Measuring and Coding Language Change: An Evolving Study in a Multi-Layer Corpus Architecture

H. HIRSCHMANN
Humboldt-Universität zu Berlin, Germany
AND
A. LÜDELING
Humboldt-Universität zu Berlin, Germany
AND
A. ZELDES
Humboldt-Universität zu Berlin, Germany

Our paper explores the possibilities of using deeply annotated, incrementally evolving comparable corpora for the study of language change, in this case for different stages from Old High German to New High German. Using the example of the evolution of German past tenses, we show how a variety of categories ranging from low to high complexity interact with the choice between competing linguistic variants. To adequately explore the influence of these categories, we use a multi-layer corpus architecture that develops together with our study. We show that a combination of quantitative and qualitative analyses can recognize relevant contextual factors, which feed into the addition of new annotation layers applying to the same data. By making our categorizations explicit as corpus annotations and our data available to other researchers, we promote an open, extensible and transparent mode of research, where both raw data and the inferential process are exposed to other researchers.

Categories and Subject Descriptors: I 2.7 **[???]** – ??? — ???.

General Terms: ???, ???

Additional Key Words and Phrases: corpus linguistics, multi-layer corpora, historical linguistics, German, tense, perfect, preterit, variation

ACM File Format:
HIRSCHMANN, H., LÜDELING, A. AND ZELDES, A. 2011. Measuring and Coding Language Change: An Evolving Study in a Multi-Layer Corpus Architecture. *ACM Journal of Computing and Cultural Heritage*,

1. INTRODUCTION:

Historical linguistics necessarily relies on corpus data and thus it is no wonder that historical linguists were among the first to discover the usefulness of electronic corpora (e.g. arguably the first electronic corpus, Roberto Busa's Index Thomisticus of the works of Thomas Aquinas, see McEnery & Wilson 2001, 20-22). Today there are many very large and tremendously useful electronic corpora of many historical languages and language stages (e.g. for English the Helsinki Corpus, Kytö 1991 and the Penn-Helsinki Parsed Corpus of Early Modern English, Kroch et al. 2004, the Tatian corpus of Old High German in Petrova et al. 2009, or the corpora in the comprehensive Perseus Digital Library for Classics, Crane 1998, to name but a few). Historical corpora serve many purposes – among them preserving and sharing data that is otherwise difficult to study or making data accessible in formats that allow new research questions (for example by combining qualitative and quantitative data in interesting ways or linking geographic or

other resources to textual data). In this paper we focus on one small aspect of working with historical corpora – the use of corpora for linguistic research. Linguistic research necessarily depends on the interpretation and classification of the data: Linguists usually do not want to speak about single occurrences of a word or phrase but to generalize over several occurrences of whatever they study in order to develop and test models and theories. Typically such linguistic classes are lemmas, parts-of-speech, or sentence types. Historical data is even more difficult to classify than modern data. For many historical texts (if they are for example written in scriptio continua) even the division into word forms is an interpretation. The interpretation can be coded along with the data as annotation. The explicit and available coding of annotations with the data allows other researchers to understand and follow an analysis. Results become reproducible – a huge step forward from the sometimes unclear and 'private' analysis of many historical studies. Many available historical corpora are, however, not annotated at all, while others come with a very specific, closed set of linguistic annotation layers. Sometimes they use a proprietary search tool and sometimes the data is not available for manipulation by researchers (often it can only be accessed through a Web interface and not be downloaded). This situation is problematic for several reasons.

Historical corpora that cannot be modified by the researcher are problematic for different reasons, e.g.:

(a) There are no linguistic classifications that are shared by all linguists. Thus, the next researcher might want to classify the same type of information (e.g. parts of speech or sentence types or even tokens) in a different way.

(b) In the course of a study it is often necessary to assume classifications or annotation layers that are specific to one research question and not annotated in the given corpus.

(c) The researcher wants to compare the corpus at hand with another corpus that has different annotation layers and wants to add the appropriate layer for comparison.

This is true for both qualitative and quantitative studies, since quantities necessarily rely on underlying categorizations (see Biber & Jones 2009 for a discussion of different types of quantitative studies). In each case where existing annotations reach their limit, the researcher is forced to perform the analysis separately, away from the corpus. This means that the analysis is not available for further study and that the results of the study are not reproducible.

If a researcher wants to add annotation layers he/she needs (a) full access to the data in a well-described, standardized format and (b) a flexible corpus architecture that can handle the addition of annotation layers, as well as (c), a powerful search tool to query these layers in conjunction. There is currently no tool which can handle all possible types of annotation (token-based annotation, spans, trees, pointers), and even if there were such a generic tool many researchers would still prefer to use a specific annotation tool which is optimized for their type of annotation and that they feel comfortable with. In this paper, we take the approach that corpus annotations should be both dedicated and extensible, by adding and merging data from different annotation tools as it becomes necessary. To this end we present state-of-the-art techniques such as meta-model based conversion of annotation formats into a common standard using the SaltNPepper converter framework (Romary & Zipser 2010) and PAULA stand-off XML (Dipper 2005), and equally complex search facilities using ANNIS2, a flexible browser based search tool for complex annotation graphs representing different types of annotation. We use these tools to build, query and extend an example corpus from different periods of German language which we will use to investigate the development of German past tense forms. Before

presenting the technical aspects of our methodology, the following section sketches out the linguistic research question our study will address, followed by an introduction of the corpus. Subsequent sections will present the corpus architecture and the search tool ANNIS, and finally we will show in several steps how the addition of annotation layers – necessitated by different aspects of the research question – is carried out and used to enrich both our results and the corpus itself.

2. RESEARCH QUESTION AND BACKGROUND

To illustrate our approach to the issues above, we will use an example of language change that can be better understood if we look at several linguistic layers at the same time. The development and competition of German past tenses has syntactic, morphological, semantic, and pragmatic aspects, but all of these are a matter of interpretation. Only if we code our interpretation directly in the corpus – as annotations – will our results be transparent and reproducible.

Modern German (MG) has, according to most grammars, a total of six tenses (compare e.g. Helbig & Buscha 2001, 25ff). Here we want to concentrate on the development of the most common past tense constructions: the preterit and the perfect (for this example study we will ignore the rather marginal German pluperfect). We are interested in both how they have evolved from the earliest documented stages of German language (Old High German, OHG) to Modern German, and what factors determine the use of one of the two tenses in a given context at a given time. The German preterit is formally similar to the English simple past tense and the perfect is formally similar to the English present perfect – but while in English the functions of past simple and present perfect are clearly different, in MG they can be used interchangeably in many contexts and it seems that the perfect is 'taking over' the place of the preterit in current usage. The preterit is formed synthetically (1), while the perfect is analytic, with two different auxiliaries (in (2a) the auxiliary *haben* 'to have' and in (2b) the auxiliary *sein* 'to be'). Simplifying somewhat, the choice of perfect auxiliary mainly depends on the transitivity of the verb; transitive verbs take *haben*, intransitive verbs take *sein*.

