ANNIS: a search tool for multi-layer annotated corpora

Amir Zeldes, Julia Ritz+, Anke Lüdeling* and Christian Chiarcos+*
*Humboldt-Universität zu Berlin +Potsdam University
amir.zeldes@rz.hu-berlin.de, julia@ling.uni-potsdam.de, anke.luedeling@rz.hu-berlin.de, chiarcos@uni-potsdam.de

Abstract

ANNIS (see Dipper & Götze 2005; Chiarcos et al. 2008) is a flexible web-based corpus architecture for search and visualization of multi-layer linguistic corpora. By multi-layer we mean that the same primary datum may be annotated independently with (i) annotations of different types (spans, DAGs with labelled edges and arbitrary pointing relations between terminals or non-terminals), and (ii) annotation structures that possibly overlap and/or conflict hierarchically. In this paper we present the different features of the architecture as well as actual use cases for corpus linguistic research on such diverse areas as information structure, learner language and discourse level phenomena.

The supported search functionalities of ANNIS2 include exact and regular expression matching on word forms and annotations, as well as complex relations between individual elements, such as all forms of overlapping, contained or adjacent annotation spans, hierarchical dominance (children, ancestors, left- or rightmost child etc.) and more. Alternatively to the query language, data can be accessed using a graphical query builder. Query matches are visualized depending on annotation types: annotations referring to tokens (e.g. lemma, POS, morphology) are shown immediately in the match list. Spans (covering one or more tokens) are displayed in a grid view, trees/graphs in a tree/graph view, and pointing relations (such as anaphoric links) in a discourse view, with same-colour highlighting for coreferent elements. Full Unicode support is provided and a media player is embedded for rendering audio files linked to the data, allowing for a large variety of corpora.

Corpus data is annotated with automatic tools (taggers, parsers etc.) or task-specific expert tools for manual annotation, and then mapped onto the interchange format PAULA (Dipper 2005), where stand-off annotations refer to the same primary data. Importers exist for many formats, including EXMARaLDA (Schmidt 2004), TigerXML (Brants & Plaehn 2000), MMAX2 (Müller & Strube 2006), RSTTool (O'Donnell 2000), PALinkA (Orasan 2003) and Toolbox (Stuart et al. 2007). Data is compiled into a relational DB for optimal performance. Query matches and their features can also be exported in the ARFF format and processed with the data mining tool WEKA (Witten & Frank 2005), which offers implementations of clustering and classification algorithms. ANNIS2 compares favourably with search functionalities in the above tools as well as other corpus search engines (EXAKT, http://www.exmaralda.org/exakt.html, TIGERSearch, Lezius,2002, CWB, Christ 1994) and other frameworks/architectures (NITE, Carletta et al. 2003, GATE, Cunningham, 2002).

1. Introduction

1.1 Motivation and definitions

Recent years have seen a move beyond traditionally inline annotated single-layered corpora towards new multi-layer (sometimes also called multilevel) architectures, offering richer, deeper and more diverse annotations (e.g. Bański & Przepiórkowski 2009; Bernsen et al. 2002; Dipper 2005; Kruijff-Korbayová & Kruijff 2004; Vieira et al. 2003; see Wittenburg 2008 for an overview of work in the multimodal context). Despite intense work on data representations and annotation tools, there has been comparatively less work on the development of architectures affording convenient access to such data (though see Section 4.2 for other work in this area). The present paper is concerned with search and visualization of multi-layer corpora for research in corpus linguistics. For the purposes of this discussion, we understand linguistic corpora to mean any collection of language data, whether written texts or speech transcriptions as well as multimodal data, that have been collected according to principled design criteria for the purposes of linguistic analysis, and annotation layers to mean any enrichment of such raw data with additional analyses, so that one may refer to each different type of analysis as an annotation layer (cf. e.g. Biber 1993, Leech 1997). While the term multi-layer itself only implies several different types of annotation, such as part of speech tagging or lemmatization (see Schmid 2008 and Fitschen & Gupta 2008 for overviews), we use this term to refer more specifically to annotations that may be created independently of each other, annotating the same phenomenon from different points of view, or different phenomena altogether. Annotations can refer both to the same spans of text or possibly to different divisions of the data (discourse structural, as in discourse referents or rhetorical units; typographic, as in chapters, paragraphs or orthographic sentences; syntactic, as in clauses or phrases; multimodal, as in audiovisual events such as gestures or changes in prosody). Multi-layer corpora and the architectures supporting them should therefore have the feature that two annotators may enrich different, possibly overlapping parts of the same corpus with different annotation values in different schemes, or even in the same scheme (i.e. repeated conflicting annotations, such as multiple syntax trees for one sentence according to different theories).

