Building and Using a Richly Annotated Interlinear Diachronic Corpus: The Case of Old High German Tatian

Svetlana Petrova[*] — Michael Solf[*] — Julia Ritz[**] — Christian Chiarcos[**] — Amir Zeldes[*]

Collaborative Research Centre 632 "Information Structure"
[*] *Humboldt Universität zu Berlin, Unter den Linden 6, D-10099 Berlin*
{s.petrova@staff.|michael.solf@|amir.zeldes@rz.}hu-berlin.de

[**] *Potsdam University, Karl-Liebknecht-Str. 24, D-14471 Potsdam*
{julia|chiarcos}@ling.uni-potsdam.de

ABSTRACT. *The present paper reports on the development and evaluation of a historical corpus designed to support detailed empirical studies on the interaction of information structure and syntax in Old High German (OHG). The creation and exploration of this corpus are part of a more general investigation concerning the role of information-structural factors in the explanation of word order variation and change in the Germanic languages. The paper also describes corpus design principles, methodologies, relevant formats and specifications, and the technical infrastructure employed during the creation of the corpus, as well as its accessibility by means of the linguistic database of information structure ANNIS.*

RÉSUMÉ. *Cet article rapporte le développement et l'évaluation d'un corpus historique conçu pour des recherches empiriques sur l'interaction entre la structure d'information et la syntaxe dans l'ancien haut-allemand. La création et l'exploration du corpus contribuent à l'investigation du rôle des conditions pragmatiques pour la typologie syntaxique, sa variation et sa mutation dans les langues germaniques. L'article décrit aussi les principes de design, les méthodologies, les formats et spécifications, et l'infrastructure technique utilisée pour créer le corpus. L'accès au corpus est obtenu par ANNIS, une base de données linguistique.*

KEYWORDS: *corpus, Old High German, information structure, syntax, multi-layer annotation.*

MOTS-CLÉS : *corpus d'ancien haut-allemand, structure d'information, syntaxe, annotation multi-niveaux.*

1. Background

The Collaborative Research Centre (German *Sonderforschungsbereich* SFB) "Information structure: the linguistic means for structuring utterances, sentences and texts" brings together scientists from different fields of linguistics and neighbouring disciplines from the University of Potsdam and Humboldt University Berlin. The overall goal of the SFB is to study on a large scale the different ways in which information-structural categories are encoded in a variety of extinct and modern languages. We define information structure (IS) as the structuring of linguistic information in order to optimise information transfer within discourse: information needs to be prepared ('packaged') in different ways depending on the goals a speaker pursues within discourse. Fundamental concepts of IS include 'topic', 'focus', 'background' and 'information status'. Broadly speaking, the topic is the entity a specific sentence is construed about, focus represents the new or newsworthy information a sentence conveys, background is that part of the sentence that is familiar to the hearer, and information status refers to different degrees of familiarity of an entity (see e.g. Krifka, 2007). The new insights gained in these fields of research are relevant for the formation of a theoretical model which accounts for the proper representation of IS in the human linguistic faculty.

Another goal of the SFB is the use and advancement of corpus technologies for complex linguistic annotations, such as the annotation of IS. Associated with this task is the project "Linguistic database for information structure: Annotation and Retrieval", henceforth *database project*. This project coordinates annotation activities in the SFB, provides services to other projects in the creation and maintenance of data collections and conducts theoretical research on multi-layer annotations (Chiarcos *et al.*, 2009c; Zeldes *et al.*, 2009). Its primary goals, however, are the development and investigation of techniques to process and exploit deeply annotated corpora with multiple kinds of annotations, such that heterogeneous resources can be accessed, queried and visualised in a unified way. The results of this research are implemented in the linguistic database ANNIS[1] described further below. For the specific facilities of ANNIS and its application to the corpus dealt with in this paper, see Sections 4 and 5.

Besides the implementation of ANNIS as a general-purpose tool for the publication, visualisation and querying of linguistic data collections, the database project conducts research on the development of multi-layer corpus architectures required for the study of IS and the technical means for their integration, merging and unified processing. This integration of complex annotations is achieved using the generic XML format PAULA (Dipper, 2005; Dipper and Götze, 2005), the native corpus format for ANNIS. PAULA is capable of representing the full spectrum of text-based linguistic annotation, in particular supporting multiple and conflicting hierarchical annotations. In order to achieve this, PAULA uses a stand-off XML architecture (see Carletta *et al.*,

1. ANNotation of Information Structure. Software freely available at `http://www.sfb632.uni-potsdam.de/~d1/annis/`.

2003). This means that annotations and primary data are stored in separate files so that multiple layers of annotation are physically independent from each other and can be organised into independent hierarchies without disrupting existing structures.

The methods developed by the database project have been applied to the processing of historical data, which presents a special challenge to the development and evaluation of a linguistic database. In cooperation with the project "The role of information structure in the development of word order regularities in Germanic", henceforth *diachronic project*, the Tatian Corpus Of Deviating EXamples (T-CODEX) has been developed as a pilot corpus for Old High German (OHG). The main goal of investigation in the diachronic project is the impact of IS on language change. A basic assumption that motivates this research is that – for the needs of communicative explicitness and rhetorical expressivity – novel word order patterns emerge which, in the course of time, may lose their special pragmatic value and become the unmarked word order pattern of a language (Hinterhölzl, 2004; Hinterhölzl, 2009).

Building upon this hypothesis, the central empirical task of the diachronic project is to find out whether there is a correlation between the information-structural properties of sentence constituents and their syntactic realisation in the clause. As IS is a complex phenomenon reflecting categories and features on various interrelated levels of pragmatic representation (Molnár, 1993; Krifka, 2007), any account of different information-structural factors leading to syntactic variation and change enforces the implementation of a multi-layer corpus architecture. A multi-layer architecture provides the possibility to search through different levels of annotation, so that we can approach the question of the extent to which the expression of information-structural categories may induce surface variation and subsequent changes in the underlying structure of the clause.

This article is structured as follows: the next section presents the data and the annotation scheme used for the creation of the corpus. The following section deals with the technical process of digitisation, annotation and documentation, touching on issues of relevant formats and annotation tools as well as the representation of metadata. Section 4 briefly presents the web interface ANNIS, which is used to access and search the corpus, and its query language AQL. Section 5 then puts the corpus and its technical architecture to use in a case study on OHG word order in subordinate clauses as a function of IS. Finally, Section 6 discusses ongoing research activities and future directions for work within the diachronic project.

2. The Pilot Corpus T-CODEX 1.0

2.1. *Objectives and Design Principles*

It is well known that Old High German is a particularly difficult ground for investigations on word order (Fleischer, 2006). This is due to the fact that the major part of OHG records consists of translations from Latin or of poetic texts. In both cases we have to be aware that the word orders attested in the records are influenced either by

the Latin original or by metrical considerations rather than reflecting genuine OHG patterns.

In this respect, the translation of Tatian's Gospel harmony from Latin into OHG offers a unique opportunity for accessing native word orders. It is one of the largest prose texts from the beginning of the OHG period, which is handed down to us in one manuscript (St. Gallen Cod. 56) written in the scriptorium of Fulda in the middle of the ninth century by at least six scribes. This text will be referred to as "the OHG Tatian" throughout this article. In the manuscript, the Latin source and the OHG translation are attested as two juxtaposed columns (see Figure 1). Only recently, it has been noticed that each line in the OHG text translates exactly the same material found in the corresponding Latin line (Masser, 1997a; Masser, 1997b). A new edition made available by (Masser, 1994) reflects these major characteristics of the manuscript and makes it possible to compare the source and target text (see Figure 1 bottom). The translating technique applied in the Tatian text imposes restrictions on the possibility of rendering genuine word order patterns in the translation, while the deviations from the Latin source can be viewed as evidence for genuine OHG structures (Dittmer and Dittmer, 1998). Recent investigations even attest a singular high value of this text for any investigation on Old High German syntax (Donhauser, 1998; Fleischer *et al.*, 2008).

