ANNIS: Complex Multilevel Annotations in a Linguistic Database

Michael Götze and Stefanie Dipper
Department of Linguistics, University of Potsdam
14415 Potsdam, Germany
{goetze,dipper}@ling.uni-potsdam.de

Abstract

We present *ANNIS*, a linguistic database that aims at facilitating the process of exploiting richly annotated language data by naive users. We describe the role of the database in our research project and the project requirements, with a special focus on aspects of multilevel annotation. We then illustrate the usability of the database by illustrative examples. We also address current challenges and next steps.

1 Introduction

Until recently, working with data that is annotated at multiple levels with different types of annotation required rather advanced computer skills, which cannot be expected from the majority of potentially interested users.

We present *ANNIS*, a linguistic database that aims at providing the infrastructure for supporting linguists in their work on multilevel annotations. We describe and illustrate the current state of our work and sketch the next steps.

In sec. 2, we present the research scenario *ANNIS* is developed for, show the role of the linguistic database therein, and sketch the major requirements it aims to fulfill. We then describe the architecture and current functionality, and discuss the way difficult aspects of multidimensional annotations are treated (sec. 3). In sec. 4, we illustrate the work with the database by three exemplary approaches. Finally, we sketch our next steps.

2 Background

Research Scenario

The database *ANNIS* is being developed in the Collaborative Research Center SFB 632 on Information Structure, which consists of 13 individual research projects from disciplines such as theoretical linguistics, psycholinguistics, first and second language acquisition, typology and historical linguistics.[1] In the research center, data of various languages is collected and annotated at the levels of phonology, morphology, syntax, semantics, and pragmatics—levels that contribute in ways yet to be determined to the information structural partitioning of discourse and utterances.

For annotation, task-specific tools are being used, e.g. *EXMARaLDA*, *annotate*, *RSTTool*, and *MMAX*.[2] Data is then converted into a standoff data interchange format, which is fed into the linguistic database *ANNIS*. *ANNIS* aims at providing functionalities for exploring and querying the data, offering suitable means for both visualization and export.

User Requirements

Central requirements evolving out of the scenario sketched above and, as we believe, for multilevel annotation in general are *Data heterogeneity*, *Data reuse*, and *Accessibility* (cf. (Dipper and Götze, 2005)).

Data heterogeneity is a result of: (i) the language data to be annotated, varying with respect to size (single sentences vs. narrations), modality (monologue vs. dialogue, text vs. speech) and language; (ii) the annotations, which use different

[1] http://www.sfb632.uni-potsdam.de/. For more information about *ANNIS*, see http://www.sfb632.uni-potsdam.de/annis/ and (Dipper et al., 2004).

[2] http://www.rrz.uni-hamburg.de/exmaralda/
http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/
http://www.wagsoft.com/RSTTool
http://mmax.eml-research.de/

data structures (attribute-value pairs, trees, pointers, etc.); and (iii) data formats that stem from different task-specific annotation tools.

Data reuse must be supported, e.g. for further or re-annotation, statistical analyses, or reuse of the data in other tools.

Accessibility of both tools and data is an obvious prerequisite for data reuse.

In the following section, we will address those aspects that are particularly relevant for these requirements and discuss their treatment in *ANNIS*.

3 ANNIS

3.1 Main Features

ANNIS is a Java servlet application that can be accessed via standard web browsers. In its current state, it is not database-backed; data is read into memory and exploited for querying and visualization in memory.[3]

Data format and interoperability The data model must be suffiently expressive for capturing the data heterogeneity sketched above, including the representation of overlapping segments, intersecting hierarchies, and alternative annotations (e.g., for ambiguous annotations). It should further facilitate the addition of new annotations.