1.2 Background for the development of ANNIS

ANNIS, which stands for ANNotation of Information Structure, is an open-source web-based architecture for search and visualization of multi-layer corpora, developed within Collaborative Research Centre 632 on Information Structure, as part of the project "Linguistic Database for Information Structure: Annotation and Retrieval" (see also Dipper & Götze 2005; Chiarcos et al. 2008). As a service provider for a variety of linguistic projects involved in research on information structure (see http://www.sfb632.uni-potsdam.de/projects_eng.html for an overview of current and completed projects), many of which use empirical data on a wide variety of languages (ranging from Old High German to Chadic and West African languages, see e.g. Petrova et al., to appear; Chiarcos et al. 2009), a central requirement for the project is offering support for diverse annotations according to multiple annotation schemes. In particular, information structure interacts with all levels of grammar (phonology, prosody, morphology and syntax, semantics and pragmatics) and manifests itself in three key areas: informational status (givenness, newness), topic-comment structure and focus vs. background (see Krifka 2007; for annotation guidelines for these areas see Dipper et al. 2007). Simultaneously annotating such different types of information can prove very difficult for one annotator using one specific annotation tool, and designing appropriate annotation applications for this challenge is both hindered by needs which cannot be predicted at the outset of novel annotation types, and also redundant, since there are suitable tools for many of the subtasks involved (e.g. syntactic or discourse annotation). We therefore suggest that independently annotating separate factors in a multi-layer fashion can afford corpus designers several advantages:

1. Distributing annotation work collaboratively, so that annotators can specialize on specific subtasks and work concurrently
2. Retroactively adding annotations to existing corpora
3. Using different annotation tools suited to different tasks
4. Allowing multiple annotations of the same type to be created and evaluated, which is important for controversial layers with different possible tag sets or low inter-annotator agreement
5. Retroactively modifying subsets of annotations or tag sets selectively becomes possible, which is particularly essential for annotation schemes that are not yet established and still undergoing consolidation

The challenges which must be overcome to reap these benefits are how to model, query and visualize data and categories relevant to linguistic research in a way that is modular and user-friendly, yet at the same time powerful, performant and expressive. For an infrastructure to deliver these possibilities, it must be able to represent and simultaneously search through such prevalent annotation types as words or spans of words bearing features (e.g. morphological or information structural information), graphs such as syntax trees, and complex interrelations such as coreference. With these goals in mind, we will now outline the features of ANNIS2 and its potential for opening new possibilities of research on richly annotated corpora. The remainder of this paper is structured as follows: The following section briefly describes the corpus search architecture itself and its query language, AQL. Section 3 introduces some use cases for multi-layer corpora using ANNIS2: error-annotated learner corpora, constituent syntactic treebanks, and anaphoric banks. Section 4 examines ANNIS2 in the context of related tools: the integration of data from different sources in PAULA XML and a comparison of the approach taken in ANNIS2 to some other relevant systems and frameworks. Section 5 concludes with an outlook on open challenges for the further open-source development of ANNIS.

2. Deploying multi-layer corpora in ANNIS2

2.1 ANNIS2 system architecture

The ANNIS2 system consists of three components: a web interface, written in JAVA using the client-side JavaScript framework ExtJS (http://www.extjs.com/); a relational database backend, using the open source database PostGreSQL, (http://www.postgresql.org/); and an RMI service for communication between the two (see Figure 1). All components build on freely available technology and are offered open-source under the Apache License 2.0 (http://www.apache.org/licenses/).