Based on these considerations, the investigation of word order and IS in the diachronic project is restricted to single instances in which the word order of the OHG text differs from the underlying Latin structure. The pilot version of the corpus contains all clauses showing such deviations occurring in the sections assigned to three different scribes. We examined the sections assigned to the first two scribes α and β exhaustively, as well as the section assigned to scribe ϵ, where violations of the line principle of the translation occur particularly often. These data comprise the raw material for the Tatian Corpus Of Deviating EXamples (T-CODEX). In total, the pilot version 1.0 of T-CODEX comprises 1,658 clauses and 9,351 tokens, approximately one-third of the entire text.

Since the extracted clauses do not form continuous text, they are stored as single documents. Each document is annotated for the various grammatical and information-structural features described in detail in Section 2.2, using the annotation tool EXMARaLDA (Schmidt, 2004), which supports multi-level annotation. We employed the annotation guidelines developed in the SFB (Dipper *et al.*, 2007), with some additional distinctions forced by the properties of the data. These modifications are documented in detail in Petrova (2009). The annotation, including the digitisation of the source data, was performed manually, as methods for an automatic or semi-automatic annotation available at the time did not meet our quality standards.

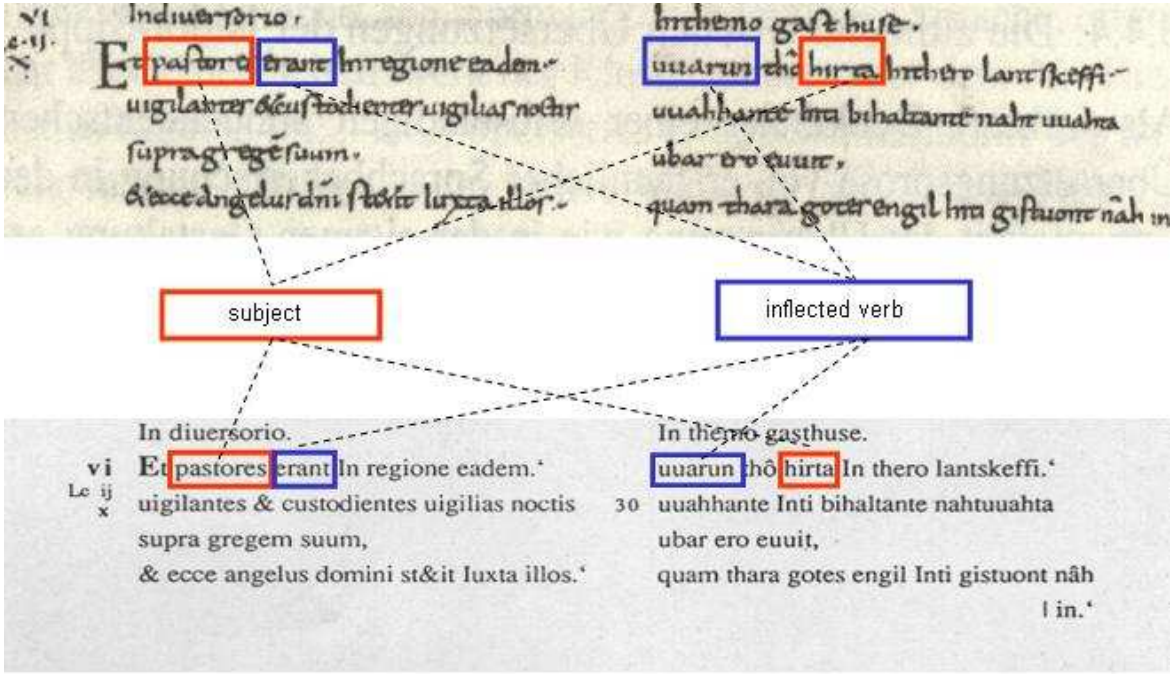


Figure 1. *Facsimile of St. Gallen Cod. 56, p. 35 (Sonderegger 2003, 130); bottom: the same text in the edition by Masser (1994, 85), "There were shepherds in the same country" (Luke 2,8)*

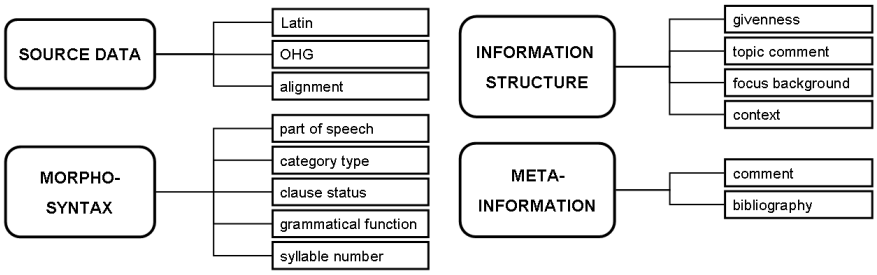


Figure 2. *Schematic representation of annotation levels in T-CODEX*

2.2. *Annotation Scheme*

The annotation of the single clauses extracted from the OHG Tatian translation is strictly tailored to the specific research goals of the diachronic project. This means that the main effort is put into the assignment of a broad range of information-structural features that will be examined as potential factors for variation in word order. As the corpus consists of clauses departing from the structure of the Latin, less emphasis is placed on the deep annotation of alignment properties.

In particular, the annotation in the pilot version 1.0 of T-CODEX provides information pertaining to the following four layers: I) source data, II) morpho-syntax, III) IS and IV) meta-information.

Layer I conveys the source data, i.e. the Latin original and its OHG translation. As a rule, we adhere to the orthographic conventions used in the text edition by Masser (1994).

Layer II contains information on the morpho-syntactic properties of the clause constituents, with special emphasis on parts of speech, syntactic category and grammatical function. Additionally, we determine clause status and the number of syllables for each constituent. The latter information is necessary to draw conclusions on the role of phonological weight for the placement of phrases in the clause.

Layer III is dedicated to IS. In line with the current approaches in the literature (Krifka, 2007, cf. also Molnár, 1993), we subscribe to the notion that IS is a complex phenomenon reflecting at least the following three independent levels of representation:

1) cognitive status, i.e. ‘given’ vs. ‘new’,

2) predication structure, i.e. ‘topic’ vs. ‘comment’, and

3) informational relevance, i.e. ‘focus’ vs. ‘background’.

As the definition of major IS categories like topic and focus is controversial in the theory, and the standard linguistic tests are difficult to apply to historical data (Petrova and Solf, 2009), we decided to annotate a wide range of features considered relevant for these categories in research.

Currently the annotation of IS is organised as follows:

1) Givenness is annotated for discourse referents only, with respect to the presence or absence of an explicit antecedent in the preceding discourse (given vs. new); additionally, referents lacking an antecedent but being inferable in the particular linguistic situation are annotated as accessible.

2) With respect to the assignment of topic-comment, we collect features considered as constitutive for topics in different frameworks: presence or absence of a bipartite structure of the predication, potential candidates for aboutness topicality, definiteness, topic-marking strategies like special syntactic constructions (German *Vorfeld*, Left Dislocation etc.) or the use of morphological markers.

3) The annotation of focus reflects novelty, contrast and emphasis as well as the use of focus markers.

To relate the clause to the preceding context, especially in terms of discourse organisation, we also annotate properties of the context, where we mainly distinguish between ’new setting’ (*Situationswechsel*) and anaphoric reference.

Finally, in Layer IV, we provide meta-information concerning manuscript page and line, scribe and concordance notes. Here, a layer for individual comments of the annotator is also provided.

With this wide range of factors annotated independently from each other, we are able to query the corpus across multiple levels and thus investigate possible interdependencies between particular features or combinations of features. This cumulative approach produces a comprehensive linguistic resource, particularly for creating an information-structural cartography of the left and right sentence periphery in OHG.

3. Assembling, Annotating, Processing and Querying T-CODEX

This section discusses challenges in the preparation of T-CODEX: the non-standardised orthography of OHG (as in most ancient languages); multiple levels of highly abstract annotations, at times requiring world knowledge (e.g. for the level of givenness); diverse types of annotations, ranging from labels applying to individual tokens, to spans of tokens and even links between segments, e.g. alignment annotation.