In our approach, we use a flexible standoff XML format, the SFB-standard interchange format, as the interface format (Dipper, 2005). In this format, primary data is stored in a file that optionally specifies a header, followed by a tag `<body>`, which contains the source text. The format makes use of generic XML elements to encode data structures and annotations: `<mark>` (markable) tags specify text positions or spans of text (or spans of other markables) that can be annotated by linguistic information. Trees and graphs are encoded by `<struct>` (structure) and `<rel>` (relation) elements, which specify local subtrees. `<feat>` (feature) tags specify the information that is annotated to markables or structures, which are referred to by `xlink` attributes. Each type of annotation is stored in a separate file, hence, competing or ambiguous annotations can be represented in a straightforward way: by distributing them over different files.

Our format allows us to represent different kinds of annotations in a uniform way. We provide importers for the export format of the annotation tools *annotate*, *EXMARaLDA*, *RST Tool*, and *MMAX*. Our *PCC* corpus (see sec. 4) imports and synchronizes the following annotations, which have been annotated by these tools: syntax, information structure, rhetorical structure, and coreference.

Visualization Suitable means for visualizing information is crucial for exploring and interpreting linguistic data. Due to the high degree of data heterogeneity, special attention has been paid to the support of visualizing various data structures. In addition, annotations may refer to segments of different sizes, e.g. syntax vs. discourse structure. Furthermore, richness of information in multilevel annotations has to be taken into account; this requires a certain degree of user-adaptivity, allowing the user to modify the way information of interest is displayed.

In *ANNIS*, we start from a basic interactive tier-based view, which allows for a compact simultaneous representation of many annotation types and whose appearance can be modified by the user in a format file. In addition, a discourse view helps the user to orient him/herself in the discourse. Further views can be added.

Query support Among the numerous requirements for a good query facility for multilevel annotation, expressiveness, efficiency, and user-friendly query-formulation appear to be the most relevant. Even a very brief discussion of these issues would go beyond the limits of this paper, the reader is instead referred to (Heid et al., 2004).

Currently, *ANNIS* uses a query language prototype which allows the user to query text and annotations, by means of regular expressions and wildcards, and various common relational operators (e.g. for stating relations in tree structures, such as dominance or sibling relations). However, the set for querying sequential relations is not sufficiently expressive, and querying co-reference relations is not supported yet. Furthermore, user support for formulating queries is rather poor.

3.2 Open Issues

Data alignment Alignment of annotations created by different annotation tools appears to be most suitable at the level of tokens. However, tools often come with their own tokenizers and mismatches do occur frequently. We currently use a

[3] For a more elaborate discussion of the basic concepts of *ANNIS*, see (Dipper et al., 2004).

