiner’) with the variants ART, PDAT, Ø etc. Because the empty form is one of the variants of this variable, it is not possible to observe this directly from the part-of-speech annotation. It is possible to solve this problem by looking at the syntactic environment for article occurrence. If a category such as this turns out to be interesting in retrospect, it is possible to add an annotation layer especially for it (in our case, since we have syntactic annotation, we do not need to do so, as we could phrase a query for articleless nominal phrases using the syntactic environment).

5. Examining underuse close up: Relative clauses

In this section we discuss how the phenomena diagnosed by rough underuse / overuse statistics can be evaluated more precisely using the rich annotation in the DDB corpora. Although the corpora at hand are extremely small for a quantitative study, comparisons with previous work on these phenomena will show the results to be plausible, while at the same time they provide estimates for the relative quantifications of competing variants, charting a gradual development in features sometimes thought to be categorical properties of particular language stages.

Categorically the development of relative clauses in German seems not especially interesting.[13] From Old High German (OHG) to Modern German (NHG) we find relative clauses in the form in (3) where the relative clause is introduced by a relative pronoun and the word order in the relative clause is that of a subordinate clause (V-final). They are sometimes considered to be the oldest dependent clauses in German (e.g. Schmidt 2004: 235). Some researchers argue that with respect to relative clauses German has not changed.

(3) Acts 1:18 (Vanheiden 2003)

Von	der	Belohnung,	die	er	für	seine	Untat	bekam,	wurde	dann	in	seinem	Namen	ein	Acker
from	the	award	that	he	for	his	misdeed	received	was	then	in	his	name	a	field

From the award that he received for his misdeed a field was bought.

Quantitatively, however, as the numbers in Table 1 suggest, there might be an interesting development. The category PRELS gradually increases over time. Does this mean that relative clauses become more frequent? If so, one would expect that they extend their domain over time, either grammatically or functionally. This will be discussed in Section 5.3 but first we need to consider a number of surface properties of relative clauses that might have a bearing on the numbers in Table 1.

5.1. Normalization

[13] For more comprehensive overviews of relative clauses in German see Lehmann (1984), Zifonun (2001), or Pittner (2009).

The data in Table 1 is normalized per token: The older language stages show an underuse of PRELS (relative pronouns) per token. However if we are interested in the occurrence of relative clauses in each period and the ways in which they are realized, this is misleading. The token-based normalization may be inappropriate in this case, since it depends on the length of sentences, and not on how many sentences in fact contain relative clauses. Using the syntactic annotation, we can establish how many PRELS appear per 100 clauses (Table 2).

Table 2: Frequencies for PRELS in a clause-based normalization

Subcorpus	**PRELS per 100 clauses**
OHG	4.62
MHG	10.25
ENHG	12.85
NHG	13.35

Normalized to clauses, PRELS appear with roughly the same frequency from MHG to NHG and only OHG shows a significantly (p<0.001) lower number of PRELS. In part, this is due to very short sentences in the OHG text (e.g. imperatives like *enti see saar* "and behold!"), which boost the amount of clauses per token compared to the other corpora.

5.2 Relativizers: Variable and variants

Not only PRELS can introduce relative clauses but also interrogatives and other elements[14], labelled as PWAVs in the tagset as in (4).

(4) Acts 3:10 (Vanheiden 2003)

Sie	wunderten	sich	über	das,	**was**	mit	ihm	geschehen	war
they	wondered	themselves	about	that	what	with	him	happened	was

They wondered at what had happened to him.

PWAVS are overused in the earlier language stages. So is it the case that perhaps in earlier language stages relative clauses were introduced by PWAVS rather than by PRELS?

Another surface phenomenon that might have relevance for the overall number of relative clauses is the fact that in older language stages (up to ENHG) we find asyndetic relative clauses, similar to English asyndetic relative clauses, as in (5) (for an overview

[14] See Pittner (2009) for an overview of the different forms of relativizers, Fleischmann (1973, 115 ff) for an overview of d-pronouns and w-pronouns, and Ágel (2010) for interesting observations on the relativizers *so* and *wo*.

of the different asyndetic forms see Gärtner 1981).[15] The head of the NP dominating this clause is the pronoun *dem*, which is in the dative (cf. the morphological annotation above the tree in Figure 5 below), marking its role in the main clause; the subject role of its referent in the subordinate clause is not marked explicitly, corresponding roughly to English: *…(he) said to [NP them [RC were there]]*. Such constructions are generally known and sometimes described controversially (cf. e.g. Schrodt 2004: 174f).