(1) *Ich arbeitete.*
I worked
"I worked"

(2a) *Ich habe gearbeitet.*
I have worked
"I worked"

(2b) *Ich bin gestolpert.*
I am stumbled
"I stumbled"

MG has developed from Old High German, which derives from Proto-Germanic, across two intermediate periods (according to the traditional division promoted by Wilhelm Scherer in the 19th century), which are referred to as Middle High German (MHG) and Early New High German (ENHG). In the early stages of OHG only synthetic tenses are attested, which means only one past tense existed – the preterit. However the analytic perfect tense developed already within the OHG period and has coexisted with the synthetic preterit since that time as a competing variant (for the developmental process of the perfect tense in OHG, see Grønvik 1986). Some authors point out that the perfect

tense emerges “at the expense” of the preterit tense (compare e.g. Reichmann & Wegera 1993, 385, or Nübling 2006, 247), which implies a complex development in which the perfect steadily increases its ground, while preterit use constantly decreases. The first perfect constructions in OHG only occur with transitive verbs and with the auxiliary *haben* ‘to have’, but in MHG the perfect tense also occurs with intransitive verbs and the auxiliary verb *sein* ‘to be’, which is still the case in MG. In OHG the auxiliary *haben* has a variant *eigan* ‘to have, own’ which disappears in that period.

At least since MHG times there have thus been two tense forms that can be used primarily to refer to past events (notwithstanding the aforementioned pluperfect or oblique ways of referring to the past, e.g. narrative present), and the question is how these two forms are distributed. It has long been debated whether the distribution is really a semantic one and the readings can be distinguished aspectually.[1] We will return to this hypothesis in Section 5.2. However it seems clear that an aspectual difference cannot explain the newer data, since in many contexts the two past tense variants can be used interchangeably in MG. Most standard grammars claim that the choice between them is guided by ‘formality’ or ‘register’ (e.g. Helbig & Buscha 2001, 129ff) – the preterit is said to be used in written, formal texts, while the perfect is used in speech or informal contexts. This means that in MG the use of the two tenses is almost never conditioned grammatically but depends on pragmatic or even extra-textual factors.[2]

The development of German temporal categories has, of course, been widely researched in previous work (compare e.g. Hilpert 2008 for the development of Germanic future constructions, and the various studies on German past tense constructions referred to in this paper). The different approaches range from purely theoretical to mainly empirical work. A noteworthy study using historical corpus data to test different hypotheses of how the German perfect tense has developed from the 11th to the 16th century is Dentler (1997). The data reported on in Dentler’s paper, however, as well as all other contributions involving empirical corpus data and statistics, are not accessible. Previous research results are therefore not easily reproducible, except by collecting the same sources again, repeating the analysis and counting from scratch, which would inevitably lead to slightly different results (since we do not know how each and every case of the relevant variables was classified in each study). We are not aware of quantitative studies of German past tenses that use electronic corpora and provide the analyzed data. We therefore feel that is important to make our analysis as explicit as possible, even where it matches analyses found in studies predating open-access, multi-layer electronic corpora.

In order to make statements about linguistic phenomena in the different German language stages, we need comparable historical corpora from the respective language stages OHG, MHG, and ENHG, which we can contrast with each other (or with modern data). To perform a reliable, reproducible quantitative analysis we must code relevant linguistic information in annotations within these corpora. The phenomenon we want to discuss has to be described with lexical (e.g. auxiliary form), morphological (e.g. inflectional status of auxiliaries and preterit morphology), syntactic (e.g. analytic verb

[1] We use the term here to refer to grammatical aspect, i.e. the way of ‘viewing the internal temporal constituency of a situation’ (Comrie 1976, 3), and in particular whether the action of a verb is related as completed (‘perfective’) or not. The aspectual reading of the tense opposition in Old German would be similar to the English distinction between past tense and perfect tenses, see e.g. Leiss 1992, 23ff.

[2] Another semantic question – whether the periphrastic perfect tense form has to be analyzed compositionally or non-compositionally (e.g. Musan 2002, 21ff) – will not be discussed in this article.

constructions), and even textual linguistic or pragmatic (e.g. contextual factors for tenses) features. The next section gives an overview of the corpus architecture we use to code these features, followed by a description of the corpora used in this study.

3. TAKING A CLOSER LOOK WITH ANNIS

With the interest in more complex phenomena, the need for similarly complex, richly annotated data arises. This desired richness includes not only simple annotations of individual words, like the part-of-speech information which have been around in electronic corpora for many years, but also syntactic analyses, discourse structural data and a potentially limitless variety of additional linguistic features, possibly specially defined for the research question at hand. Such an extensible collection of annotations not only allows access to frequencies and detailed information about elements and structures that are present, it further offers a deeper comprehension of absent phenomena and their possible substitutes, since it is easier to search for an explicit annotation than for the absence of some phenomenon. Therefore, multi-layered, heterogeneous annotation proves to be a powerful ally in the study of linguistic variation, e.g. the choice between competing tenses.

Ideally, a multi-layer architecture should allow a corpus to grow dynamically with the needs of its users. Different layers of annotation can be developed collaboratively by different researchers with different expertise and integrated into a common multi-layer resource. In order to realize such a flexible architecture we require a facility for altering and extending annotation without disrupting existing data structures. A traditional corpus architecture which adds annotations in-line, i.e. after each word or using XML tags around word forms in a single XML file, makes this difficult, since files cannot be edited easily while hiding already existent annotations and current format structures may be easily disrupted. To circumvent this problem, the concept of stand-off XML (Carletta et al. 2003) has been developed, wherein different annotation levels can be kept in separate XML files pointing at the unaltered raw data. This means that annotation layers which point at data externally can be added and removed, without disturbing other annotations. It is even possible to add several versions of the same type of annotation (e.g. competing syntactic analyses), or structures whose hierarchies conflict, which would be impossible in standard in-line XML. In our case, we use the stand-off format PAULA XML, which is generic and extensible to arbitrary novel annotations (see Dipper 2005 for more details).

Though stand-off XML is a useful tool for representing different layers of annotation, dedicated annotation tools that are comfortable to work with generally use their own format, typically a form of in-line XML. For example, spans of text can be annotated using the tool EXMARaLDA (Schmidt 2004) and saved in EXMARaLDA XML, but this format is unsuitable for the representation of syntax trees, which can be stored in a different format, e.g. Tiger XML (Lezius et al. 2002). At the same time, we wish to work with the different layers simultaneously, for example by running queries which examine correspondences between multiple layers. For this reason, we must use a common meta-model capable of representing all of our data at once. This is realized using the converter framework SaltNPepper (Zipser & Romary 2010 and Zipser 2009), which makes it possible to integrate new import modules which recognize new formats as they become necessary (that is, once we desire a new type of annotation for our data which requires a new annotation tool). Data from different sources can then be merged in the common

meta-model to allow concurrent queries based on multiple annotation layers simultaneously.