3.1. *Digitisation*

The source data described in Section 2.1 was digitised primarily based on Masser (1994). Since 2006, we have been able to adjust Masser's readings using a digital facsimile made accessible by the *e-codices project* of the University of Fribourg[2], and we also used a microfilm of the manuscript. Finally, we occasionally employed the older Tatian edition by Sievers (1892).

The data was digitised according to the following principles: Punctuation marks were ignored. Line breaks are marked with a slash "/" whenever they appear within a sentence. Ligations appearing within the text edition are represented in the source line. Accents and other diacritics are represented as in the text edition, despite the fact that research has so far not succeeded in proving a particular prosodic function for these.

Digitisation was performed manually, i.e. the data was typed directly into EXMARaLDA, the tool used for annotation. Besides the greater precision as compared to the application of OCR techniques[3], this allowed us to include editorial information (alignment with Latin and bibliographic information) in the annotations of the OHG text in one pass. The OHG text constitutes the primary layer, and its tokenisation represents the minimal granularity available to annotations in a so-called timeline. All annotations that apply to one or more tokens then refer to this timeline.

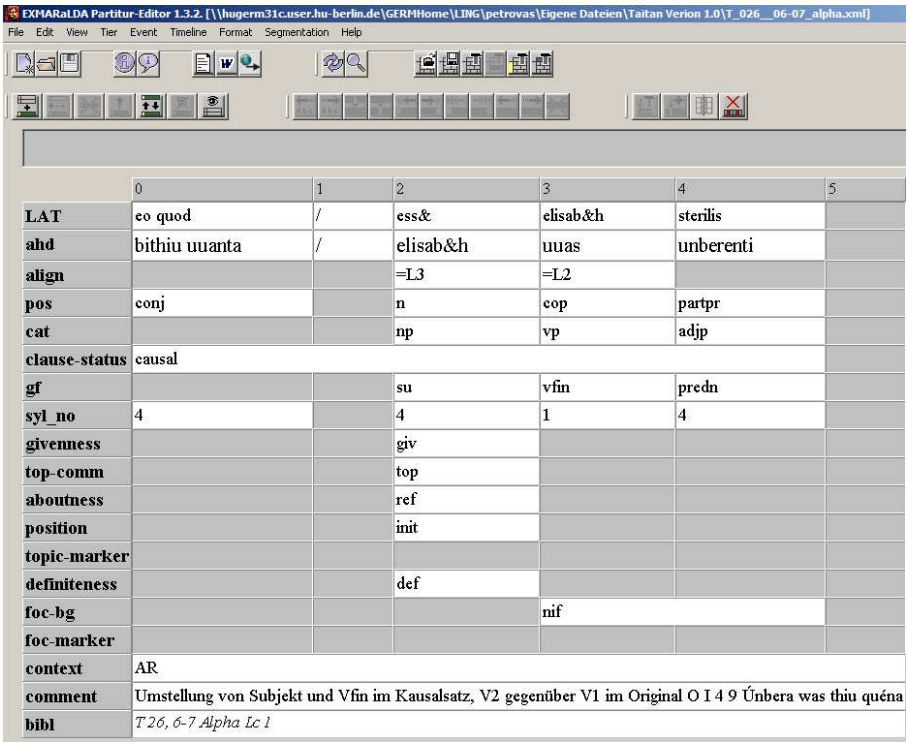
Corresponding tokens in OHG and Latin are represented in an interlinear form. If an OHG token does not have a Latin equivalent, the cell on the Latin tier is left empty. In case of permutation against the Latin original, a token alignment is provided in an extra tier (one-to-one wherever possible). Alignment is annotated as follows: if the OHG token in timeline position t does not correspond to the Latin token at the same position, an annotation is added with the value '=L<t>', where <t> is the position of the corresponding Latin token. This annotation type is interpreted as a link when imported into the interchange format PAULA (see Section 3.4).

3.2. *Annotation Process with EXMARaLDA*

Each document in T-CODEX was annotated for various grammatical and IS features (as described in Section 2.2) as a separate EXMARaLDA file. Figure 3 shows a complete document entry. EXMARaLDA documents are encoded in an XML format that allows for the annotation of feature/value pairs for sequences of tokens, where se-

2. `http://www.e-codices.unifr.ch/en/list/one/csg/0056`.

3. For a Tatian digitisation project using optical character recognition see http://lexicon.ff.cuni.cz/texts/ohg_sievers_tatian_about.html. Although OCR is a time-saving technique, it introduces a certain amount of noise and requires extensive manual corrections. For our purposes, where the text was not only to be typed but also filtered according to linguistic criteria and enriched with edition-relevant annotations, the benefits of OCR would not have been substantial.



	0	1	2	3	4	5
LAT	eo quod	/	ess&	elisab&h	sterilis	
ahd	bithiu uuanta	/	elisab&h	uuas	unberenti	
align			=L3	=L2		
pos	conj		n	cop	partpr	
cat			np	vp	adjp	
clause-status	causal					
gf			su	vfin	predn	
syl_no	4		4	1	4	
givenness			giv			
top-comm			top			
aboutness			ref			
position			init			
topic-marker						
definiteness			def			
foc-bg				nif		
foc-marker						
context	AR					
comment	Umstellung von Subjekt und Vfin im Kausalsatz, V2 gegenüber V1 im Original O I 4 9 Únbera was thiu quéna					
bibl	*T 26, 6-7 Alpha Lc 1*					

Figure 3. *EXMARaLDA transcription of document T 26, 6-7 ("because Elizabeth was barren", Luke 1,7)*

quences on different layers may overlap. The tier containing the OHG word forms is interpreted as the token layer, i.e. the layer constituting the minimally granular entities available to annotation.

3.3. *Annotation Documentation*

Usually, linguistic annotations are documented in human-readable form, e.g. as a technical report, in a reference publication or a handbook. While this is certainly the most intuitive type of documentation, there are good reasons to create documentations that are machine-readable as well. Firstly, information can be used to explore corpora based on conceptual information – more or less independently from a specific tagset – so that researchers unfamiliar with a specific scheme may get a first glance at a resource without having to study the documentation (Chiarcos *et al.*, 2009a). Sec-

ondly, the formal specification of annotation documentation according to a specific scheme allows one to generate annotation documentation in a highly consistent way (Chiarcos, 2008). And thirdly, it is particularly fruitful if multiple annotation schemes are defined with reference to a single terminological reference (as suggested by a number of initiatives such as EAGLES[4], GOLD[5], DCR[6] or TDS[7]). For this purpose, we employ OLiA (Ontologies of Linguistic Annotation), a modular architecture described in Chiarcos (2008). The OLiA *reference model* represents a terminological backbone, consisting of (i) a taxonomy of linguistic categories (implemented as OWL[8] classes such as *Noun*, *CommonNoun*), (ii) a taxonomy of grammatical features (OWL classes, e.g. *Accusative*), and (iii) relations (OWL properties, e.g. *hasCase*). Annotation models representing an annotation scheme are then linked to this reference model.

T-CODEX is accompanied by an annotation model for its annotation, allowing the use of an XSL transformation specifically developed for the visualisation of OLiA annotation models: the ontological specifications can be converted into HTML and be used for documentation purposes (see Figure 4).

3.4. *Conversion to PAULA and Corpus Organisation*

Once the primary data has been annotated, it is converted to the interchange format PAULA[9] (Dipper, 2005; Dipper and Götze, 2005) used for a wide variety of corpora at the SFB. PAULA is a generic XML format for the representation of different types of linguistic annotations, integrating annotations and metadata from various source formats. It also represents the native format of the linguistic database ANNIS, which integrates corpus data from diverse sources (Chiarcos *et al.*, 2009a).