(5) Hench 1890, ch. XXIII, line 10

enti	quad	za	dem	dar	uuarun
and	said	to	them	there	were

and said to them who were there

These surface phenomena suggest that the relevant variable is not 'part of speech' but 'relativizer', the variants being PRELS, PWAV and Ø. The new variable cannot simply be inferred from the available part-of-speech annotation: PWAV has other readings and Ø cannot be found at all in the POS annotation.

Here we can use the syntactic annotation layer: relative clauses are marked with the edge label RC (for Relative Clause) in the syntactic trees, as in Figure 3.

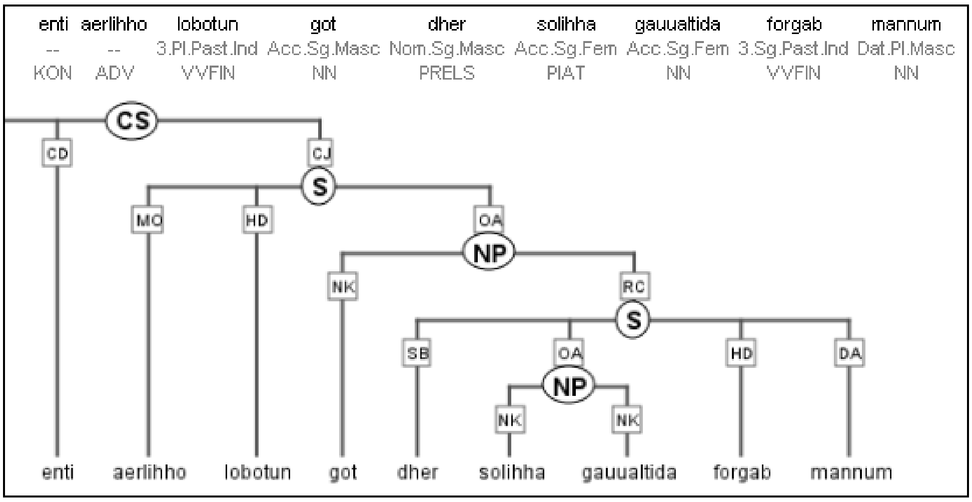


Figure 3: OHG relative clause with a relative pronoun *dher* as a subject: *enti aerlihho lobotun got dher solihha gauualtida forgab mannum* "and indeed they praised God who gave such power to men". (Hench 1890, ch. I, line 22)

[15] The word 'asyndetic' is used for different phenomena. We use it here only for those relative clauses that have no relativizer. Relative clauses that have no reference NP or PP are called free relative clauses. Free relative clauses occur in standard NHG but not asyndetic RCs. There are NHG dialects that have asyndetic relative clauses, see Fleischer (2005).

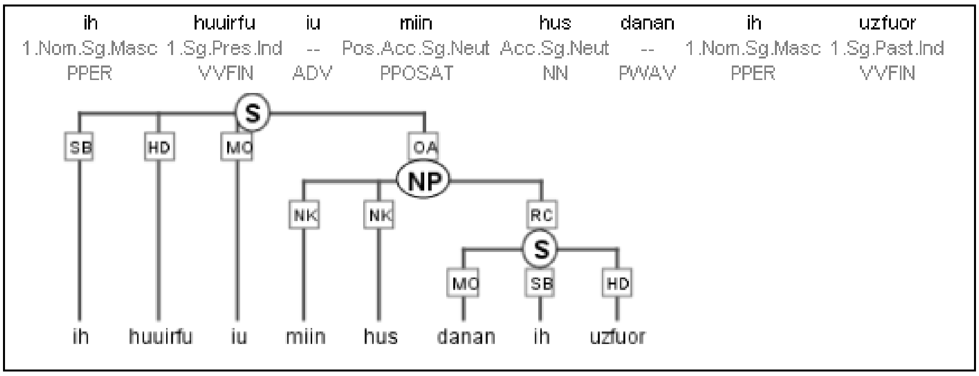


Figure 4: OHG relative clause with uninflected pronominal adverb: *ih huuirfu iu miin hus danan ih uzfuor*, lit. “I return now to my house, whence I departed”(Hench 1890, ch. VII, line 13)