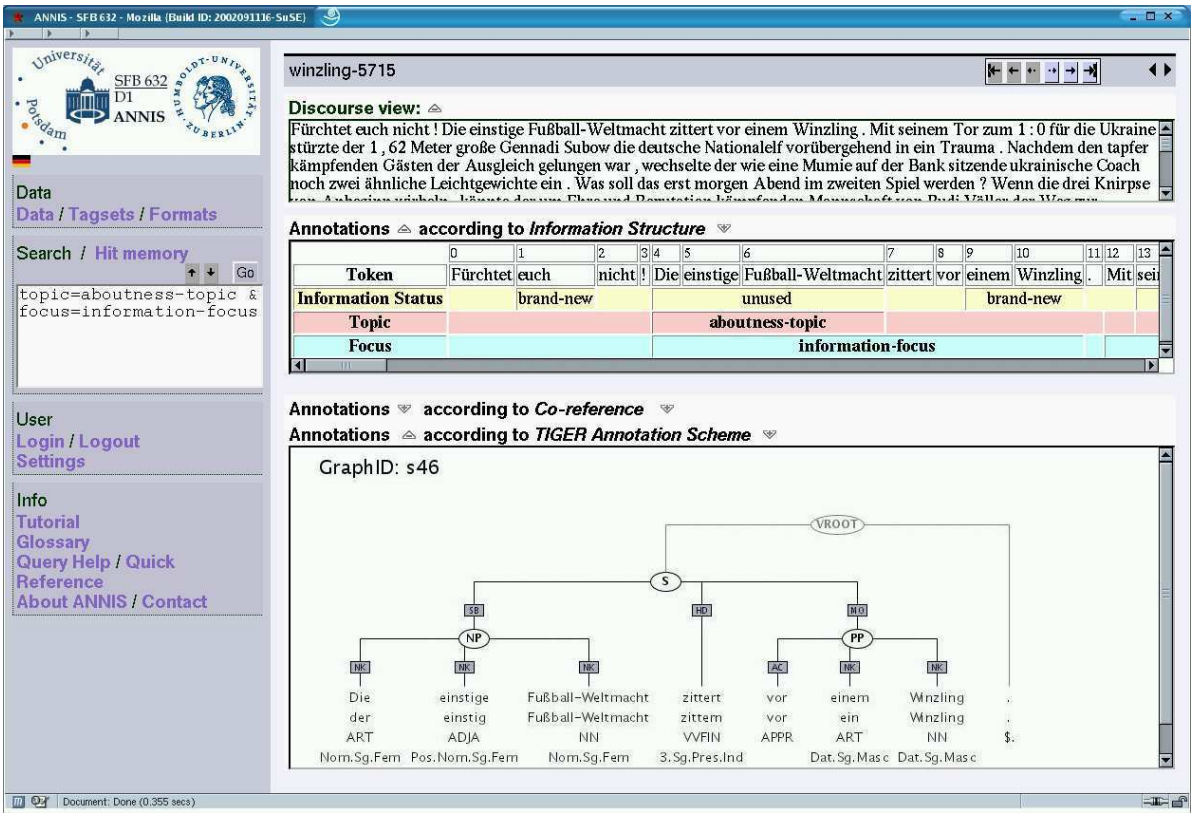

Figure 1: The ANNIS user interface, displaying data from the *PCC*

simple script that checks for text and token identity in the standoff files that we generate from the output of the individual tools. However, all mismatches have to be corrected manually. At least for white-space differences, an automatic fixing procedure should be feasible (similar to the one implemented by (Witt et al., 2005)).

Efficient Querying Current querying is restricted to rather small amounts of data, and complex queries may take some time until finishing the search.

Overlapping elements and intersecting hierarchies The query language does not yet support comfortable searching for overlapping elements. However, exactly what kinds of queries on overlapping segments or intersecting relations should be supported is an open question.

4 Use Cases

We illustrate the use of *ANNIS* in linguistic research, exemplified with research questions from three different linguistic areas.

Historical investigations The project B4: *The role of information structure in the development of word order regularities in Germanic* investigates the verb-second phenomenon, which occurred in certain Germanic languages only (e.g., it did in Modern German, but not in Modern English). One of their findings is that verb placement in the Old High German translation of Tatian correlates with discourse relations: verb-initial sentences usually occur in narrative contexts and signal continuation of the story. In contrast, verb-second sentences indicate subordinative relations (Hinterhölzl and Petrova, 2005).

Typological studies In the research project *D2: Typology of Information Structure* (cf., e.g., (Götze et al., To appear)), a typological questionnaire is designed, with which language data can be elicited using largely language-independent methods. Currently, data from 13 different languages is elicited and annotated with information from various linguistic levels (morphosyntax, phonology, semantics, and information structure).

An interesting query might look for nominal phrases (*const=np*) that are new in the discourse (*given=new*) and belong to the (information-) focus of a sentence (*focus=ans*), e.g. for investigating the phonological realization of these.

The according query has the form: *const=np & given=new & focus=ans & #1_=_#2*.[4]

Queries in *ANNIS* can be restricted to subsets of a corpus, by queries such as *focus=ans & doc=*81-11**, which searches for all answer foci in the data that has been elicited by means of the task 81-11 in the questionnaire, yielding matching data from all languages in our database.