The PAULA format distinguishes three types of nodes: *tokens*, *markables* and *structs*. Tokens are the basic segments of linguistic annotations as considered here. Markables are annotation elements that extend over a certain (and possibly discontinuous) span of tokens. Structs represent hierarchical annotations, i.e., one struct may contain not only tokens, but also markables and other structs. Nodes may be connected by edges ('*relations*'), in particular dominance (between one struct and any of its child elements), or a pointing relation (links between nodes of any type without implying a hierarchical relationship). Any of these nodes or edges can be provided with *features*.

Based on these data structures, a PAULA document is a collection of files with the primary data contained in one file and, for each annotation layer, files contain-

4. `http://www.ilc.cnr.it/EAGLES96/home.html`.
5. `http://emeld.org/tools/ontology.cfm`.
6. `http://www.isocat.org/`.
7. `http://languagelink.let.uu.nl/tds/index.html`.
8. OWL - Web Ontology Language, http://www.w3.org/TR/owl-ref/.
9. PAULA stands for German *Potsdamer AUstauschformat Linguistischer Annotationen*, 'Potsdam Interchange Format for Linguistic Annotations'.

Figure 4. *Representation of definiteness annotation in T-CODEX using OLiA*

ing the annotated elements (tokens, markables, structs and relations) and the actual annotations (feature/value pairs) for each level. These files are bound together in a separate master file, the *AnnoSet*. An entire PAULA corpus, such as T-CODEX, is defined as a set of *AnnoSets*. Due to the physical separation of primary data and the various annotations, a stand-off architecture arises which allows annotation of overlapping elements and conflicting hierarchies, but also the replacement or addition of new annotation levels at a later time without disrupting existing data.

3.5. *Metadata*

As proposed by Rehm *et al.* (2008), a corpus can contain metadata at the following different levels:

setting (author(s), time and place of origin of the raw data);

raw data (nature of raw data, e.g., manuscript, book, audio or video recording of a conversation, etc.);

primary data (data that is the object of annotations, e.g., transcribed speech, digital texts, etc.);

annotations (information added to primary data, e.g. parts of speech);

corpus as a whole (primary data with one or more annotation levels, e.g. underlying edition, author(s) of annotations).

For T-CODEX, metadata sets were created using the eTEI XML scheme (Rehm *et al.*, 2008), an extension to the document header suggested by the TEI[10]. These sets were then included in the *AnnoSet* of the corpus, as information relevant to the corpus as a whole, rather than repeated in every document of the corpus, which would result in a highly redundant representation.

4. Accessing T-CODEX with ANNIS

4.1. *General background*

ANNIS[11] is a Java-based corpus search web application that allows users to visualise, query and mine corpora annotated at multiple levels, i.e. with (i) annotations of different types (spans, pointing relations, trees/graphs with labelled edges), and (ii) annotation structures possibly overlapping and/or conflicting.

Although ANNIS uses PAULA as a native format, a variety of converters coupled with merging tools allow ANNIS to exploit data stemming from diverse manual and automatic tools (taggers, parsers, etc.), which can be specialised on their respective tasks. Importers into PAULA exist for various formats, including EXMARaLDA (Schmidt, 2004), Tiger-XML (i.e. annotate (Brants and Plaehn, 2000) and Synpathy[12]), MMAX2 (Müller and Strube, 2006), RSTTool (O'Donnell, 2000), PALinkA (Orasan, 2003) and Toolbox[13]. Internally, the data in ANNIS is then compiled into a relational database for reasons of performance (see below).

Query matches can be visualised directly in ANNIS or exported along with selected features in the ARFF format, the input format of the data mining tool WEKA (Witten and Frank, 2005), which offers implementations of various clustering and classification algorithms.

10. Text Encoding Initiative, `http://www.tei-c.org/index.xml`.
11. See Chiarcos *et al.* (2009a) for technical details.
12. `http://www.lat-mpi.eu/tools/synpathy/`.
13. `http://www.sil.org/computing/toolbox/`.

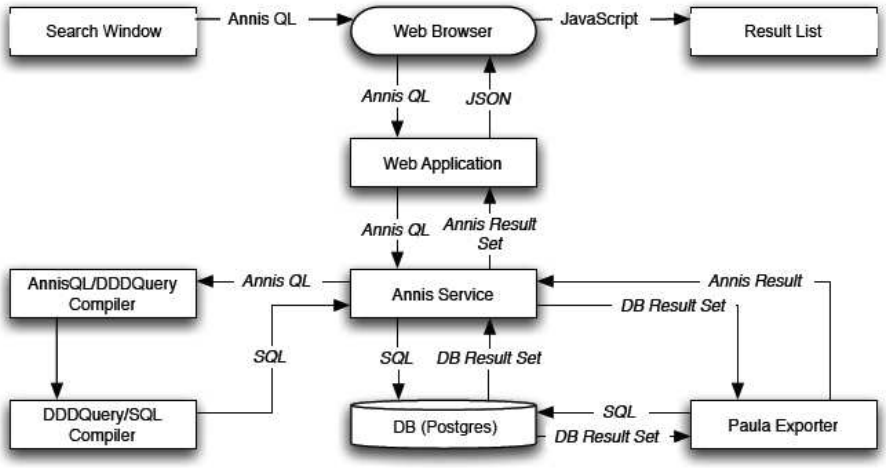


Figure 5. *Application logic of Annis2*

4.2. *System Architecture*

The ANNIS architecture is based on a multi-tier architecture roughly dividable into a back-end employing a relational database, a window-based web interface running on AJAX and a set of middleware services mediating between the two (Figure 5). For reasons of performance and scalability, the original XML documents are not searched using an XML database, but are rather compiled into a relational format, exploiting the advantages of off-the-shelf relational databases on an open source platform (specifically, PostgreSQL is used). Among other features, the database natively supports Unicode and regular expression searches. In order to reconcile the rather complex SQL syntax necessary for querying the complex structures in the relational schema with the user's need for a comprehensible query language, queries are formulated in the far simpler ANNIS Query Language (AQL)[14] and compiled in terms of nodes and edges into SQL queries. The AQL syntax is similar to and partly based on NXT Search and the Nite Query Language (Heid *et al.*, 2004; Carletta *et al.*, 2005) as well as TIGERSearch (Lezius, 2002). AQL makes reference to nodes[15] being searched

14. See `http://www.sfb632.uni-potsdam.de/~d1/annis/` for a query language documentation.

15. Search constraints on nodes are formulated as follows: `tok` for any token, i.e. terminal node, `node` for any non-terminal node and annotations in the form of feature value pairs, e.g. `tok="bithiu"` or `pos="PRONPRS"`. For constraints to the token level, a quoted string (`"bithiu"`) may be used as a shortcut.

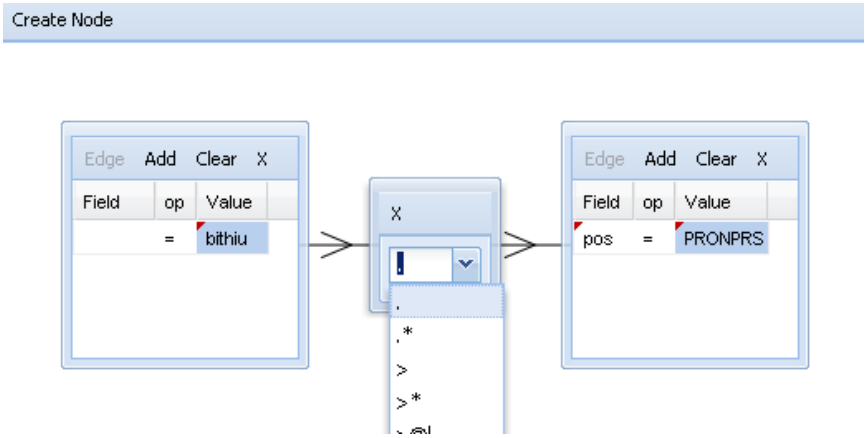


Figure 6. *Query Builder representing Query (1)*

for and the relations which must hold between them in order to produce a match. For example, query (1) searches for two nodes: the first is the OHG word form *bithiu* 'because' and the second is an annotation of the type `pos` (part of speech) with the value `PRONPRS` (a personal pronoun), where both elements are adjacent, as indicated by the operator '.' between the numeric references corresponding to the two elements.