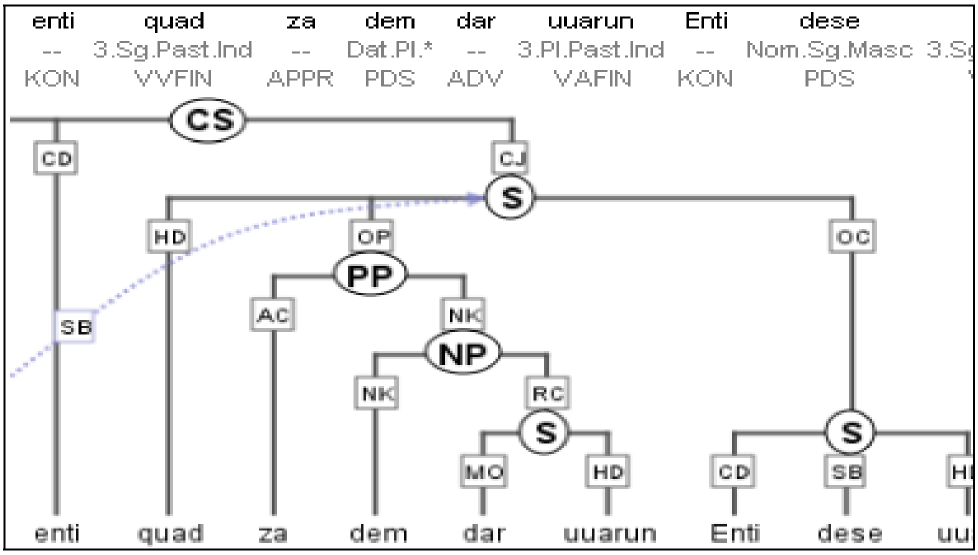


Figure 5: OHG asyndetic relative clause: *enti quad za dem dar uuarun*, lit. “and [he] said to them were there”.(Hench 1890, ch. I, line 22)

Figure 3 shows an OHG relative clause with PRELS at the left edge of the clause. This structure corresponds to the structure of NHG relative clauses. Structures like the one in Figure 4 which are introduced by an uninflected pronominal adverb (PWAV) are also found in all language stages (see e.g. Paul 2007, 405ff). The marked clause in Figure 5 is asyndetic.

Recall that we began with the observation that relative pronouns (PRELS) are underused in the older language stages. We then saw that our diagnostics used a wrong normalization and the wrong variable. We corrected both mistakes by making reference to multiple annotation levels. If we now look at the variable RC across the language stages we find that there is still a significant (p<0.005) underuse of relative clauses per 100 clauses (Table 3). In Section 5.3 we explore possible reasons for this.

Table 3: Frequencies of RC types in each period. Figures in brackets are normalized per 100 clauses.

	with PRELS	**with PWAV**	**asyndetic**	**total RC**
OHG	**26 (4,62)**	**1 (0,18)**	**12 (2,13)**	**39 (6,93)**
MHG	**30 (10,60)**	**6 (2,12)**	**0**	**36 (12,72)**
ENHG	**34 (11,81)**	**3 (1,04)**	**0**	**37 (12,85)**
NHG	**61 (13,35)**	**2 (0,44)**	**0**	**63 (13,79)**