In order to search through our data in this way, we use ANNIS (Zeldes et al. 2009, http://www.sfb632.uni-potsdam.de/d1/annis/), a flexible web-based corpus architecture that allows users to query and visualize deeply-annotated data. ANNIS supports search and visualization of annotations applying to tokens, spans of tokens, and generalized directed acyclic graphs (DAGs) with labeled edges (such as syntax trees), as well as arbitrary pointing relations between nodes in the graph, and metadata. Using the query language AQL (ANNIS Query Language), it is possible to address multiple layers of annotation in a wide variety of constellations of graph-topological relations (e.g. annotations encompassing, overlapping or dominating the same text, one another etc., starting and ending at certain points, or connected by labeled edges).

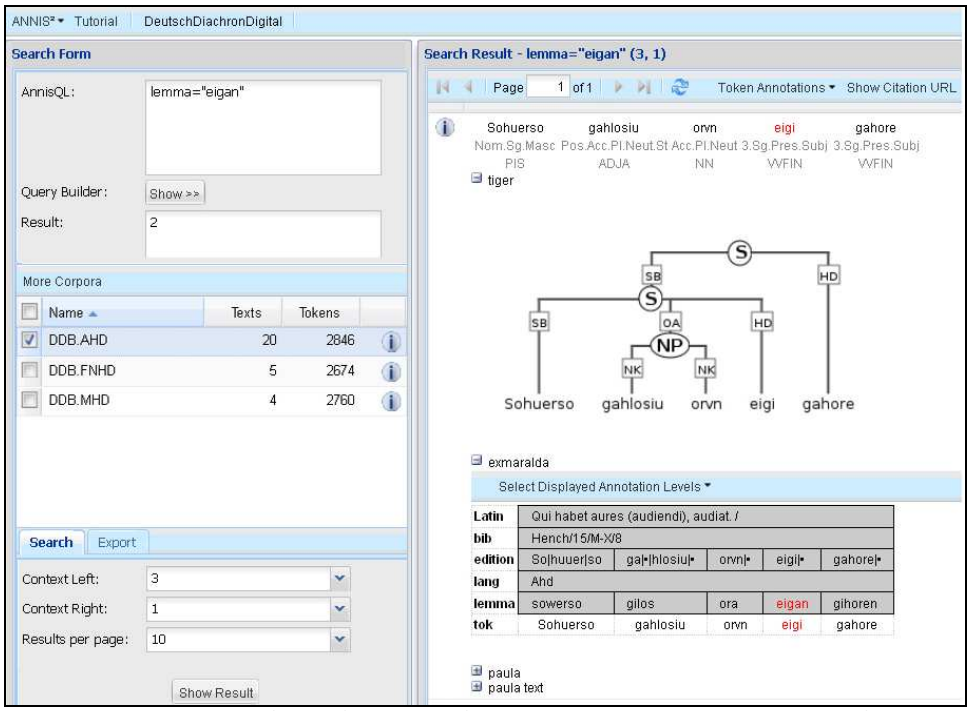

Fig. 1: The ANNIS user interface. Left: selection of corpora (OHG corpus selected) and search form. Right: display of results, with a syntax tree and grid of span annotations expanded for one search result in OHG.

Using multiple annotation layers at once, we can for example take a closer look at the lexical environments in which different tenses in OHG appear while searching for the different types of phrases or sentences in which they are embedded according to the annotated syntactic structure. If we want to separate types of results in a query which cannot be defined by our current annotation, we can then expand our corpus with additional layers describing these types using an appropriate dedicated annotation tool, and the new annotation(s) will be merged with existing data in the multi-layer architecture (see section 5.2). Once suitable annotation levels are made available and queries for relevant structures have been formulated, search results can be exported for inspection, and annotation features of interest can also be exported in a tabular format for

quantitative evaluation (e.g. in the Attribute Relation File Format (ARFF) used in the machine learning tool WEKA, Witten & Frank 2005). Finally, interesting queries can be referenced using deep links, which may be cited in publications to make data reproducible and examinable for other researchers (see examples below).

With this methodology at hand, we now return to our research question. How can we collect data for a quantitative study of the diachronic development of German past tenses? What kind of annotation scheme is suitable for this study? And how can our results enrich the corpus to allow even more accurate explanations of the phenomena we find?

4. CORPUS BASED APPROACH AND CORPUS DATA

In order to trace the development of German past tense constructions empirically we require a suitable comparable corpus, as described above. For this purpose we must make sure not only that our initial annotations reflect categories which have been considered relevant in previous work, but also that the annotation scheme is comparable between language stages. Only by making sure we are counting 'the same thing' in each case can we make qualified assumptions that either confirm or refute traditional accounts. In a first step, we will want to confirm the assumed distribution familiar from older work: the perfect should appear in MHG and gradually gain ground. In a second step, we will want to test in how far the distribution of preterit and perfect constructions in our data correlates with certain aspectual and pragmatic factors.

In this example study we will use three very small, but deeply annotated comparable corpora for OHG, MHG, and ENHG, which form the DDB Treebank (Hirschmann & Linde 2011). The small size of these corpora is of course not ideal for quantitative work, but it makes it possible to develop a dynamic annotation scheme that can easily be extended: ideas can be tested using careful manual annotation and any design decisions can be adapted and carried over to the other language stages with relative ease. If an annotation scheme proves its worth, a larger corpus can then be modeled on the initial sample. Needless to say, results based directly on this corpus thus have a very restricted scope and should only be taken to illustrate possible directions, and more importantly some methodological points of corpus design. The texts we have chosen to annotate are divided into the following subcorpora:

a) Subcorpus Old High German, consisting of a part of the Monsee Fragments (written end of the 8th century), which contain the Gospel of Matthew, based on an edition by George Allison Hench (1890). The subcorpus consists of 2846 tokens.
b) Subcorpus Middle High German, consisting of a collection of Middle High German sermons, called Specculum ecclesiae (written end of the 12th century), based on an edition by Gert Mellenbourn (1944). The subcorpus consists of 2760 tokens.
c) Subcorpus Early New High German, 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 2674 tokens.

Each subcorpus was initially annotated with what we considered to be minimal syntactic and morphological annotations: normalized lemmatization (unified across spelling variants in each period) part-of-speech, inflectional morphology, phrase structure, grammatical functions, and the bibliographical source (position in the manuscript/edition). The corpus and all subsequent annotations described below are available at http://korpling.german.hu-berlin.de/ddd/search.html.

5. EXAMINING THE DATA: AN EVOLVING CORPUS STUDY

In this section we will study the development of German past tenses using the DDB and an incremental approach to multi-layer annotation. Section 5.1 is concerned with the quantitative change patterns of the two tense forms. Here we show that we need the syntactic annotation layer (a graph) as well as part-of-speech and morphological annotation layers and lemmatization (token and span annotations) in order to identify the constructions we are interested in. These annotation layers are not specific to the present study but can be used for many other research questions, increasing the long-term utility of the corpus beyond this study. In Section 5.2 we look at context variables that influence the choice of each form. We show that it is necessary to add annotation layers that are specific to our given research question to the more general annotation scheme of the corpus. This requires use of specialized annotation tools in conjunction with the SaltNPepper framework and the ANNIS search architecture.