(1) `"bithiu" & pos="PRONPRS" & #1 . #2`

In order to facilitate the formulation of such queries, ANNIS also provides a graphical query builder (Figure 6), where nodes are represented by boxes and relations between elements by edges which carry the operator labels. A wide variety of operators is available for querying these relations, including different types of direct and indirect adjacency, overlap and hierarchy between annotations. Section 5 discusses more complex queries used to study the influence of Information Structure as annotated in the corpus on the development of German word order. Once the desired query is formulated, the system retrieves all instances that meet the constraints, along with the total number of search results, which can be used in quantitative investigations. The matching instances in context and the corresponding annotations may be visualised in a number of ways, including 'key word in context'-style token annotations (for levels annotating individual tokens only), annotations of arbitrary spans of continuous or discontinuous tokens, hierarchical trees or directed acyclic graphs (used for syntactic annotation), text-wide views used to annotate discourse referents and coreferent expressions, and even multimodal data using an embedded player.

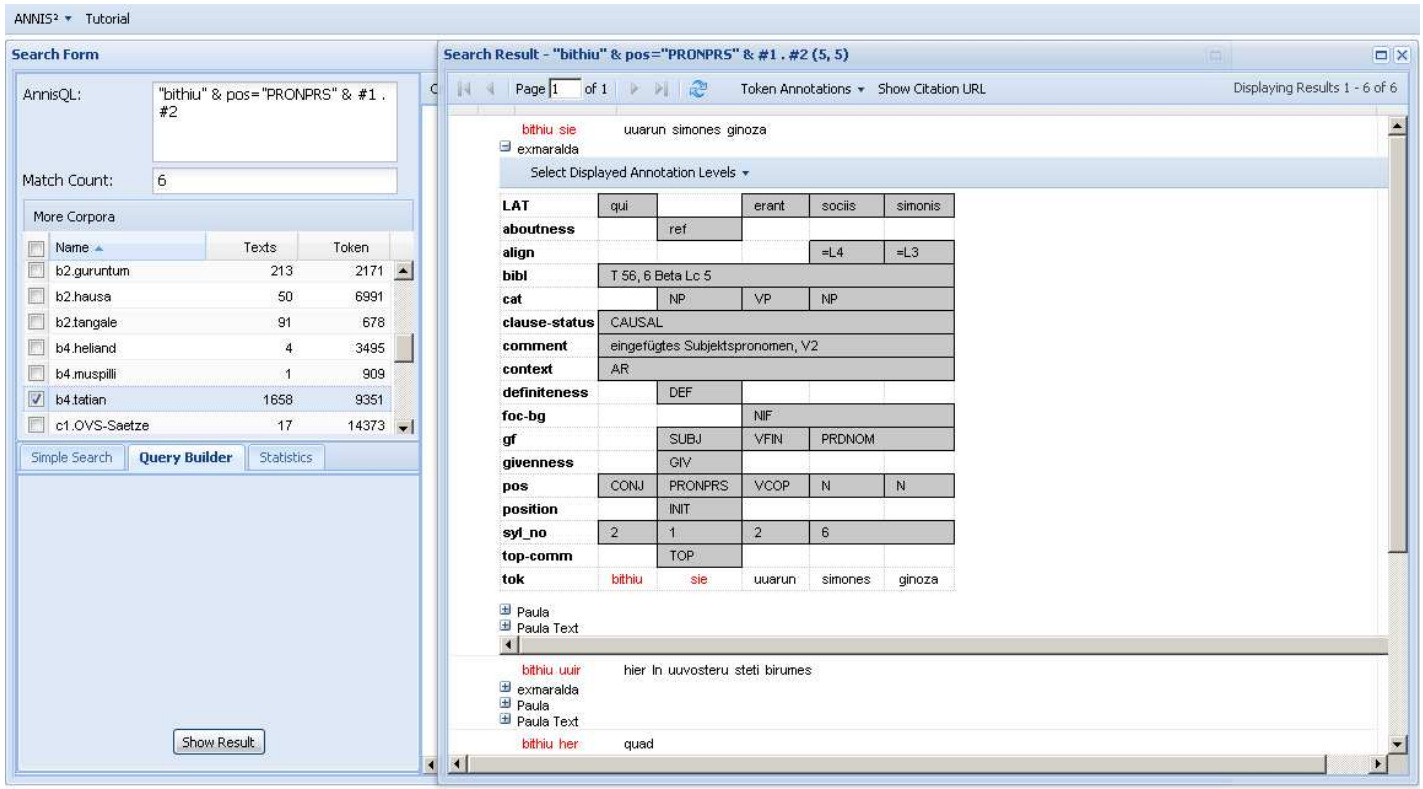


Figure 7. *Result list for Query (1) with grid view of the annotations ("because they were partners with Simon", Luke 5,10)*

Since T-CODEX primarily comprises non-hierarchical annotations (i.e. annotations to [spans of] tokens), a flexible grid view is employed to visualise all annotation levels at once (Figure 7; the user can specify which levels to show or hide). This view is familiar to users participating in the corpus annotation, since it closely conforms to the interface of the EXMARaLDA tool used for annotation, and allows users to recognise overlapping relationships between the annotations and the primary data. The system is also capable of generating URLs pointing at query results for use in publications, thus facilitating collaboration and reproducibility of results. In order to support versioning and collaboration of users with different access rights, the system requires user authentication, so that these links are only usable for licensed users.

5. Case Study

This section describes how the corpus is being exploited for different linguistic questions concerning the interaction of IS and word order variation in the diachronic project. Different aspects of this issue have been addressed, e.g. by Hinterhölzl and Petrova (to appear) in a study on variation in root clauses, and an investigation on variation in subordinate clauses in OHG (Petrova, to appear; Petrova and Hinterhölzl, forthcoming).

Here, we provide an example related to the study of word order variation in subordinate clauses. As one of the most striking properties of OHG syntax, variation in subordinate clauses has often been discussed in the historical treatment of German sentence structure, as part of the description of the rules and principles that determine the position of the finite verb in the earlier stages of the language. The same kind of variation is also attested in the earlier stages of the remaining Germanic languages, and has therefore received special attention in recent generative literature. Three basic accounts have been put forward to explain the issue.

According to the first account, the early Germanic languages display a uniform SOV base order (see Lenerz, 1984 for OHG, v. Kemenade, 1987 for Old English, Erickson, 1997 for Old Saxon), but allow for massive extraposition to the right of the verb in subordinate clauses. In line with this model, word order variation is explained as the result of rightward movement of different types of constituents (heavy DPs and VPs) while the verb always remains in its basic position at the end of the clause (v. Kemenade, 1987; Tomaselli, 1995; Axel, 2007).

This account has been challenged for Old English (OE) by Pintzuk (1991) who discovered evidence for postverbal phrases, e.g. pronouns and light adverbs, which are excluded from extraposition in modern SOV languages. To explain structural variation in the data, Pintzuk claimed that the head-complement parameter in IP and VP was not fixed in OE (also called a double base). According to this model, verb-medial orders in OE are explained partly as the result of leftward movement of the finite verb to a clause-medial IP, and partly as instances of basic VO. A refinement is proposed by Fuß and Trips (2002), who retain the idea of a double base in the VP but assume an

optional movement of the finite verb to a head-initial *v*P rather than to IP in subordinate clauses.[16]

A rather different account is proposed by Hróarsdóttir (2000) and Hinterhölzl (2004), who study word order variation in Old Icelandic and Old High German, respectively. They argue that variation in word order is a variation on the surface that is due to the expression of information-structural categories. Adopting the Universal Base Hypothesis (UBS, see Kayne, 1994), they assume that there is no variation in the base, deriving different surface orders from a unique order Spec-Head-Compl and leftward movement only. A brief overview of the results of the diachronic project's evaluation of T-CODEX via ANNIS is given in (Donhauser, 2007). Here, we restrict our search to complement and adverbial clauses for reasons of space[17].