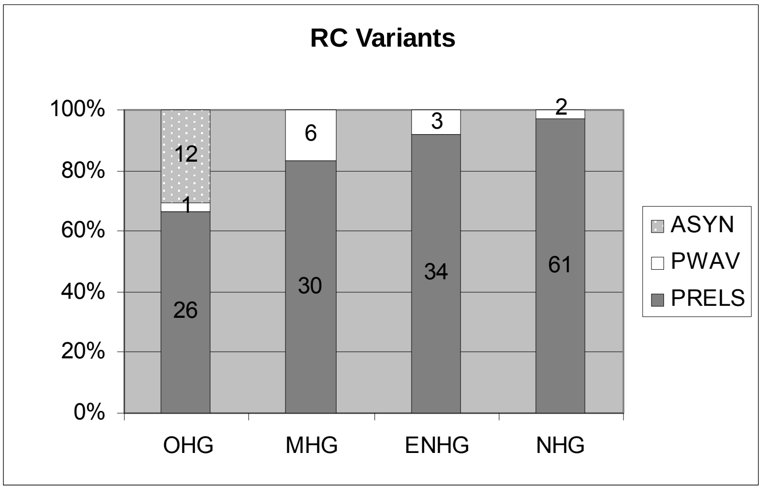


Figure 6: Proportion of RC types in each period

Additionally we see that the proportions of the variants of the variable ‘relativizer’ have changed over time (Figure 6). It has always been dominated by the variant with an inflecting relative pronoun (PRELS), though the latter has clearly become more dominant over time at the expense of both asyndetic clauses, which are attested only in OHG, and PWAV, which is not available in standard NHG except in adverbial clauses, e.g. *ein Ort, wo …* ‘a place where…’ beside the PRELS variant *ein Ort, an dem* ‘a place in which’. This quantitative description expands and complements categorical descriptions in previous accounts of relative clauses, and reference grammars which often give little or no idea of the absoluteness or rapidity of change in such structures, but rather just list attested constructions in every period. What would be needed now is an analysis of the parameters that influence the choice between variants in each stage. For this study our corpus is unfortunately too small.

5.3. Expansion of Relative Clauses?

The significant increase of relative clauses between OHG and the newer language stages could be due to a number of factors. It could be the case that relative clauses extended their semantic, syntactic, or information structural functions. It could also be the case that relative clauses already had all the possible functions in OHG but the same functions were more frequently realized by other variants. To test these alternatives one would again have to define variables with their variants, ideally marking everything directly in the corpus. We will show this for the syntactic functions below. Semantic variables are more difficult to define and annotate. The distinction between restrictive and appositive modification is a notoriously difficult one and many authors have suggested that they should have different syntactic analyses.[16] The syntactic scheme chosen for the analysis of our test corpora does not make this distinction. In order to analyse the semantic function we would have to add annotation layers. We will not go into this issue here in depth but rather state that already in OHG there are clearly restrictive relative clauses (6) as well as clearly appositive relative clauses (7).

(6) (Hench 1890, ch. IV, line 24)

huuelih	iuuuer	ist	der	man	der	ein	scaf	habet	…
who	(of)you	is	the	man	that	a	sheep	had	

who of you is the one, who had a sheep …

(7) (Hench 1890, ch. I, line 23)

aerlihho	lobotun	got	dher	solihha	gauualtida	forgab	mannum
truly	praised	god	who	such	power	gave	man

they truly praised god who gave such power to mankind

Finally we would like to examine another hypothesis that might help us understand the extension of the frequency of relative clauses between OHG and later stages in German, by looking at the reference word/phrase of the variable RELATIVIZER (generalizing over the variants PRELS, PWAV and Ø). Could it be the case that the grammatical possibilities of the relativizers have changed? Here we need to check (a) the part-of-speech of the reference word, (b) the phrasal category of the reference phrase and the syntactic function of the reference phrase. All of this information is already present in the annotation (distributed over several annotation layers). The absolute numbers or the numbers normalized per tokens or clause are not conclusive (Table 4). Relativizers seem to be able to refer to any reference element in any language stage. There are interesting ‘outliers’ here and there but no continuous development to include or exclude a category over time.[17]

[16] Various syntactic analyses have been suggested, the basic idea being that restrictive relative clauses should be adjuncts to N’ while appositive relative clauses are adjuncts to NP (or DP).

[17] Some of the rows point to interesting developments nevertheless (why are there so few relative clauses modifying demonstrative pronouns in NHG?, why do we have so many relative clauses modifying

Table 5 shows that the corpora in DDB contain no relative clauses modifying sentences or VPs, all relativizers refer to either an NP or a PP. Here we do find an interesting difference between NHG and the other language stages: NHG uses significantly (p<0.005) more relative clauses to modify PPs than the others.