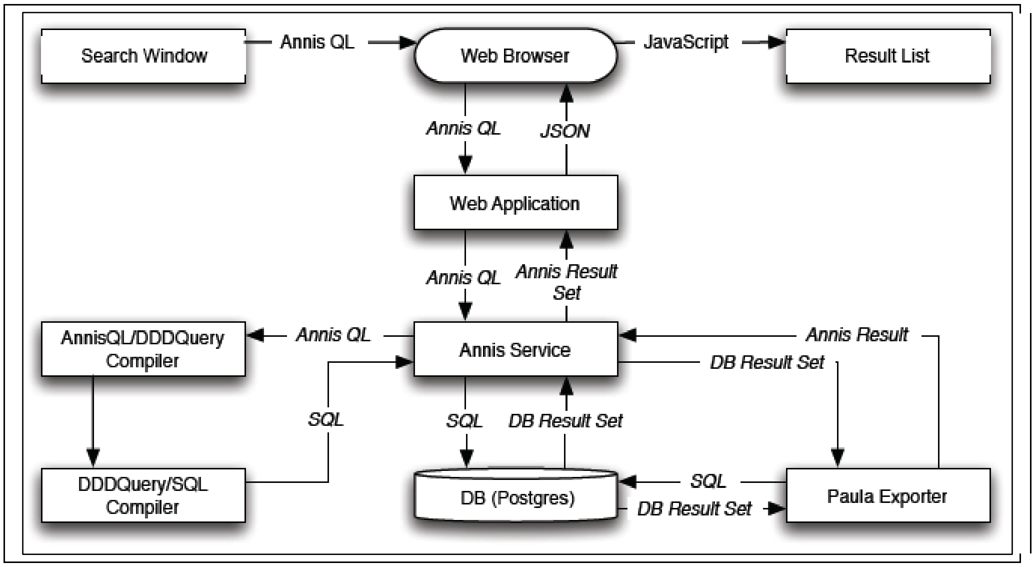


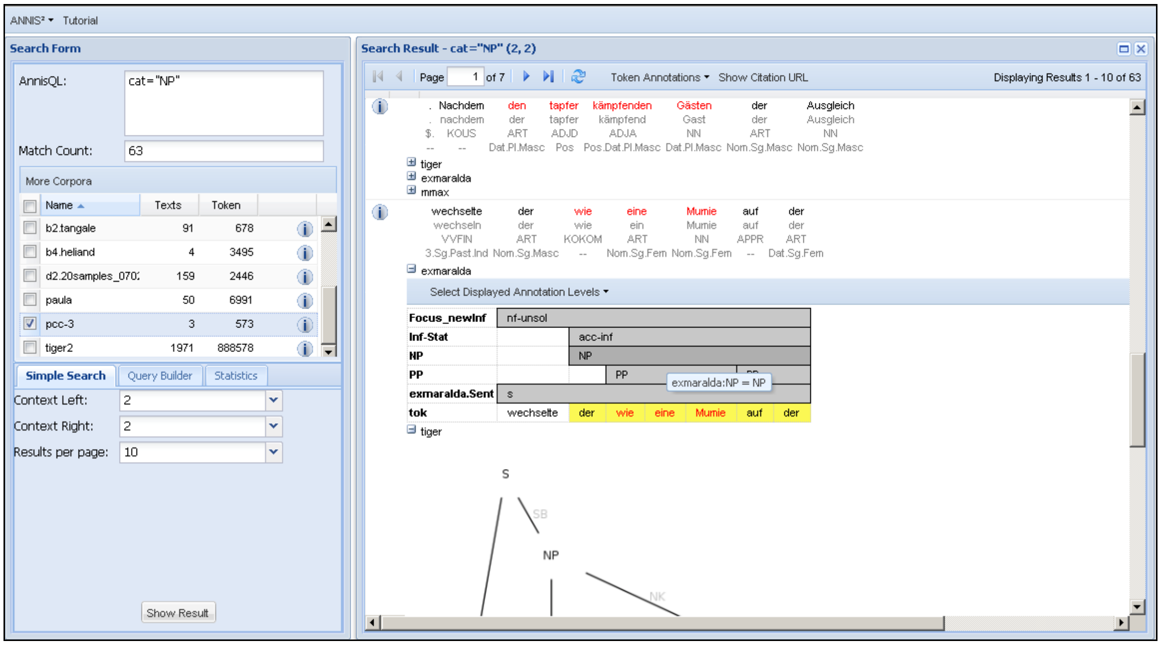
Fig. 1: Structure of ANNIS2

The web interface offers users a familiar window-based environment to input queries and view results in the corpora to which they are allowed access (see Figure 2). The corpus list is customizable, permitting users to define a list of 'favourite' corpora, hiding undesired corpus data. Queries are entered using the ANNIS Query Language (AQL, see Section 2.2). Optionally, a graphical query builder can be used to generate AQL queries. In both cases, ANNIS allows using both exact and regular expressions to search for graphs with nodes corresponding to text and annotations in multiple corpora, as well as metadata at the corpus, sub-corpus and document levels. Queries may also be saved as citable links to allow other researchers with access rights to reproduce published results.