5.1 PRETERIT VERSUS PERFECT IN OHG, MHD, AND ENHG

We begin by finding preterit and perfect forms in our subcorpora, in order to find out whether the assumptions stated in Section 2 – a gradual increase of the perfect and a decrease of the preterit – are borne out in our data. The annotations for the synthetic preterit forms can be assigned directly to a token: the annotation scheme for inflectional morphology specifies tense, mood etc. for each verb, allowing us to find all non-auxiliary preterital verbs (for the auxiliary verbs we must consider that they may themselves be part of an analytic construction, though there are of course preterital cases of ‘have’ and ‘be’ in the strong sense). We therefore need part-of-speech annotation and the assignment of inflectional morphology so that we can search lexical verbs in the past tense indicative mood.

The analytic perfect forms can only be found reliably using the syntactic annotation. It is not possible to do this based on morphological forms alone because there are similar forms with different functions. Consider examples (3a-c) from Modern German (comparable problems arise in the older language stages). All three examples in (3) contain a finite auxiliary and a participle. Only (3a) is a perfect. (3b) is a statal passive and (3c) is a predicative construction.

(3a) *Wir sind gekommen.*
we are come
„We have come.“

(3b) *Wir sind geheilt.* (statal passive)
we are cured
“We are cured.”

(3c) *Wir sind verloren.* (predicative)
we are lost

"We are lost."

A purely form-based search or a search for part-of-speech categories in combination with morphological information will therefore lead to many false positives.[3] The underlying structures in (3) do however lead to different syntactic analyses, which are expressed by different structures in our annotation scheme. The syntactic and morphological annotations are closely related to the Tiger annotation scheme (Albert et al. 2003) and the Tiger morphology scheme (Crysmann et al. 2005) of the Tiger treebank (Brants et al. 2002). The necessary query must thus cover several annotation layers.

Figure 2 shows a query for perfect tenses involving the auxiliary *haben* in MHG and how the annotations are visualized in ANNIS.

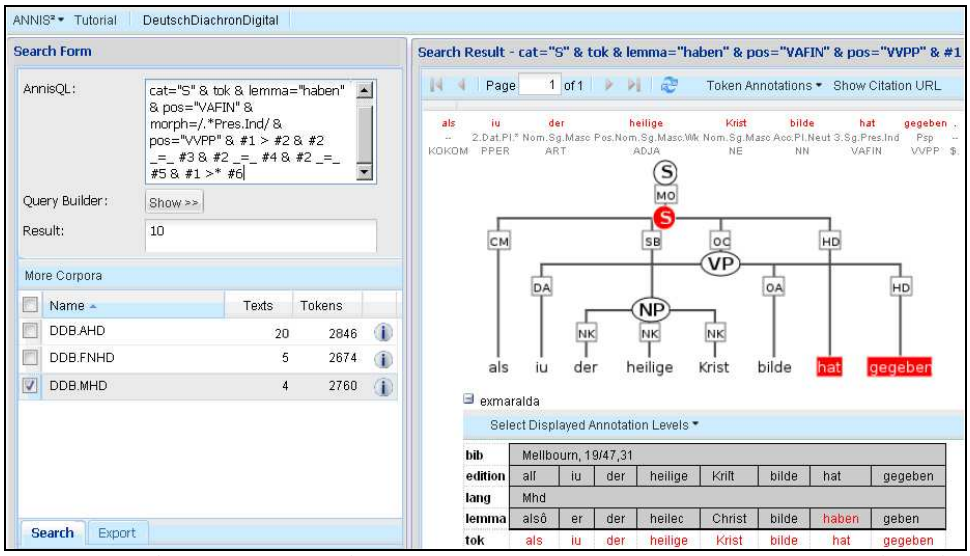

Fig. 2: Example for the visualization of a MHG clause matching a complex search query (a query for perfect clauses with the auxiliary *haben* in the MHG subcorpus: a clause should contain a token which has the lemma *haben* and which at the same time is an auxiliary verb in indicative present; the same clause should contain a past participle main verb). The screenshot shows the search query (top left), the number of hits for the query (below), the selected corpus (below), and one matching clause with various annotations.

The result of each search is given in Table 1. Before looking at the numbers we must make two important methodological points about the base of normalization and query formulation. Very often corpus counts are normalized per *n* tokens (this can be calculated without additional annotations). Token-based normalization is, however, often not appropriate (see Lüdeling et al. to appear). If a variable can be expressed only once per, say, noun phrase (like the form of the determiner) or clause (like finite verbs) the normalization base needs to be the noun phrase or clause, respectively. This requires the appropriate annotations to be available in the corpus, even if they are not directly the subject of our search. The normalization base must be made explicit and motivated

[3] Another problem is that orthographic forms in older German periods vary significantly so that searches for word forms are difficult. For this reason we originally added an annotation layer with orthographically normalized lemmas, allowing an easier search for auxiliary verbs regardless of orthography. It is also possible to normalize lemmas across time periods using hyper-lemmas, i.e. an annotation unifying the OHG verb *wesan* 'to be' with its MHG or ENHG counterparts *sîn* and *sein* respectively. In our study the variants were so few that we searched for the different normalized lemmas instead.

together with the counts – in this case we normalize per clause, since each clause can have its own tense (but not each token).

The second point, query formulation, may seem obvious: the way a query is formulated influences results. In many cases there are several ways of formulating the query and sometimes there is a trade-off between recall and precision. In this case, we aim for high precision by defining perfect constructions as combining a specific auxiliary head (a finite, indicative, present tense auxiliary verb) with a past participle main verb, but we might lower recall by missing e.g. elliptical cases which do not contain both elements but still could be considered perfect constructions (trying to find these using a simple query will inevitably lead to false positives). At this juncture we must decide if these cases are important enough to merit additional annotation (see the next section) or not. In this case we ignore such unclear cases. But the only way to ensure a transparent study is to make our assumptions explicit. We therefore give all queries in the appendix together with embedded links to the search in ANNIS, making our results easily reproducible.

language stage	frequency of preterit constructions	frequency of perfect constructions involving *haben* or *eigan*	frequency of perfect constructions involving *wesan*, *sîn*, or *sein*
OHG	36.1 (203)	0 (0)	0 (0)
MHG	28.3 (80)	3.5 (10)	1.1 (3)
ENHG	2.4 (7)	15.6 (45)	3.5 (10)

Tab. 1: Normalized frequencies of preterit indicative constructions, perfect constructions involving the lemmas *haben* or *eigan* 'to have', and perfect constructions involving the lemmas *wesan, sîn,* or *sein* 'to be' in OHG, MHD, and ENHG (occurrences per 100 clauses; absolute frequencies in brackets).