Table 4: Categories of the reference elements (absolute / normalized per 1000 tokens / normalized per 1000 clauses, normalized numbers rounded)

	OHG	MHG	ENHG	NHG
NN (noun)	14 / 3,9 / 26	21 / 8,5 / 74	12 / 4,5 / 42	22 / 6,2 / 48
PDS (demonstrative pronoun)	17 / 4,7 / 31	9 / 3,7 / 31	10 / 3,7 / 34	7 / 2 / 15
NE (proper name)	2 / 0,6 / 4	1 / 0,4 / 4	0 / 0 / 0	5 / 1,4 / 11
PPER (personal pronoun)	0 / 0 / 0	1 / 0,4 / 4	3 / 1,1 / 10	1 / 0,3 / 2
PWS (interrogative pronoun)	1 / 0,3 / 2	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0
PIS (indefinite pronoun)	4 / 1,1 / 7	2 / 0,8 / 7	4 / 1,5 / 14	12 / 3,6 / 26

Table 5: Categories of the reference phrase (absolute / normalized per 1000 tokens / normalized per 1000 clauses, normalized numbers rounded).

	OHG	MHG	ENHG	NHG
NP	37 / 10 / 67	31 / 12 / 109	27 / 10 / 93	34 / 10 / 74
PP	2 / 0,5 / 4	5 / 2 / 18	5 / 1,9 / 17	13 / 3,6 / 28
S/VP	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0

We have shown that relativizers in OHG already have all the functions and syntactic possibilities they have in later language stages. The only conclusive quantitative difference to NHG (modification of PPs) is small and cannot explain the numbers in Tables 1 and 2. We again need to step back and pursue another road. We need to look at the basis for our numbers. Could it be that relative clauses have not changed much but something else, namely the *chance* to use a relative clause, has changed? In Table 5 we saw that in all our corpora relative clauses modify NPs and PPs. Does OHG simply have fewer NPs and PPs and thus fewer chances for modification by relative clause? If we look at ‘chances’ per 100 tokens we see a significant difference between OHG and the other language stages (Table 7).

Table 7: Chances for relative clauses / tokens

	OHG	MHG	ENHG	NHG
NP+PP/Token	0,140926641	0,207813129	0,183688739	0,19613878

indefinite pronouns in NHG? etc.). In the absence of continuous developments the observed differences point to the fact that the corpora (even if the parameters are kept constant) show idiosyncracies.

The difference in Table 7 (together with the difference in Table 5) finally explains why OHG differs quantitatively from the other language stages: the main environment licensing relative clauses is itself less common in that subcorpus.

To summarize: we started from a seemingly continuous development (relative pronouns increase over time) in Table 1. Only by looking at different variables and variants and different normalization bases we were able to see that the development was not continuous at all and has (almost) nothing to do with relative clauses. It is an epiphenomenon of a different development (change in the frequency of NPs and PPs between OHG and the later language stages). At the same time, the rising dominance of the PRELS relativizer variant is evident and can be charted gradually and neatly to its status today, after supplanting asyndetic clauses and gaining some of the functions previously available to PWAV.

6. Conclusion

In this paper we showed how a deeply annotated diachronic corpus can help to detect and study language change. As has long been known and studied, language change is always gradual – at each given point in time a linguistic variable can be expressed by several variants. The questions to be studied are (a) What is a linguistically interesting variable? (b) What are the relevant variants? (c) How are the variants distributed synchronically and what triggers the use of each variant? (d) How does the distribution of variants change over time and which features trigger this change?

It has often been shown (see e.g. Rissanen 2008, Kroch 2001) that corpus studies – where variants can be studied in their context – help in answering all of these questions. In our paper we looked at questions (a), (b) and (d) and discussed the choice of variables and variants in a small sample study. As mentioned above, a variable is always an abstraction over several variants. We have shown that in the course of a study we need to change the level of abstraction in order to understand the phenomenon. While this is probably done in most diachronic studies, it usually remains implicit. In a multi-layer corpus this can be made explicit. This becomes especially clear in those studies that involve categories that have no overt exponent.