To begin with, query (2) yields the set of subordinate clauses in which the finite verb (VFIN) is in clause-final position (190 instances). Query (3) on the other hand yields the set of subordinate clauses containing post-verbal material (128 instances, see Figure 8).

(2) `clause-status=/SUBORD.*/ & gf="VFIN" & #2 _r_ #1`

(3) `clause-status=/SUBORD.*/ & gf="VFIN" & gf & #1 _i_ #2 & #1 _i_ #3 & #2 . #3`

Further, we are able to specify the type and the phonological weight of post-verbal phrases in the result set of query (3). Here, we find heavy DPs (4) but also light objects containing a single noun only (5), as well as predicative nouns or adjectives (6), non-finite main verbs (7) and PPs (8).[18]

(4) *thô her* **gisah** *[manage thero pharisęorum / Inti sadducęorum]*
when he saw crowd DT.GEN Pharisees / and Sadducees
when he saw lots of the Pharisees and the Sadducees (T 46, 2-4)
lat. Uidens multos phariseorum / & sadduceorum

(5) *thaz sie* **gabin** *[obphar]*
that they gave sacrifice
that they gave a gift (T 37, 19)
lat. & ut darent hostiam

16. For a similar account of OHG see Weiß (2006).
17. Relative and causal clauses, which display some ambiguities in OHG, will be ignored here. Note that these types of clauses are dealt with in detail in the remaining publications of the project.
18. Verbs are highlighted by bold type, post-verbal constituents are in square brackets.

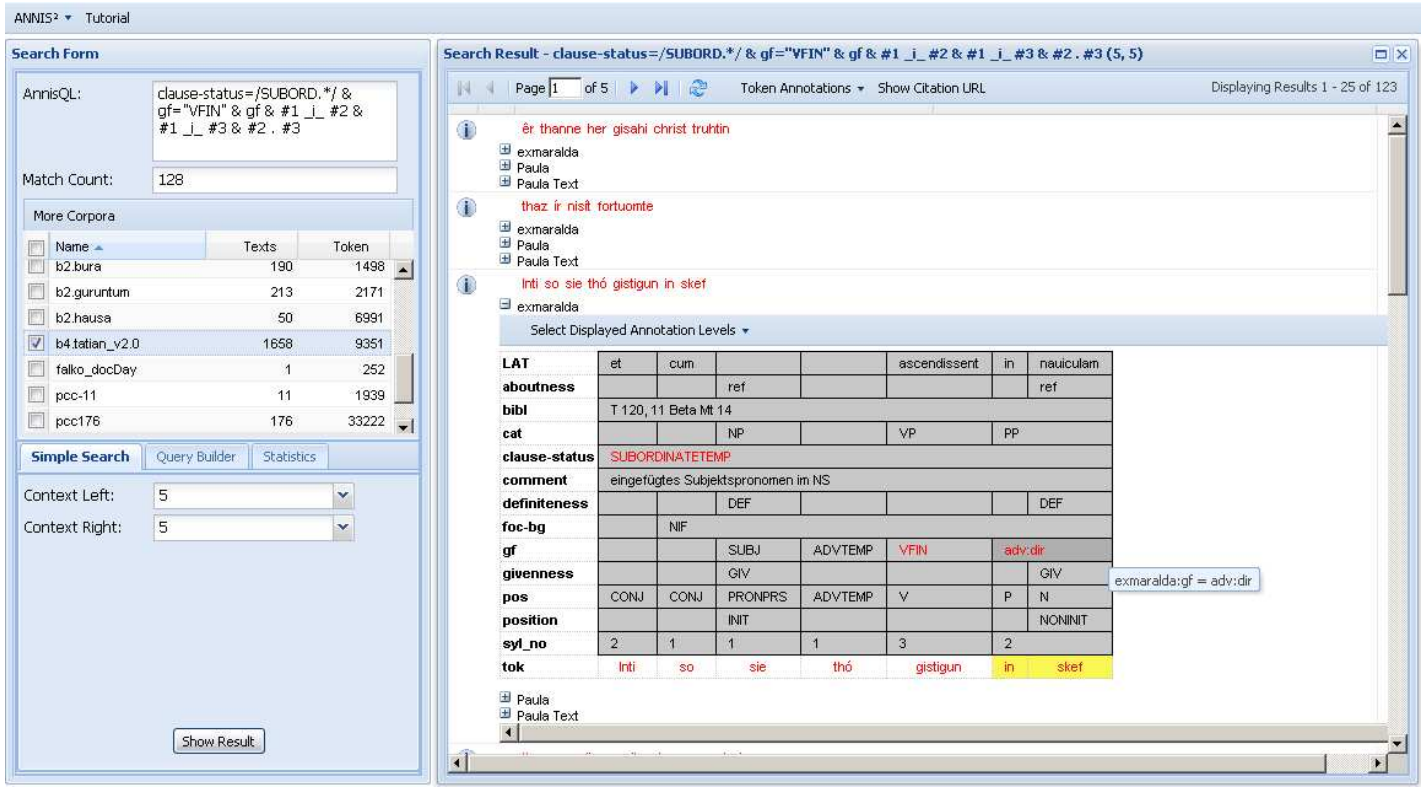


Figure 8. *Subordinate clauses with postverbal elements (result set for query (3); grid view of clause "and when they got into the boat", Matthew 14,32)*

(6) *thaz sie* **hiezzin** *[boanerges]*
that they be called Boanerges
that they be called Boanerges (T 59, 22)
lat. boanerges

(7) *thaz ír* **nisít** *[fortuomte]*
that you not be judged
that you are not judged [as well] (T 71, 17)
lat. ut non iudicemini

(8) *mit thiu hér thó* **ingieng** *[in capharnaum]*
with this he there trod in into Capharnaum
when he went to Capharnaum (T 83, 8)
lat. Cum autem introiss& capharnaum

This sheds doubt on the hypothesis that variation is due to extraposition, as some of the post-verbal constituents, e.g. light objects, predicative nouns and adjectives as well as directional PPs, as the one in (8), do not move to the right in SOV languages. Thus, we consider the hypothesis that such orders follow information-structural patterns. Based on this hypothesis, we look at the position of objects with different IS properties in the clause. Additionally, we aim at retrieving only those cases in which the difference in word order between Latin and OHG affects exactly the verb-object order in the clause. In other terms, we are looking for subordinate clauses in which a pre-verbal object in the Latin original is realised postverbally in OHG and vice versa.

For this purpose we search i) for OHG objects following the finite verb, and ii) for OHG objects preceding the finite verb independently of the word order in the original. The formulation of queries is given in (9) and (10). Query (9) produced only one hit, cf. (11), while query (10) produced 46 instances, one of which is given in (12).