Discourse studies The Potsdam Commentary Corpus, *PCC* (Stede, 2004), consists of 173 newspaper commentaries, annotated for morphosyntax, coreference, discourse structure according to Rhetorical Structure Theory, and information structure.

A question of interest here is the information-structural pattern of sentences introducing discourse segments that elaborate on another part of the discourse: *elaboration & rel=satellite & (cat=vroot & aboutness-topic) & #1 > #2 & #2_=_#3*. Another research issue is the relationship of coreference and discourse structure. However, querying for coreference relations is not supported yet.

5 Future Work

Currently we are working on integrating a native XML database into our system. To make processing more efficient, we are developing an internal inline representation of the standoff interchange format, encoding overlapping segments by means of milestones or fragments (Barnard et al., 1995).

Furthermore, the query language will be extended to cover different kinds of queries on sequential relations as well as coreference relations. Finally, we will add basic statistical means to the query facility, which, e.g., can point to rare and, hence, potentially interesting feature combinations.

6 Demo

In our demonstration of *ANNIS*, we will show example data from the PCC, Old High German, and data elicited by the typological questionnaire. We then illustrate by means of example queries how the researchers make use of our database in their daily work, as described above. This includes presenting the visualization and querying facilities of *ANNIS*.

[4]The expression *#n* refers to the *n*th constraint stated in the query; the binary operator _=_ requires extensional identity (Dipper et al., 2004).

References

David Barnard, Lou Burnard, Jean-Pierre Gaspart, Lynne A. Price, C. M. Sperberg-McQueen, and Giovanni Batista Varile. 1995. Hierarchical encoding of text: Technical problems and SGML solutions. *Text Encoding Initiative: Background and Context. Special Issue of Computers and the Humanities*, 29(211–231).

Stefanie Dipper and Michael Götze. 2005. Accessing heterogeneous linguistic data — generic XML-based representation and flexible visualization. In *Proceedings of the 2nd Language & Technology Conference 2005*.

Stefanie Dipper, Michael Götze, Manfred Stede, and Tillmann Wegst. 2004. ANNIS: A linguistic database for exploring information structure. In Shinichiro Ishihara, Michaela Schmitz, and Anne Schwarz, editors, *Interdisciplinary Studies on Information Structure (ISIS)*, volume 1, pages 245–279. Universitätsverlag Potsdam, Potsdam, Germany.

Stefanie Dipper. 2005. XML-based stand-off representation and exploitation of multi-level linguistic annotation. In *Proceedings of Berliner XML Tage 2005 (BXML 2005)*, pages 39–50, Berlin, Germany.

Michael Götze, Torsten Roloff, Stavros Skopeteas, and Ruben Stoel. To appear. Exploring a cross-linguistic production data corpus. In *Proceedings of the Sixth International Tbilisi Symposium on Language, Logic and Computation*. Batumi, Georgia.

Ulrich Heid, Holger Voormann, Jan-Torsten Milde, Ulrike Gut, Katrin Erk, and Sebastian Padó. 2004. Querying both time-aligned and hierarchical corpora with NXT Search. In *Proceedings of the Forth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1455–1458, Lisbon.

Roland Hinterhölzl and Svetlana Petrova. 2005. Rhetorical relations and verb placement in early germanic languages. Evidence from the Old High German Tatian translation (9th century). In M. Stede, C. Chiarcos, M. Grabski, and L. Lagerwerf, editors, *Salience in Discourse. Multidisciplinary Approaches to Discourse*, pages 71–79.

Manfred Stede. 2004. The Potsdam Commentary Corpus. In *Proceedings of the ACL Workshop on Discourse Annotation*, pages 96–102, Barcelona.

Andreas Witt, Daniela Goecke, Felix Sasaki, and Harald Lüngen. 2005. Unification of XML documents with concurrent markup. *Literary and Linguistic Computing*, 20(1):103–116.