A further insight is gained from the use of our pilot corpus as testing grounds for a specific multi-layer annotation scheme for historical High German. With the experience gathered in defining the interesting variables in each period and annotating them in a comparable way across time we can now turn to the development of further resources using the same scheme. The next step is therefore to build further corpora, in which we can test our hypotheses pre-hoc and confirm the validity of our analyses by confronting them with the data. Then we can use quantitative methods to find change candidates and pursue different hypotheses. We saw in our study that one crucial factor was the normalization base. In a multi-layer corpus we can add different normalization bases (tokens, clauses, chances) at any point in the analysis.

References

All URLS were checked on June 28, 2010.

Ágel, V. 2010. “+/- Wandel. Am Beispiel der Relativpartikeln *so* und *wo*”. Bittner, D. and L. Gaeta (eds.) *Kodierungstechniken im Wandel. Das Zusammenspiel von Analytik und Synthese im Gegenwartsdeutschen*. Berlin: de Gruyter, 199-222.

Baroni, M. and S. Bernardini. 2006. “A new approach to the study of translationese: Machine-learning the difference between original and translated text”. *Literary and Linguistic Computing* 21(3), 259-274.

Biber, D. 1988. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. 2009. “Multi-Dimensional Approaches”. Lüdeling, A. and M. Kytö (eds.) *Corpus Linguistics. An International Handbook*. Vol 2. Berlin: Mouton de Gruyter, 822-855.

Brants, S., S. Dipper, S. Hansen, W. Lezius and G. Smith. 2002. “The TIGER Treebank”. *Proceedings of TLT-02*. Sozopol, Bulgaria.

Carletta, J., S. Evert, U. Heid, J. Kilgour, J. Robertson and H. Voormann. 2003. “The NITE XML Toolkit: Flexible Annotation for Multi-modal Language Data.” *Behavior Research Methods, Instruments, and Computers* 35(3), 353-363.

Demske, U. 2007. “Das MERCURIUS-Projekt: Eine Baumbank für das Frühneuhochdeutsche.“ *Sprachkorpora: Datenmengen und Erkenntnisfortschritt*. Kallmeyer and Zifonun (eds). Berlin: Walter de Gruyter, 91-104.

Dipper, S. 2005. “XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation”. *Proceedings of Berliner XML Tage 2005 (BXML 2005)*. Berlin, Germany, 39-50.

Evert, S. 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart. Online at http://elib.uni-stuttgart.de/opus/volltexte/2005/2371/.

Evert, S. 2006. “How random is a corpus? The library metaphor”. *Zeitschrift für Anglistik und Amerikanistik* 54(2), 177-190.

Fleischer, J. 2005. “Relativsätze in den Dialekten des Deutsche: Vergleich und Typologie.“ *Linguistic Online 24/3(05)*, 171-186. Online at: http://www.linguistik-online.com/24_05/fleischer.html

Fleischmann, K. 1973. Verbstellung und Relieftheorie. Ein Versuch zur Geschichte des deutschen Nebensatzes. München: Wilhelm Fink Verlag.

Garside, R., Leech, G., and McEnery, A. (eds.). 1997. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman.

Gärtner, K. 1981. “Asyndetische Relativsätze in der Geschichte des Deutschen”. *Zeitschrift für Germanistische Linguistik* 9(2), 152–163.

Granger, S., J. Hung, J. and S. Petch-Tyson (eds.). 2002. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins.

Greenbaum, S. and G. Nelson. 2009. *An Introduction to English Grammar*. 3rd edition. London: Pearson.

Kilgarriff, A. 2001. “Comparing Corpora”. *International Journal of Corpus Linguistics* 6 (1), 1-37.

Klein, D. and C.D. Manning. 2003. “Accurate Unlexicalized Parsing”. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 423-430.

Kroch, A. (this volume) *Using parsed corpora to compare the evolution of word order in English and French*. XXX

Kroymann, E., S. Thiebes, A. Lüdeling and U. Leser. 2004. *Eine vergleichende Analyse von historischen und diachronen digitalen Korpora*. Technical Report 174 des Instituts für Informatik der Humboldt-Universität zu Berlin. Online at http://www2.informatik.hu-berlin.de/Forschung_Lehre/wbi/publications/2004/tr174_corpora.pdf

Labov, W. 1966. *The Social Stratification of English in New York City*. Washington, D.C.: Center for Applied Linguistics, 2006. Second edition: Cambridge: Cambridge U. Press.