Once a query has been entered, it is validated and subsequently translated into SQL by the ANNIS Service using the DDDQuery Language as an intermediate language (see Dipper et al. 2004; Faulstich et al. 2005; Chiarcos et al. 2008). For performance reasons, we use the open source PostGreSQL database instead of searching directly through underlying XML data (see Section 4.1), which is converted into a relational format by the graph-based converter Pepper, supplied with the software. Graphs are normalized and segregated according to edge types before being indexed using pre- and postorder sequences (see Grust et al. 2004). This allows for fast searches along XPath axes (searches for nodes' parents, descendents etc.) and the separation of graph paths according to edge types (e.g. searching separately through syntactic dominance or coreference paths, see the queries in Sections 3.2 and 3.3).

After retrieving query results, the database sends all hits along with the annotations applying to them to the ANNIS Service, which can either export matches in the ARFF format for further processing with the data mining tool WEKA (Witten & Frank 2005), or pass them on to the web interface for visualization. Since multi-layer corpora combine data from different sources and types of annotations, search matches are first visualized in an adjustable context window in the Key-Word-in-Context style

(KWiC, cf. Wynne 2008) showing only annotations applying to the token stream (e.g. part of speech, lemma, morphology etc.). Further layers are visualized on demand by expanding them with a mouse click, including aligned multimodal data, such as audio files corresponding to the retrieved matches. Selected annotation fields may be hidden within the annotation layers if they are not of interest or distracting. Figure 2, for example, shows two corpus hits, for the second of which a grid visualization of span annotations and a graph visualization for a syntax tree have been expanded (the visualizations are described in more depth within the use cases in Section 3).



AnnisQL: cat="NP"

Match Count: 63

	Name ▲	Texts	Token
□	b2.tangale	91	678
□	b4.heliand	4	3495
□	d2.20samples_070:	159	2446
□	paula	50	6991
☑	pcc-3	3	573
□	tiger2	1971	888578

Context Left: 2

Context Right: 2

Results per page: 10

Page 1 of 7 — Displaying Results 1 - 10 of 63

```
  .  Nachdem    den       tapfer   kämpfenden      Gästen       der          Ausgleich
  .  nachdem    der       tapfer   kämpfend        Gast         der          Ausgleich
 $.  KOUS       ART       ADJD     ADJA            NN           ART          NN
 --  --         Dat.Pl.Masc  Pos   Pos.Dat.Pl.Masc Dat.Pl.Masc  Nom.Sg.Masc  Nom.Sg.Masc
⊞ tiger
⊞ exmaralda
⊞ mmax

 wechselte     der          wie     eine         Mumie        auf    der
 wechseln      der          wie     ein          Mumie        auf    der
 VVFIN         ART          KOKOM   ART          NN           APPR   ART
 3.Sg.Past.Ind Nom.Sg.Masc  --      Nom.Sg.Fem   Nom.Sg.Fem   --     Dat.Sg.Fem
⊟ exmaralda
```

Focus_newInf	nf-unsol						
Inf-Stat		acc-inf					
NP		NP					
PP			PP			PP	
exmaralda.Sent	s						
tok	wechselte	der	wie	eine	Mumie	auf	der

⊟ tiger

Fig. 2: Search results with an annotation grid and a syntax tree expanded.