As Table 1 shows, our expectations for relative tense frequencies are borne out: Preterit forms decrease significantly between OHG and ENHG, while both perfect constructions increase significantly.

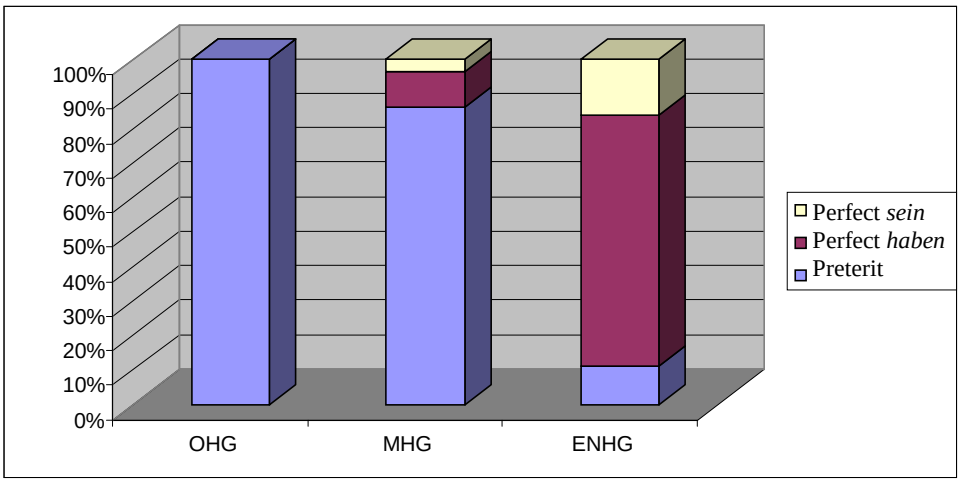

Fig. 3: Distributions of the three different past tense constructions in OHG, MHG and ENHG

Figure 3 – which shows the same data proportionally – illustrates that using corpus data we can see more than the general tendency: We can show *how much* more pronounced change is in each developmental stage (it would be better to have more data points, of course, annotated in the same scheme): While there are no attested perfect constructions in the 8th century, 400 years later (in MHG times) both perfect constructions that Modern German has today exist, but they are very rare in our data (only about 10% of all past tense constructions). In the ENHG text this ratio between preterit and perfect is almost reversed (only about 15% of all past tense constructions are realized in the preterit). This dramatic change occurs in a period that is actually shorter than the first period.[4] Since then the rate of change has slowed down – Modern German still uses the preterit.

The results lead to further questions: Why are perfect constructions in the MHG text present, but used very rarely? If they are used so rarely they can generally be considered highly marked. What licenses these marked constructions? In which contexts do they occur? We know that in Modern German perfect and preterit use is at least partly dependent on register, but is this already the case in MHG and ENHG?

5.2 A CLOSER LOOK AT CONTEXTUAL FACTORS

We have seen that for linguistic research it is necessary to annotate the data in question on several levels. We have also seen that we need different data formats (token-based annotation, syntax trees etc.) which can only be combined and searched in a suitable format (see above). The annotation layers we talked about so far (part-of-speech, lemma, morphological analysis, syntax) are not unusual[5] and can be used for many studies. In those layers every token is annotated. To answer the very specific research questions formulated at the end of the previous subsection, it is necessary to add annotation layers that are tailored to these questions and only pertain to some of the tokens or clauses.

In order to illustrate this we test two hypotheses about the distribution of perfect and preterit verbs in our data. The first hypothesis is that the perfect is used in communicative contexts and the preterit is preferred in narrative contexts (see also Dentler 1997, 58ff). This corresponds to the observation in Section 2 that in Modern German the perfect is used in spoken/informal contexts and the preterit is more typical to written/formal contexts. Since the MHG and ENHG texts are sermons, they contain sections in which the speaker directly addresses or includes the audience and other sections where he narrates or comments on biblical stories. In order to test the hypothesis we export the syntactically annotated data using the SaltNPepper framework to the span based EXMARaLDA format, so that each clause in each text is marked as a flat span for annotation. SaltNPepper necessarily loses information in the process, since EXMARaLDA XML cannot express the hierarchical ordering of constituents inherent in the syntactic annotation; however, that hierarchy does not interest us for the present purpose (but needless to say, the original syntactic clause annotation remains untouched). While not suitable for syntactic annotation, the EXMARaLDA tool is very convenient for discourse annotations, i.e. annotating spans of text as having certain properties. We therefore annotate the exported text with an additional layer ‘context’, merge it back with

[4] Similar rates of change (logistic curves) have actually been observed for many linguistic phenomena, see e.g. Kroch (1989).

[5] There are historical treebanks for some historical languages, e.g. English (Kroch et al. 2004), Latin (Bamman et al. 2009), or Greek (Bamman & Crane 2006). Apart from the very small DDB there is no freely available Treebank of historical German (Demske et al. 2004 describe the Mercurius treebank for ENHG - the treebank is as yet not published).

the original data and reimport it into ANNIS. The contexts are divided into two main classes: COM for communicative contexts and NAR for narrative contexts, which are assigned to the token spans of clauses. The distinction is fairly easy to operationalize: COM contexts contain vocatives, first or second person plural verbs or first or second person personal or possessive pronouns whereas NAR contexts convey third person information:

(4a)

tok	daz	er	durh	uns	komen	ist	ze	der	marter
edition	daz	er	dvrh	unſ	komen	iſt	ze	der	marter
lemma	daz	er	durch	er	komen	sîn	ze	der	marter
cont	NAR								
bib						Mellbourn, 19/46,23			

that he through us come is to the torture

(MHG subcorpus)

(4b)

tok	Lieben	,	disen	tak	den	hat	got	selbe	geheret	unde	gewihet
edition	Lieben	,	diſen	tak	den	hat	got	ſelbe	geheret	vnde	gewihet
lemma	liebe	,	dirre	tac	der	haben	got	selp	hêren	unde	wîhen
cont	NAR										
bib								Mellbourn, 6/21,21			

Dear this day this has god himself honoured and blessed

(MHG subcorpus)

(4c)

tok	Daz	sungen	die	engele	,
edition	Daz	ſvngen	die	engele	,
lemma	der	singen	der	engel	,
cont	COM				
bib			Mellbourn, 6/21,9		

this sang the angels

(MHG subcorpus)

(4d)

tok	Der	brunne	,	da	der	bach	uzran	,	der
edition	Der	brvnne	,	da	der	bach	vz|ran	,	der
lemma	der	brunne	,	da	der	bach	ûzrennen	,	der
cont	COM								
bib	Mellbourn, 6/21,17								

the well where the river ran out it

(MHG subcorpus)