(9)
```
clause-status=/SUBORD.*/ & gf=/.*O/ & gf="VFIN" & align &
align & #1 _i_ #2 & #1 _i_ #3 & #2 _i_ #4 & #3 _i_ #5 &
#3 .* #2
```

(10)
```
clause-status=/SUBORD.*/ & gf=/.*O/ & gf="VFIN" & align &
align & #1 _i_ #2 & #1 _i_ #3 & #2 _i_ #4 & #3 _i_ #5 &
#2 .* #3
```

(11) *thaz mannes sun / **hab&** [giuualt in erdu / zifurlazenne sunta]*
that man.GEN son / had power in earth / to forgive sin
that the Son of Man had the power to forgive the sins in Earth (T 89, 26-28)
lat. quod filius hominis / potestatem hab& in terra / dimittere pecca

(12) *thaz her [uueralt]* **tuome**
that he world.ACC judged
that he judged the world (T 197, 31)
lat. ut iudic& mundum

In the next step, we determine the information-structural value of the objects in the lists that the query produced according to the annotation scheme described in Section 2.2. We look for post-verbal objects annotated as *given* (13), i.e. which refer to an antecedent explicitly mentioned in the preceding context, as opposed to ones annotated as *new* (14), and likewise for preverbal *given* (15) and *new* objects (16):

(13)
```
clause-status=/SUBORD.*/ & gf=/.*O/ & gf="VFIN" & align &
align & #1 _i_ #2 & #1 _i_ #3 & #2 _i_ #4 & #3 _i_ #5 &
#3 .* #2 & givenness="GIV" & #2 _i_ #6
```

(14)
```
clause-status=/SUBORD.*/ & gf=/.*O/ & gf="VFIN" & align &
align & #1 _i_ #2 & #1 _i_ #3 & #2 _i_ #4 & #3 _i_ #5 &
#3 .* #2 & givenness="NEW" & #2 _i_ #6
```

(15)
```
clause-status=/SUBORD.*/ & gf=/.*O/ & gf="VFIN" & align &
align& givenness="GIV" & #1 _i_ #2 & #1 _i_ #3 & #2 _i_ #4 &
#3 _i_ #5 & #2 .* #3 & #2 _i_ #6
```

(16)
```
clause-status=/SUBORD.*/ & gf=/.*O/ & gf="VFIN" & align &
align & givenness="NEW" &#1 _i_ #2 & #1 _i_ #3 & #2 _i_ #4 &
#3 _i_ #5 & #2 .* #3 & #2 _i_ #6
```

Queries (13) and (16) produce no hits, while (14) and (15) retrieve 1 and 39 instances respectively.

It turns out that in OHG, in contrast to the Latin original, *given* objects are regularly placed before the finite verb (cf. query (15)). On the other hand the results of query (14) suggest that these objects appear post-verbally when they convey novel information. We can thus identify the preverbal material as background, and postverbal material as being part of the new information focus in the clause. There is no evidence for objects which represent a new discourse entity and at the same time precede the finite verb. Nevertheless, there is no one-to-one correspondence between focus and new information. This can be shown by query (17), which finds cases of preverbal objects annotated as contrastive focus (cf. (18)).

(17)
```
clause-status=/SUBORD.*/ & gf=/.*O/ & gf="VFIN" & align &
align & foc-bg="CF" & #1 _i_ #2 & #1 _i_ #3 & #2 _i_ #4 &
#3 _i_ #5 & #2 .* #3 & #2 _i_ #6
```

(18) (context: thane thu fastes/ salbo thin houbit/ Inti thin annuzi thuah="when you fast, anoint your head and wash your face", T 68, 28-30)

zithiu thaz	*thu*	*[mannon]*	**nisís**	*gisehán*	/	*fastenti*	*úzouh*	*thinemo*
so that	you	men.DAT	not be	seen	/	fasting	but	your.DAT
fater								
father								

in order to appear fasting not to men but to your father (T 68, 31-32)

lat. ne uidearis hominibus / ieiunans. Sed patri tuo

Given the fact that we find such objects as well, we must conclude that in OHG focussed objects may hold two different positions with respect to the finite verb. This corresponds to the findings of investigations which claim the existence of two distinct focus domains in OHG (cf. Hinterhölzl, 2004).

6. Discussion and ongoing research activities

In the diachronic project, T-CODEX 1.0 is currently being further analysed. The overall goal is to extend the search through different levels of the annotation in order to detect factors or combinations of factors favouring special word order patterns. Furthermore, we intend to look for differences in the quantitative distribution of competing patterns among the individual scribes. In this way, the corpus is of enormous value for detecting some ordering principles in an apparently unordered system and for identifying domains in which the establishment of general rules first applied. The enhancement of the search options in ANNIS and the integration of methods for statistical analyses are supposed to help explain the enormous variation in word order in early Germanic.

The methods of multi-level corpus annotation and retrieval developed at the SFB are indispensable for ongoing research in the field of corpus linguistic and automatic processing of historical data. The experience from this research will be implemented in the creation of a broad, fully annotated reference corpus covering the entire written tradition of the Old German period, which was recently launched at Humboldt University Berlin and the universities in Jena and Frankfurt/Main[19]. The methods we have developed so far will be enhanced in two directions: i) adding layers of annotation to the Latin parts of the records, providing lemma information and a full annotation of inflectional morphology, and ii) combining different annotation tools, e.g. for the an-

19. Referenzkorpus Altdeutsch (750-1050), funded by the German Research Foundation, funding period 2008-2013, principal investigators Karin Donhauser (Berlin), Jost Gippert (Frankfurt/M.) and Rosemarie Lühr (Jena).

notation of rhetorical relations or syntactic trees[20], which are all searchable in parallel via ANNIS.

As ANNIS provides full Unicode support, the annotation framework presented in this paper can also be applied to other languages, as has been done e.g. in a project that compares languages differing typologically rather than with respect to their stage, focussing on IS in African Languages (Chiarcos *et al.*, 2009b).

7. References

Axel K., *Studies in Old High German syntax: left sentence periphery, verb placement and verb second*, John Benjamins, Amsterdam/Philadelphia, 2007.

Brants T., Plaehn O., "Interactive corpus annotation", *Proceedings of LREC 2000*, Athens, Greece, 2000.

Carletta J., Evert S., Heid U., Kilgour J., "The NITE XML Toolkit: data model and query", *Language Resources and Evaluation Journal*, vol. 39, n° 4, p. 313-334, December, 2005.

Carletta J., Kilgour J., O'Donnell T., Evert S., Voormann H., "The NITE Object Model Library for Handling Structured Linguistic Annotation on Multimodal Data Sets.", *Proceedings of the EACL Workshop on Language Technology and the Semantic Web (Third Workshop on NLP and XML, NLPXML-2003)*, 2003.

Chiarcos C., "An ontology of linguistic annotations", *LDV Forum*, vol. 23, n° 1, p. 1-16, 2008.

Chiarcos C., Dipper S., Götze M., Leser U., Lüdeling A., Ritz J., Stede M., "A Flexible Framework for Integrating Annotations from Different Tools and Tagsets", *TAL (Traitement automatique des langues)*, 2009a.

Chiarcos C., Fiedler I., Grubic M., Haida A., Hartmann K., Ritz J., Schwarz A., Zeldes A., Zimmermann M., "Information Structure in African Languages: Corpora and Tools", *Proceedings of the EACL 2009 Workshop on African Language Technology*, Athens, Greece, 2009b.

Chiarcos C., Ritz J., Stede M., "By all these lovely tokens... Merging Conflicting Tokenizations", *Proceedings of the Third Linguistic Annotation Workshop (LAW III) 2009*, Suntec, Singapore, 2009c.

Dipper S., "XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation", *in* R. Eckstein, R. Tolksdorf (eds), *Proceedings of Berliner XML Tage*, p. 39-50, 2005.

Dipper S., Götze M., "Accessing Heterogeneous Linguistic Data — Generic XML-based Representation and Flexible Visualization", *Proceedings of the 2nd Language & Technology Conference 2005*, 2005.

20. Methods for combining different annotation tools in a historical corpus have been successfully implemented in a cross-disciplinary research project "Complex Databases" in the Linguistics Department and the Department of Computer Science at Humboldt University Berlin, funding period 2003-2008, principal investigators Karin Donhauser, Anke Lüdeling and Ulf Leser.

Dipper S., Götze M., Skopeteas S. (eds), *Information Structure in Cross-Linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics, and Information Structure*, ISIS Working papers of the SFB 632 (7), Universitätsverlag Potsdam, 2007.

Dittmer A., Dittmer E., *Studien zur Wortstellung - Satzgliedstellung in der althochdeutschen Tatianübersetzung*, Vandenhoek & Ruprecht, Göttingen, 1998.

Donhauser K., "Negationssyntax im Althochdeutschen. Ein sprachhistorisches Rätsel und der Weg zu seiner Lösung", *in* K. Donhauser, L. Eichinger (eds), *Deutsche Grammatik. Thema in Variationen*, Carl Winter, Heidelberg, p. 283-298, 1998.