2.2 Queries in AQL

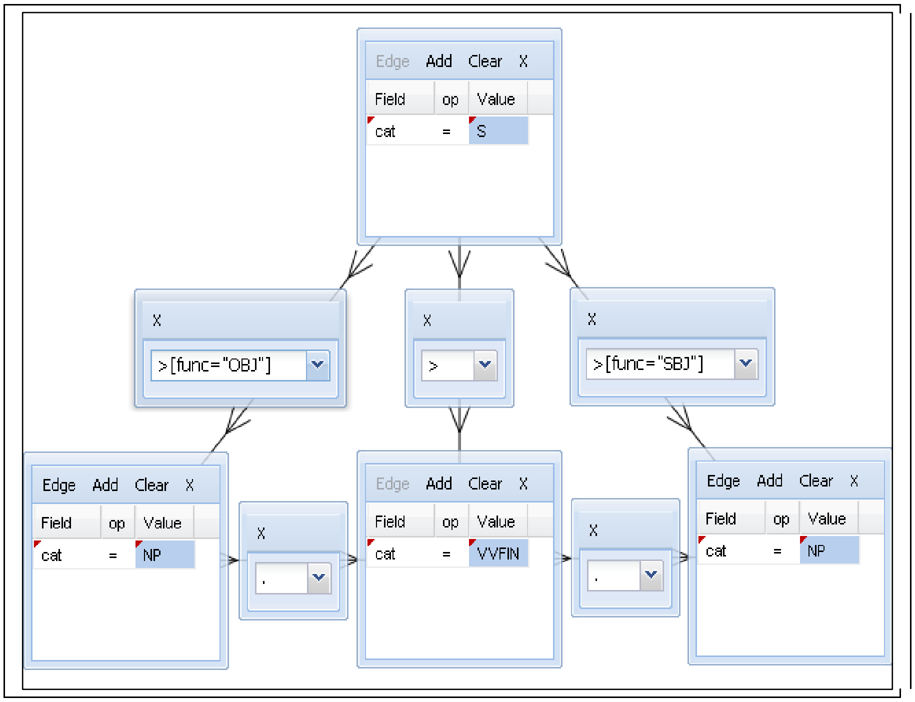
In order to carry out searches on diverse data structures, ANNIS2 uses a simple yet versatile query language based on the definition of search elements, such as tokens (usually word-forms), non-terminal nodes and annotations, and the relationships between them (cf. NiteQL, Heid et al. 2004, Carletta et al. 2005; and the TIGERSearch query language, Lezius 2002). Each element is specified as a key-value pair, as in the search for a lemma in (1), or a Regular Expression matching a variety of possible values, as in (2). When multiple elements are declared, they are conjoined with '&', and a relationship between the first element (designated as #1) and the second element (#2) must be established using an operator, as in (3), where the lemma 'house' is said to follow an article with the CLAWS7 (see http://ucrel.lancs.ac.uk/claws7tags.html) part of speech tag 'AT', using the 'direct precedence' operator, '.' (dot). Searches are case sensitive by default.

(1) `lemma="house"`
(2) `lemma=/[Hh]ous(e|ing)/`
(3) `pos="AT" & lemma="house" & #1 . #2`

Note that the designations ‘lemma’, ‘pos’ or any other annotation names are arbitrary and not inherent to the system – each corpus may use different labels and tag sets. A selection of the most important operators, such as syntactic dominance and overlapping annotation spans, is introduced throughout the use cases in Section 3. For a complete list of operators see http://www.sfb632.uni-potsdam.de/d1/annis/. Finally, it is possible to specify metadata conditions which must apply to the returned matches. These are also key-value pairs preceded by the prefix *meta::* and which may not be bound by operators, as in (4), which searches in texts where *meta::genre* is annotated as ‘academic’ for a verb (tags starting with ‘VB’) followed by a particle (tag ‘RP’) within 2 to 5 tokens (with the operator .2,5), this time using the Penn tag set (Bies et al. 1995, see also the queries in Section 3.3):

(4)
```
penn:pos=/VB.*/ &
penn:pos="RP" &
#1 .2,5 #2 &
meta::genre="academic"
```

Namespaces like *penn:* before annotation names like *pos* are optional. They may be used to distinguish between multiple annotations of the same name, e.g. *penn:pos* vs. *claws:pos*. In such cases, searching for *pos* alone would find hits in both annotation sets simultaneously. Since keeping track of variable numbers (#1, #2, etc.) may become difficult, a graphical query builder is also provided in the ANNIS2 web-interface (Figure 3). The query builder closely corresponds to the structure of the query language, defining search nodes as small boxes with key-value pairs, and the relationships between those nodes using edges labeled with the selected operator from a list.



cat = S

>[func="OBJ"]

>

>[func="SBJ"]

cat = NP

.

cat = VVFIN

.

cat = NP

Fig. 3: The ANNIS2 query builder representing a search for a sentence (cat="S") dominating (the operator >) an object NP, a finite verb and a subject NP, in that order.

Further, more complex examples are discussed below and in the tutorial distributed with ANNIS2.

3 Use cases

The following three sub-sections aim to exemplarily showcase the search and visualization strategies in ANNIS2, by going through three types of increasingly complex data: flat span annotations, hierarchical syntax trees and directed pointing relations used for coreference annotation. Each use case focuses on a different area of linguistic research with different types of corpora: learner language in learner corpora, syntax in a treebank and discourse annotations in an anaphoric bank.