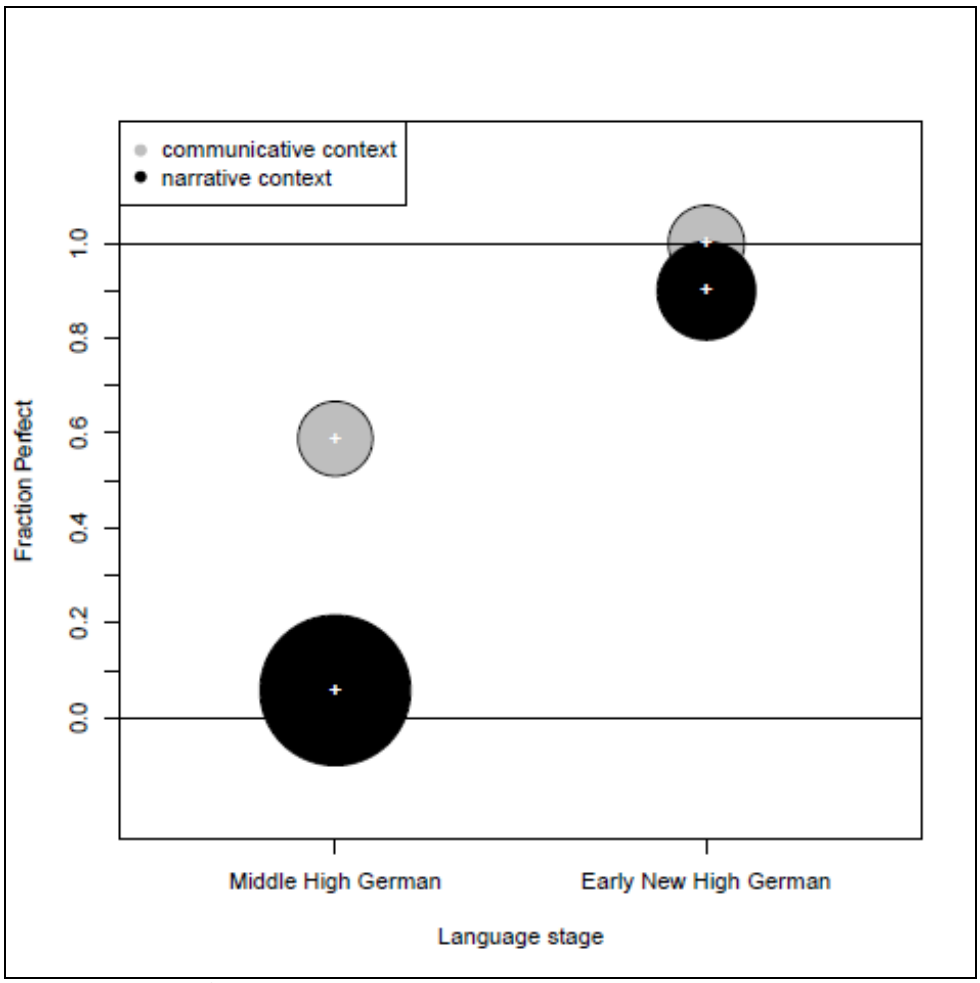

Fig. 4: Preterit and perfect constructions in communicative (grey) and narrative (black) contexts in MHG and ENHG. Frequencies of perfect and preterit tenses are expressed in fractions (1.0 means 100% perfect constructions, 0.0 means 100% preterit constructions per language stage). The sizes of the blobs imply the relative quantities of the respective contexts.
Normalized frequencies (occurrences per 1000 clauses): 24.7 preterit tenses and 35.3 perfect tenses in communicative contexts, 229.7 preterit tenses and 14.1 perfect tenses in narrative contexts in MHG; 0 preterit tenses and 62.5 perfect tenses in communicative contexts, 10.4 preterit tenses and 93.8 perfect tenses in narrative contexts in ENHG.

Figure 4 shows that our hypothesis holds. In MHG we see that the preterit occurs significantly more often in the narrative contexts than in the communicative contexts while the perfect (unexpectedly) occurs in both.

In ENHG there is a massive difference in function and Figure 4 shows exactly what we expect: the perfect is the default tense for both contexts. If preterit occurs, it appears in NAR contexts (there is not a single occurrence of preterit in a COM context). ENHG thus behaves similarly to what we expect of Modern German, whereas MHG seems to be quite different from these periods. Our contextual classes do not seem to be able to explain the distribution of perfect constructions in the older period.

We therefore need to test a second hypothesis: The literature claims that in earlier stages of German aspectual distinctions are a relevant trigger for preterit and perfect constructions. Perfect constructions emerged from present tense constructions which were reinterpreted and grammaticalized as a regular past tense construction (compare Grønvik 1986, Moya 2010). Presumably, the perfect tense has per se had a resultative reading[6] and later on developed general (including imperfective) readings. We want to test whether aspectual restrictions existed before the perfect tense became the default past tense construction. Following our assumptions we would expect that in MHG (where the preterit is still the unmarked past tense), the perfect is exclusively used for resultative readings with relevance to the present context (as a direct result of the presence of a present auxiliary, much like the English present perfect), whereas in ENHG, where the perfect tense has become the default past tense, it can occur in both resultative and non-resultative contexts.

To test this hypothesis we again need to add an annotation layer (aspect) in which we assign resultative and non-resultative contexts to the MHG and ENHG data and measure the frequency of perfect and preterit constructions in each context. This is then again merged into the corpus, allowing for a simultaneous search of each past tense construction in conjunction with the different contexts. The results of this analysis are visualized in Figure 5.

[6] That is to say it emphasizes the 'successful completion of a situation' (Comrie 1976, 20), implying relevance of the state resulting from the verbal action for the current discourse.