Labov, W. 2001. *Principles of Linguistic Change. Vol 2: Social Factors*. Oxford: Blackwell.

Labov, W. 2004. “Quantitative Analysis of Linguistic Variation.” Ammon, U., N. Dittmar, P. Mattheier and P. Trudgill (eds.) *HSK Sociolinguistics/Sociolinguisik Vol I.* Berlin: de Gruyter, 6-21.

Leech, G. 1993. “Corpus Annotation Schemes”. *Literary and Linguistic Computing* 8(4), 275-281.

Leech, G., N. Smith, M. Hundt and C. Mair. 2008. *Changes in Contemporary English: a Corpus-Based Study*. Cambridge: Cambridge University Press.

Lehmann, C. 1984. *Der Relativsatz. Typologie seiner Strukturen – Theorie seiner Funktionen – Kompendium seiner Grammatik*. Tübingen: G. Narr

Lüdeling, A., T. Poschenrieder, L.C. Faulstich. 2005. “DeutschDiachronDigital - Ein diachrones Korpus des Deutschen.“ *Jahrbuch für Computerphilologie 2004*, 119-136. Online at http://computerphilologie.uni-muenchen.de/ejournal.html.

Mair, C. 2009. “Corpora and Study of recent Change in Language”. *Corpus Linguistics. An International Handbook*. Vol 2. Lüdeling & Kytö (eds). Berlin: Mouton de Gruyter, 1109-1125.

Kommentar [l1]: Comment to the editor : This title will need to be checked – this was the title of Kroch’s talk at the conference

Pittner, K. 2009 “Relativum“. Hoffman, L. (ed.) *Handbuch der deutschen Wortarten*, Berlin: de Gruyter, 727-758.

Paul, H. 2007. *Mittelhochdeutsche Grammatik.* (Revised by T. Klein, H.-J. Solms and K.-P. Wegera.) Tübingen: Niemeyer.

Resnik, P., M. Broman Olsen and M. Diab. 1999. “The Bible as a parallel corpus: Annotating the “book of 2000 tongues””. *Computers and the Humanities* 33, 129-153.

Rissanen, M. 2008. “Corpus Linguistics and Historical Linguistics.” Lüdeling, A. and M. Kytö (eds.) *Corpus Linguistics. An International Handbook*. Vol 1. Berlin: Mouton de Gruyter, 53-68.

Sag, I. and T. Wasow. To appear. Performance-Compatible Competence Grammar. Borsley, R. D. and K. Börjars *New Models of Grammar*. Blackwells. Manuscript available at: http://www.stanford.edu/~wasow/procpap2.pdf.

Schmid, H. 1994. “Probabilistic Part-of-Speech Tagging Using Decision Trees”. *Proceedings of the Conference on New Methods in Language Processing*. Manchester, UK

Schmidt, W. 2004^{9}. *Geschichte der deutschen Sprache*. Stuttgart: S. Hirzel Verlag.

Schrodt, R. 2004. *Althochdeutsche Grammatik II.* Tübingen: Niemeyer.

Wasow, T. 2007. Gradient Data and Gradient Grammars. *Proceedings of the 43rd Annual Meeting of the Chicago Linguistics Society*, 255-271.

Zeldes, A. 2007. Machine Translation between Language Stages: Extracting Historical Grammar from a Parallel Diachronic Corpus of Polish. *Proceedings of Corpus Linguistics 2007, Birmingham, 27-30 July, 2007*. Online at: http://www.corpus.bham.ac.uk/corplingproceedings07/paper/60_Paper.pdf.

Zeldes, A., A. Lüdeling and H. Hirschmann. 2008. “What's Hard? Quantitative Evidence for Difficult Constructions in German Learner Data”. *Quantitative Investigations in Theoretical Linguistics 3 (QITL-3)*. Helsinki, Finnland, 2-4 June 2008.