3.1 Falko – an error annotated learner corpus of German

As an area of research dealing with complex, diverse and non-standard phenomena, the study of learner language is a natural scenario for the use of multi-layer corpora. In this section we will show the use of searches for spans of annotated tokens in learner data, using the operators:

- '.' (precedence, A . B means that the last token covered by A directly precedes the first token of B)
- '.*' (indirect precedence, A .* B means that the tokens under A precede the first token of B, though there may be more tokens between A and B)
- '_i_' (inclusion, A _i_ B means that the token span annotated by A includes at least the same tokens as the span annotated by B)
- '_o_' (overlap, A _o_ B means that at least some tokens are annotated by both A and B)
- '_=_' (identical coverage, A _=_ B means that A and B annotate the same span of tokens)
- '_l_' (left aligned, A _l_ B means that the spans of A and B begin at the same token)
- '_r_' (right aligned, A _r_ B means that the spans of A and B end at the same token)

An extension of the precedence operator also allows a search for an exact number or range of tokens between A and B:

- '.*n*' (e.g. A .2 B means that there are exactly 2 tokens between A and B)
- '.*n*,*m*' (e.g. A .2,5 B means there are between 2 and 5 tokens between A and B)

As an example for using these operators on corpus data, we consider the case of the error-annotated learner corpus of German as a foreign language, Falko (http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko, Lüdeling et al. 2008). The Falko corpus consists of multiple sub-corpora exemplifying different methodologies in learner corpus design: the Essays sub-corpus contains argumentative texts on one of four topics selected to match topics from the International Corpus of Learner English (ICLE, Granger 1993) for comparability. It contains texts collected from advanced German learners of over 30 native languages, as well as a comparable corpus collected from German natives. The Summaries sub-corpus contains summaries of scientific texts, similarly collected from diverse advanced learners and comparable natives – this corpus is used in the examples below. The sub-corpus Falko GU contains longitudinal studies of learners over 4 years of study, with no comparable control group. Finally, an extension of the Essays corpus is currently being prepared as part of the WHIG Project (What's Hard in German, a jointly funded DFG/AHRC project), in cooperation with Bangor University. A key interest in research on Falko revolves around the question of identifying structural acquisition difficulties in L2 German as manifested in advanced learners, who have already mastered the basics of German morphosyntax.

The following example queries show the use of ANNIS2 in two of the main paradigms in learner corpus research, namely error analysis and contrastive interlanguage analysis (see Granger et al. 2002 for an overview). The error analysis in Falko is based on the assumption that the annotation of errors always implies a target hypothesis describing what the annotator believes a native would have said in the learner's stead (Lüdeling 2008). Since target hypotheses are highly subjective, learner texts in the corpus contain spans representing target hypotheses for every erroneous segment. The Falko corpus also contains other types of spans (see Doolittle 2008), including an annotation scheme for topological fields designating, among other things, the pre- and postverbal areas of main clauses and the configuration of material around the German 'sentence brackets' in main and subordinate clauses (for the topological model of German syntax see the overview in Dürscheid 2007). Using these annotations it is possible, for example, to search for learner errors restricted to the so-called German 'Mittelfeld' (middle-field), the domain between the finite verb and its possible infinitive complements, which allows for particularly complex word-order variation:

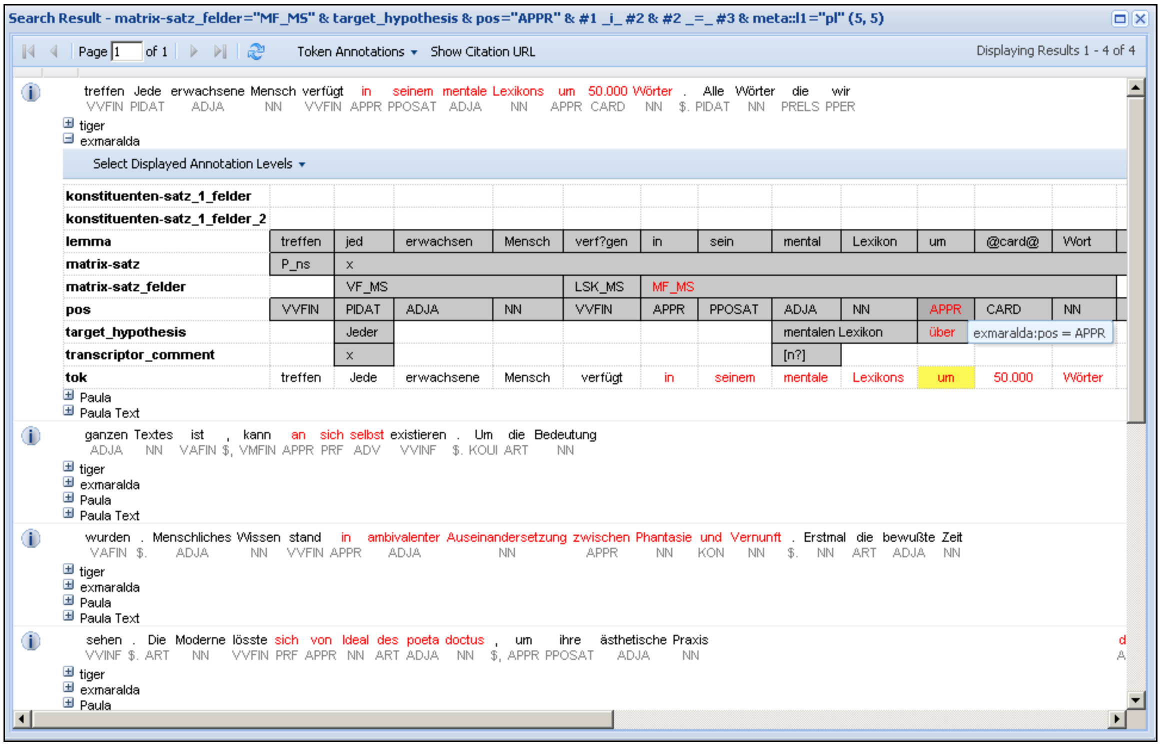
```
(5) matrix-satz_felder="MF_MS" &
    target_hypothesis &
    #1 _i_ #2 &
    meta::l1=/[^(de)]/
```

Query (5) searches for a middle-field in a main clause (the value MF_MS) and an error target hypothesis. The third line specifies that the area encompassed by the first element must include (the operator '_i_') the area encompassed by the second element. Finally, the metadatum l1 (the author's native language) is specified to match a regular expression ruling out German (de, the ISO 2-character designation for *Deutsch* 'German'). Note that the second element is specified as an annotation name with no value, meaning that an annotation must be present, but its value is irrelevant. The configuration #1 _i_ #2 means that the middle-field may overlap with the error partly or completely, but the error may not be larger than the middle-field. It is also possible to specify exactly identical coverage with the operator '_=_' as in query (6), which searches for errors exactly overlapping a preposition in middle-fields produced by native Polish learners of German:

```
(6) matrix-satz_felder="MF_MS" &
    target_hypothesis &
    pos="APPR" &
    #1 _i_ #2 &
```

```
#2 _=_ #3 &
meta::l1="pl"
```

This time a third element has been added, with the POS tag corresponding to a preposition. The fifth row adds the condition that the second element, the target hypothesis, correspond exactly ('_=_') to the span covered by the preposition. The results of query (6) are shown in Figure 4.


Search Result - matrix-satz_felder="MF_MS" & target_hypothesis & pos="APPR" & #1 _i_ #2 & #2 _=_ #3 & meta::l1="pl" (5, 5)

Page 1 of 1
Token Annotations
Show Citation URL
Displaying Results 1 - 4 of 4

treffen Jede erwachsene Mensch verfügt in seinem mentale Lexikons um 50.000 Wörter . Alle Wörter die wir
VVFIN PIDAT ADJA NN VVFIN APPR PPOSAT ADJA NN APPR CARD NN $. PIDAT NN PRELS PPER

tiger
exmaralda
Select Displayed Annotation Levels

konstituenten-satz_1_felder												
konstituenten-satz_1_felder_2												
lemma	treffen	jed	erwachsen	Mensch	verf?gen	in	sein	mental	Lexikon	um	@card@	Wort
matrix-satz	P_ns	x										
matrix-satz_felder		VF_MS			LSK_MS	MF_MS						
pos	VVFIN	PIDAT	ADJA	NN	VVFIN	APPR	PPOSAT	ADJA	NN	APPR	CARD	NN
target_hypothesis		Jeder						mentalen Lexikon		über	exmaralda:pos = APPR	
transcriptor_comment		x						[n?]				
tok	treffen	Jede	erwachsene	Mensch	verfügt	in	seinem	mentale	Lexikons	um	50.000	Wörter