Donhauser K., "Zur informationsstrukturellen Annotation sprachhistorischer Texte", *Sprache und Datenverarbeitung*, vol. 31, p. 39-45, 2007.

Erickson J., "Some observations on word order in Old Saxon", *in* K. H. Ramers, M. Schwarz (eds), *Sprache im Fokus. Festschrift für Heinz Vater zum 65. Geburtstag*, Niemeyer, Tübingen, p. 95-105, 1997.

Fleischer J., "Zur Methodologie althochdeutscher Syntaxforschung", *Beiträge zur Geschichte der deutschen Sprache und Literatur*, vol. 128, p. 25-69, 2006.

Fleischer J., Hinterhölzl R., Solf M., "Zum Quellenwert des AHD-Tatian für die Syntaxforschung: Überlegungen auf der Basis von Wortstellungsphänomenen", *Zeitschrift für germanistische Linguistik*, vol. 36, p. 210-239, 2008.

Fuß E., Trips C., "Variation and change in Old and Middle English - on the validity of the Double Base Hypothesis", *Journal of Comparative Germanic Linguistics*, vol. 4, p. 171-224, 2002.

Heid U., Voormann H., Milde J.-T., Gut U., Erk K., Padó S., "Querying both time-aligned and hierarchical corpora with NXT search", *Proceedings of LREC-2004*, Lisbon, Portugal, 2004.

Hinterhölzl R., "Language Change versus Grammar Change: What diachronic data reveal about the distinction between core grammar and periphery", *in* C. Trips, E. Fuß (eds), *Diachronic Clues to Synchronic Grammar*, John Benjamins, Amsterdam/Philadelphia, p. 131-160, 2004.

Hinterhölzl R., "The role of information structure in word order variation and word order change", *in* R. Hinterhölzl, S. Petrova (eds), *New Approaches to word order variation in Germanic*, Mouton de Gruyter, Berlin, p. 45-66, 2009.

Hinterhölzl R., Petrova S., "Rhetorical Relations and Verb Placement in Old High German Tatian Salience", *in* C. Chiarcos, B. Claus, M. Grabski (eds), *Salience. Multidisciplinary perspectives on its function in discourse*, de Gruyter, Berlin, to appear.

Hróarsdóttir T., *Word order change in Icelandic: from OV to VO*, John Benjamins, Amsterdam/Philadelphia, 2000.

Kayne R., *The Antisymmetry of Syntax*, MIT Press, Cambridge, MA/London, 1994.

Krifka M., "Basic notions of information structure", *in* C. Féry, G. Fanselow, M. Krifka (eds), *The Notions of Information Structure*, Interdisciplinary Studies on Information Structure (ISIS) 6, Working Papers of the SFB 632, Universitätsverlag Potsdam, Potsdam, p. 13-56, 2007.

Lenerz J., *Syntaktischer Wandel und Grammatiktheorie. Eine Untersuchung an Beispielen aus der Sprachgeschichte des Deutschen*, Niemeyer, Tübingen, 1984.

Lezius W., Ein Suchwerkzeug für syntaktisch annotierte Textkorpora (German), Ph.D. thesis, University of Stuttgart, Institut für Maschinelle Sprachverarbeitung, 2002.

Masser A., *Die lateinisch-althochdeutsche Tatianbilingue Stiftsbibliothek St. Gallen Cod. 56*, Vandenhoek & Ruprecht, Göttingen, 1994.

Masser A., "Syntaxprobleme im althochdeutschen Tatian", *in* Y. Desportes (ed.), *Semantik der syntaktischen Beziehungen. Akten des Pariser Kolloquiums zur Erforschung des Althochdeutschen 1994*, Carl Winter, Heidelberg, p. 123-140, 1997a.

Masser A., "Wege zu gesprochenem Althochdeutsch", *in* E. Glaser, M. Schlaefer (eds), *Grammatica Ianua Artium. Festschrift für Rolf Bergmann zum 60. Geburtstag*, Carl Winter, Heidelberg, p. 49-70, 1997b.

Molnár V., "Zur Pragmatik und Grammatik des TOPIK-Begriffs", *in* M. Reis (ed.), *Wortstellung und Informationsstruktur*, Niemeyer, Tübingen, p. 155-202, 1993.

Müller C., Strube M., "Multi-level annotation of linguistic data with MMAX2", *in* S. Braun, K. Kohn, J. Mukherjee (eds), *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, Peter Lang, Frankfurt a.M., Germany, 2006.

O'Donnell M., "RSTTool 2.4 – A Markup Tool for Rhetorical Structure Theory", *Proceedings of the International Natural Language Generation Conference (INLG'2000)*, Mitzpe Ramon, Israel, p. 253-256, 2000.

Orasan C., "PALinkA: a highly customisable tool for discourse annotation", *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, Sapporo, Japan, p. 39-43, 2003.

Petrova S., *Guidelines for the annotation of the OHG Tatian (Ms)*, Humboldt University Berlin. 2009.

Petrova S., "Information structure and word order variation in the Old High German Tatian", *in* R. Hinterhölzl, S. Petrova (eds), *Salience. Multidisciplinary perspectives on its function in discourse*, Mouton de Gruyter, Berlin, p. 251-279, to appear.

Petrova S., Hinterhölzl R., "Evidence for two types of focus positions in Old High German", *in* G. Ferraresi, R. Lühr (eds), *Proceedings of the 29th Annual Meeting of the German Linguistics Society 2006*, Siegen, forthcoming.

Petrova S., Solf M., "On the Methods of Information-Structural Analysis in Texts from Historical Corpora: A Case Study on Old High German", *in* R. Hinterhölzl, S. Petrova (eds), *Information structure and language change: New approaches to word order variation in Germanic*, Mouton de Gruyter, Berlin, p. 121-160, 2009.

Pintzuk S., Phrase structures in competition: variation and change in Old English word order, Ph.D. thesis, University of Pennsylvania, 1991.

Rehm G., Eckart R., Chiarcos C., Dellert J., "Ontology-Based XQuerying of XML-Encoded Language Resources on Multiple Annotation Layers", *Proceedings of LREC 2008*, Marrakech, Morocco, 2008.

Schmidt T., "EXMARaLDA – ein System zur computergestützten Diskurstranskription", *in* A. Mehler, H. Lobin (eds), *Automatische Textanalyse. Systeme und Methoden zur Annotation und Analyse natürlichsprachlicher Texte*, Verlag für Sozialwissenschaften, Wiesbaden, p. 203-218, 2004.

Sievers E. (ed.), *Tatian. Lateinisch und altdeutsch mit ausführlichem Glossar*, reprint 1960 of 1st edn, Schöningh, Paderborn etc., 1892.

Sonderegger S., *Althochdeutsche Sprache und Literatur. Eine Einführung in das älteste Deutsch. Darstellung und Grammatik*, 3rd edn, Mouton de Gruyter, Berlin/New York, 2003.

Tomaselli A., "Cases of Verb Third in Old High German", *in* A. Battye, I. Roberts (eds), *Clause Structure and Language Change*, Oxford University Press, New York/Oxford, p. 345-369, 1995.

v. Kemenade A., *Syntactic Case and Morphological Case in the History of English*, Foris Publications, Dordrecht, 1987.

Weiß H., *Die rechte Peripherie im Althochdeutschen. Zur Verbstellung in dass-Sätzen*, Universität Jena. 2006.

Witten I. H., Frank E., *Data mining: Practical machine learning tools and techniques*, 2nd edn, Morgan Kaufman, San Francisco, 2005.

Zeldes A., Hirschmann H., Lüdeling A., *Multilevel Learner Corpora,*. March 10-14, 2009, Presentation at the Workshop on Automatic Analysis of Learner Language 2009 (AALL'09), CALICO '09, Arizona State University, Tempe, AZ, 2009.

SERVICE ÉDITORIAL – HERMES-LAVOISIER
14 rue de Provigny, F-94236 Cachan cedex
Tél. : 01-47-40-67-67
E-mail : revues@lavoisier.fr
Serveur web : http://www.revuesonline.com