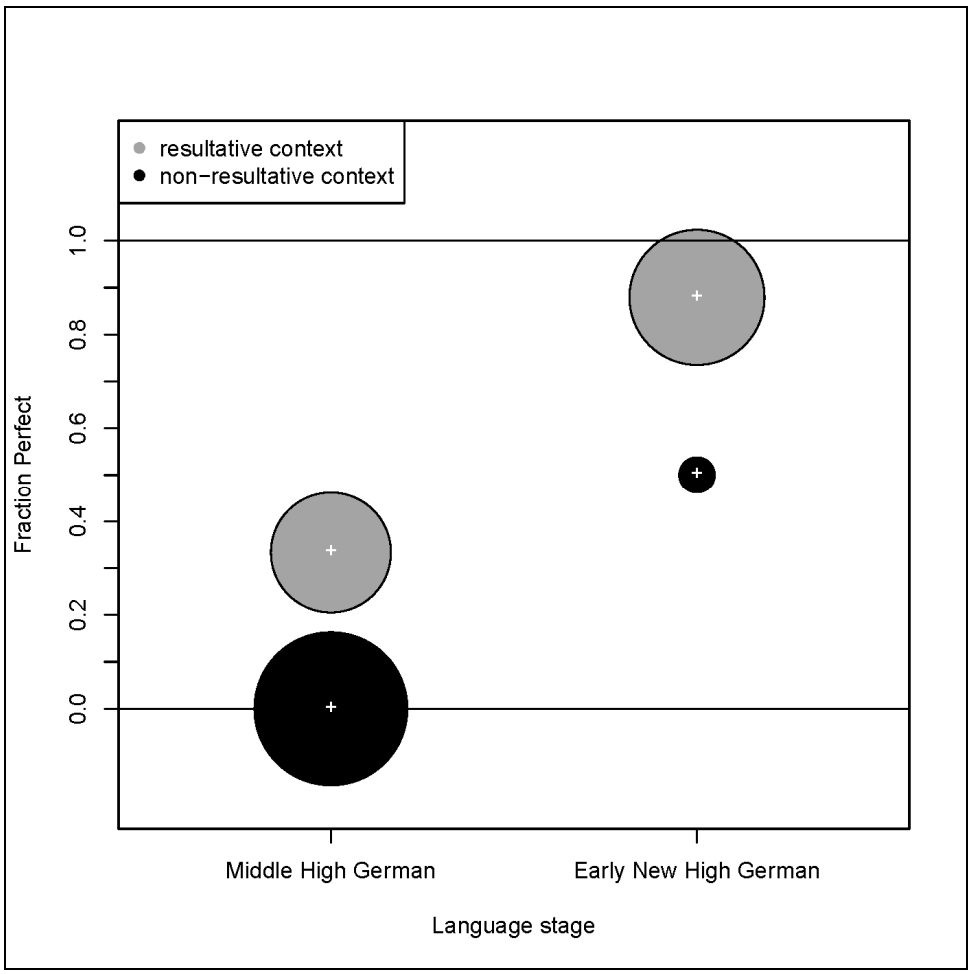

Fig. 5: Preterit and perfect constructions in resultative (grey) and non-resultative (black) contexts in MHG and ENHG. Frequencies of perfect and preterit tenses are expressed in fractions (1.0 means 100% perfect constructions, 0.0 means 100% preterit constructions per language stage). The sizes of the blobs imply the relative quantities of the respective aspectual contexts.
Normalized frequencies (occurrences per 1000 clauses): 106 preterit tenses and 53 perfect tenses in resultative contexts, 258 preterit tenses and 0 perfect tenses in non-resultative contexts in MHG; 24.3 preterit tenses and 177.1 perfect tenses in resultative contexts, 6.9 preterit and 6.9 perfect tenses in non-resultative contexts in FNHG.

In MHG there are no instances of perfect constructions in non-resultative contexts (black blob is at zero level), although there is a high ratio of non-resultative contexts in the MHG data (size of the black blob). Non-resultative readings are never expressed by perfect tense constructions, which can be regarded as evidence for our hypothesis that in MHG the perfect tense (still) has a clear resultative reading.

Figure 5 shows clearly that in ENHG perfect constructions increase drastically for both resultative and non-resultative readings, which means that the perfect tense becomes the unmarked past tense construction, taking over both aspectual contexts. The two different contextual categories, communicative vs. narrative contexts, and resultative vs. non-resultative readings, thus complement each other in explaining the division of perfect

and preterit tenses in MHG and ENHG. None of the two factors can explain the distribution of the two past tense constructions by themselves, but both contribute to a complex explanation taken together.

6. CONCLUSION

Historical linguists have been using electronic corpora for several years, but most of the actual linguistic analysis is still not coded back into the corpora to ensure that it is transparent and reproducible. In this study we have shown how deeply annotated corpora can be used in historical linguistics to find the contextual factors responsible for variation in language change. We have shown that it is often necessary to make further analyses in already annotated corpus data to answer certain research questions, and that the nature of these annotations can become clear in the course of the investigation itself. For additional categories to become truly useful, it is necessary to integrate them into the corpus. These explicit analyses can be used not only to reproduce previous results, but also for further studies, which can extend the corpus further with other researchers' own annotations.

Using the development and distribution of German preterit and perfect tenses as a test case we have shown how this methodology can be realized: we extract information from an initial annotation scheme, export relevant data using SaltNPepper into an appropriate format for further annotation in a dedicated tool (in this case EXMARaLDA), then merge the extended data back into a corpus in stand-off XML and finally re-import all annotations into ANNIS for research and publication.

We hope that future research in corpus-based historical linguistics will increasingly make data freely available, and aim to contribute to the dissemination of relevant tools and methodologies.

REFERENCES

ALBERT, S. ET AL. 2003. Tiger Annotationsschema. Technical Report. Universität Potsdam, Universität des Saarlandes, Universität Stuttgart.
(http://www.ifi.uzh.ch/cl/volk/treebank_course/tiger_annot . pdf).

BIBER, D. AND JONES, J. 2009. Quantitative Methods in Corpus Linguistics. In: LÜDELING, A. AND KYTÖ, M. (eds.) Corpus Linguistics. An International Handbook. Vol 2. Berlin: Mouton de Gruyter. pp. 1286-1304

BAMMAN, D. AND CRANE, G. 2006. The Design and Use of a Latin Dependency Treebank. Proceedings of the Fifth International Workshop on Treebanks and Linguistic Theories (TLT 2006). Prague. pp. 67-78.

BAMMAN, D., MAMBRINI, F. AND CRANE, G. 2009. An Ownership Model of Annotation: The Ancient Greek Dependency Treebank. Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT8). Milan.

BRANTS, S., DIPPER, S., HANSEN, S., LEZIUS, W. AND SMITH, G. 2002. The TIGER Treebank. In: Proceedings of TLT-02. Sozopol, Bulgaria.

CARLETTA, J., EVERT, S., HEID, U., KILGOUR, J., ROBERTSON, J. AND VOORMANN, H. 2003. The NITE XML Toolkit: Flexible Annotation for Multi-modal Language Data. Behavior Research Methods, Instruments, and Computers 35(3), 353-363.

COMRIE, B. 1976. Aspect. An Introduction to the Study of Verbal Aspect and Related Problems. Cambridge: Cambridge University Press.

CRANE, G. 1998. The Perseus Project and Beyond: How Building a Digital Library Challenges the Humanities and Technology. D-Lib Magazine, January 1998.

CRYSMANN, B., HANSEN-SCHIRRA, S., SMITH, G. AND ZIEGLER-EISELE, D. 2005. TIGER Morphologie-Annotationsschema. Technical Report, Universität Potsdam, Universität Saarbrücken.

DEMSKE, U., FRANK, N., LAUFER, S. AND STIEMER, H. 2004. Syntactic interpretation of an Early New High German corpus. Proceedings of the Third Workshop on Treebanks and Linguistic Theories (TLT 2004). pp 175–182.

DENTLER, S. 1997. Zur Perfekterneuerung im Mittelhochdeutschen. Die Erweiterung des zeitreferentiellen Funktionsbereichs von Perfektfügungen. Göteborg: Acta Universitatis Gothoburgensis.