Zeldes, A., J. Ritz, A. Lüdeling and C. Chiarcos. 2009. “ANNIS: A search tool for multi-layer annotated corpora”. *Proceedings of Corpus Linguistics 2009.* July 20-23, Liverpool, UK.

Zeldes, A., J. Richling, H. Hirschmann and A. Lüdeling. Submitted. “ Measuring Syntactic Change: Underuse and Overuse Statistics in a Multi-Layer Historical Corpus of German”

Zifonun, G. 2001. Grammatik des Deutschen im europäischen Vergleich: Der Relativsatz. Mannheim: Institut für Deutsche Sprache.

Corpus Editions

OHG: Hench, G.A. 1890. *The Monsee Fragments*. Strassburg: Karl J. Trübner

MHG: Mellenbourn, G. 1944. *Speculum ecclisiae*. (Lunder Germanistische Forschungen 12.) Lund: Gleerup.

ENHG: Diel, M., B. Fisseni, W. Lenders and H.-C. Schmitz. 2002. *XML-Kodierung des Bonner Frühneuhochdeutschkorpus*. Bonn: IKP-Arbeitsbericht NF 02.

NHG: Vanheiden, K.-H. 2003. *Neue evangelistische Übertragung*. Available from: http://www.kh-vanheiden.de/.

Index

Index of Proper Nouns (about 5 items) : Humboldt-University Berlin, DeutscheDiachroneBaumbank (DDB)

Names (about 5 items) : LABOV, W.

Index of Subjects (about 10 items): annotation, multi-layer corpus architecture, relative clause, variationism, treebank, overuse/underuse

INFORMATION

Anke LÜDELING

Professor, Institute for German Language and Linguistics, Humboldt-University Berlin

Lüdeling, A. and M. Kytö (eds.) 2008/2009 *Corpus Linguistics. An International Handbook.* 2 Vols. Berlin: Mouton de Gruyter.

Lüdeling, A. and A. Zeldes. 2008 Three Views on Corpora: Corpus Linguistics, Literary Computing, and Computational Linguistics. *Jahrbuch für Computerphilologie 9*, 149-178 . Online at http://computerphilologie.tu-darmstadt.de/jg07/luedzeldes.html

Hagen HIRSCHMANN

Researcher, Institute for German Language and Linguistics, Humboldt-University Berlin

Hirschmann, H. to appear „Eine für Korpora relevante Subklassifikation adverbieller Wortarten". Konopka, M., J. Kubczak, C. Mair, F. Štícha and U.H. Waßner. (eds.), *Grammar & Corpora / Grammatik und Korpora 2009. Third International Conference / Dritte Internationale Konferenz, Mannheim, 22.-24.09.2009*. Tübingen: Gunter Narr Verlag.

Hirschmann, H. S. Doolittle and A. Lüdeling 2007. Syntactic annotation of non-canonical linguistic structures. *Proceedings of Corpus Linguistics 2007*, Birmingham. Online at http://www.corpus.bham.ac.uk/corplingproceedings07/paper/128_Paper.pdf

Lüdeling, A., S. Doolittle, H. Hirschmann, K. Schmidt and M. Walter 2008. Das Lernerkorpus Falko. *Deutsch als Fremdsprache 2(2008),* 67-73.

Amir ZELDES

Researcher, Institute for German Language and Linguistics, Humboldt-University Berlin

Lüdeling, A. and A. Zeldes. 2008 Three Views on Corpora: Corpus Linguistics, Literary Computing, and Computational Linguistics. Jahrbuch für Computerphilologie 9, 149-178 . Online at http://computerphilologie.tu-darmstadt.de/jg07/luedzeldes.html

Zeldes, A., J. Ritz, A. Lüdeling and C. Chiarcos. 2009. "ANNIS: A search tool for multi-layer annotated corpora". *Proceedings of Corpus Linguistics 2009.* July 20-23, Liverpool, UK.

Zeldes, A. 2007. Machine Translation between Language Stages: Extracting Historical Grammar from a Parallel Diachronic Corpus of Polish. *Proceedings of Corpus Linguistics 2007, Birmingham, 27-30 July, 2007*. Online at: http://www.corpus.bham.ac.uk/corplingproceedings07/paper/60_Paper.pdf.