DIEL, M., FISSENI, B., LENDERS, W. AND SCHMITZ, H.-C. 2002. XML-Kodierung des Bonner Frühneuhochdeutschkorpus. Technical Report. Bonn University.

DIPPER, S. 2005. XML-Based Stand-Off Representation and Exploitation of Multi-Level Linguistic Annotation. In: Proceedings of Berliner XML Tage 2005 (BXML 2005). Berlin, Germany, 39-50.

GRØNVIK, O. 1986. Über den Ursprung und die Entwicklung der aktiven Perfekt- und Plusquamperfektkonstruktionen des Hochdeutschen und ihre Eigenart innerhalb des germanischen Sprachraumes. Oslo: Solum Forlag.

HELBIG, G. AND BUSCHA, J. 2001. Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht. Berlin et al.: Langenscheidt.

HILPERT, M. 2008. Germanic Future Constructions. A usage-based approach to language change. Amsterdam/Philadelphia: John Benjamins

HIRSCHMANN, H. AND LINDE, S. 2011. Annotationsguidelines zur Deutschen Diachronen Baumbank. Technical Report. Humboldt-Universität zu Berlin.

KROCH, A. 1989. Reflexes of Grammar in Patterns of Language Change. In: Language Variation and Change 1, 199-244.

KROCH, A., SANTORINI, B. AND DELFS, L. (eds.) 2004. The Penn-Helsinki Parsed Corpus of Early Modern English. University of Pennsylvania, Philadelphia: Department of Linguistics.

KYTÖ, M. 1991. Manual to the Diachronic Part of the Helsinki Corpus of English Texts: Coding Conventions and Lists of Source Texts. Department of English, University of Helsinki.

LEISS, E. (1992), Die Verbalkategorien des Deutschen. Ein Beitrag zur Theorie der sprachlichen Kategorisierung. Berlin: Mouton de Gruyter (Studia linguistica Germanica, 31).

LEZIUS, W., BIESINGER, H. AND GERSTENBERGER, C. 2002. TIGER-XML Quick Reference Guide (Tech. Rep.). IMS, University of Stuttgart.

LÜDELING, A., HIRSCHMANN, H., ZELDES, A. to appear. Variationism and Underuse Statistics in the Analysis of the Development of Relative Clauses in German. In: KAWAGUCHI, Y., MINEGISHI, M. AND VIERECK, W. (eds.) Corpus Analysis and Diachronic Linguistics. Amsterdam: John Benjamins.

MCENERY, T. AND WILSON, A. 2001. Corpus Linguistics. 2nd ed. Edinburgh: Edinburgh University Press.

MOYA, I. G. 2010. Eine variationistische Analyse der Entstehung und Entwicklung des deutschen *haben*-Perfekts. Bachelor Thesis. Humboldt-Universität zu Berlin.

MUSAN, R. 2002. The German Perfect. Dordrecht: Kluwer Academic Publishers.

NÜBLING, D. 2006. Historische Sprachwissenschaft des Deutschen. Eine Einführung in die Prinzipien des Sprachwandels. In cooperation with DAMMEL, A., DUKE, J., AND SZCZEPANIAK, R. Narr, Tübingen

PETROVA, S., SOLF, M., RITZ, J., CHIARCOS, C. & ZELDES, A. 2009. Building and Using a Richly Annotated Interlinear Diachronic Corpus: The Case of Old High German Tatian. Traitement automatique des langues 50(2), 47-71.

REICHMANN, O. AND WEGERA, K.-P. (eds.) 1993. Frühneuhochdeutsche Grammatik. Tübingen: Niemeyer.

RESNIK, P., OLSEN, M. B. AND DIAB, M. 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. In: LÜDELING, A., AND KYTÖ, M. (eds.) Corpus Linguistics. An International Handbook. Vol 1. Berlin: Mouton de Gruyter, 53-68.

SCHMID, H. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of the Conference on New Methods in Language Processing. Manchester, UK

WITTEN, I. H. AND FRANK, E. 2005. Data mining: Practical Machine Learning Tools and Techniques, 2nd Ed. San Francisco: Morgan Kaufman.

ZELDES, A., RITZ, J., LÜDELING, A. AND CHIARCOS, C. 2009. ANNIS: A search tool for multi-layer annotated corpora. In: Proceedings of Corpus Linguistics 2009. July 20-23, Liverpool, UK.

ZIPSER F. 2009. Entwicklung eines Konverterframeworks für linguistisch annotierte Daten auf Basis eines gemeinsamen (Meta-)modells. Master Thesis, Humboldt-Universität zu Berlin, (https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/mitarbeiter-innen/florian/pdf/diplomarbeit.pdf)

ZIPSER F. AND ROMARY L. 2010. A model oriented approach to the mapping of annotation formats using standards. In: Proceedings of the Workshop on Language Resource and Language Technology Standards, LREC 2010. Malta.

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., FISSENI, B., LENDERS, W. AND SCHMITZ, H.-C. (2002), XML-Kodierung des Bonner Frühneuhochdeutschkorpus. IKP-Arbeitsbericht NF 02, Bonn.

ANNIS SEARCH QUERIES:

Main verbs or modal verbs in preterit indicative tense:
*pos=/V(V|M)FIN/ & morph=/.*Past.Ind/ & #1 _=_ #2*
Link to query

Perfect constructions with auxiliary *haben – to have*:
*cat="S" & tok & lemma=/(haben|eigan)/ & pos="VAFIN" & morph=/.*Pres.Ind/ & pos="VVPP" & #1 > #2 & #2 _=_ #3 & #2 _=_ #4 & #2 _=_ #5 & #1 >* #6*
Link to query

Perfect constructions with auxiliary *sein – to be*:
*cat="S" & cat="VP" & tok & lemma=/(wesan|sîn|sein)/ & pos="VAFIN" & morph=/.*Pres.Ind/ & pos="VVPP" & #1 > #3 & #3 _=_ #4 & #3 _=_ #5 & #3 _=_ #6 & #1 >[func="OC"] #2 & #2 > #7*
Link to query

Preterit/perfect constructions in communicative/narrative contexts:
*tense="PRET" & context="COM" & morph=/.*Ind/ & tok & #1 _=_ #2 & #3 _=_ #4 & #1 _i_ #4*
Replace "COM" by "NAR" to formulate the query for narrative contexts; replace "PRET" by "PERF" to search for perfect constructions instead.
While annotating the communicative and narrative contexts, we also annotated the tense of the respective clause as a span to be able to formulate the query more easily.
Link to query

Preterit/perfect constructions in resultative/non-resultative contexts:
*tense="PRET" & aspect="RES" & morph=/.*Ind/ & tok & #1 _=_ #2 & #3 _=_ #4 & #1 _i_ #4*
Replace "RES" by "NONRES" to formulate the query for non-resultative contexts; replace "PRET" by "PERF" to search for perfect constructions instead.
Link to query