ganzen Textes ist , kann an sich selbst existieren . Um die Bedeutung
ADJA NN VAFIN $, VMFIN APPR PRF ADV VVINF $. KOUI ART NN

wurden . Menschliches Wissen stand in ambivalenter Auseinandersetzung zwischen Phantasie und Vernunft . Erstmal die bewußte Zeit
VAFIN $. ADJA NN VVFIN APPR ADJA NN APPR NN KON NN $. NN ART ADJA NN

sehen . Die Moderne lösste sich von Ideal des poeta doctus , um ihre ästhetische Praxis d
VVINF $. ART NN VVFIN PRF APPR NN ART ADJA NN $, APPR PPOSAT ADJA NN A

Fig. 4: Results of a query for middle-field errors overlapping prepositions exactly. A gridview is expanded to show an error using the preposition *um* instead of *über* for *verfügt … über 50.000 Wörter* ‘has 50,000 words at his disposal’.

Finally we consider an example of interlanguage analysis in a search for adverb chains. Chaining multiple adverbs has been shown to be particularly infrequent in learner data (cf. Hirschmann 2009, Zeldes et al. 2008), possibly because of the variability of possible word orders. Query (7) finds such chains in middle-field initial position by left aligning the first of two adverbs with the middle-field:

```
(7) matrix-satz_felder="MF_MS" &
    pos="ADV" &
    pos="ADV" &
    #1 _l_ #2 &
    #2 . #3 &
    meta::l1=/[^(de)]/
```

Here the middle field (#1) is said to begin at the same point as the first ADV (the operator '_l_' for left alignment), and the first ADV precedes the second ADV (#2 . #3, direct precedence). By comparing frequencies from learners of different native languages and native German speakers, contrastive quantitative analyses can reveal if the structure searched for is under- or overused, and which native languages are involved, in the case of L1 interference. Russian natives, for example, show a middle-field initial overuse of adverb chains at a rate of 1.84 compared to native speakers, whereas Polish natives show an underuse at a rate of 0.76. A qualitative study of results suggests the possibility that the Russian writers tend to use the postverbal position more rarely for the subject, whereas this word order in the Polish and native texts forces subsequent adverbs further into the middle-field.

3.2 TIGER – a constituent syntactic treebank

In this use case we concentrate on syntactic queries, which can also be run in conjunction with span-based searches. TIGER (Brants et al. 2002) is a treebank consisting of approx. 900,000 tokens (50,000 sentences) of German newspaper text which has been semi-automatically tagged with parts of speech according to the STTS tagset (Schiller et al. 1999), morphological information, and syntactic structure. The corpus has served as an empirical basis for works in various fields, including studies on German syntax (Kempen & Harbusch 2005), training of parsers (Buchholz & Marsi 2006), and lexical semantics (Schulte im Walde 2006). The advantage of porting it to ANNIS is the possibility of adding additional annotation layers (e.g. coreference annotation, cf. Kunz & Hansen-Schirra 2003), as well as accessing heterogeneous resources in a uniform way. This section will focus on queries to syntactic trees. ANNIS provides the following operators:

- '>' (direct dominance, A>B means there is an edge from node A to node B, case (a) in Figure 5),
- '>*' (indirect dominance, A>*B means there is a path (i.e. sequence of edges) from A to B, case (b)),
- '>@l' (leftmost child, A>@l B means B is the leftmost child of A with respect to token order, case (c))
- '>@r' (rightmost child, analogous to leftmost child, case (d))
- '$' (common parent, A $ B means A and B have a common parent node, case (e))
- '$.*' (common ancestor, A $ B means A and B have a common ancestor node, case (f))

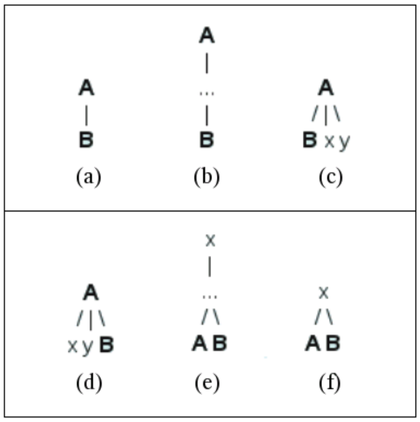


Fig. 5: Illustrations to dominance operators in ANNIS

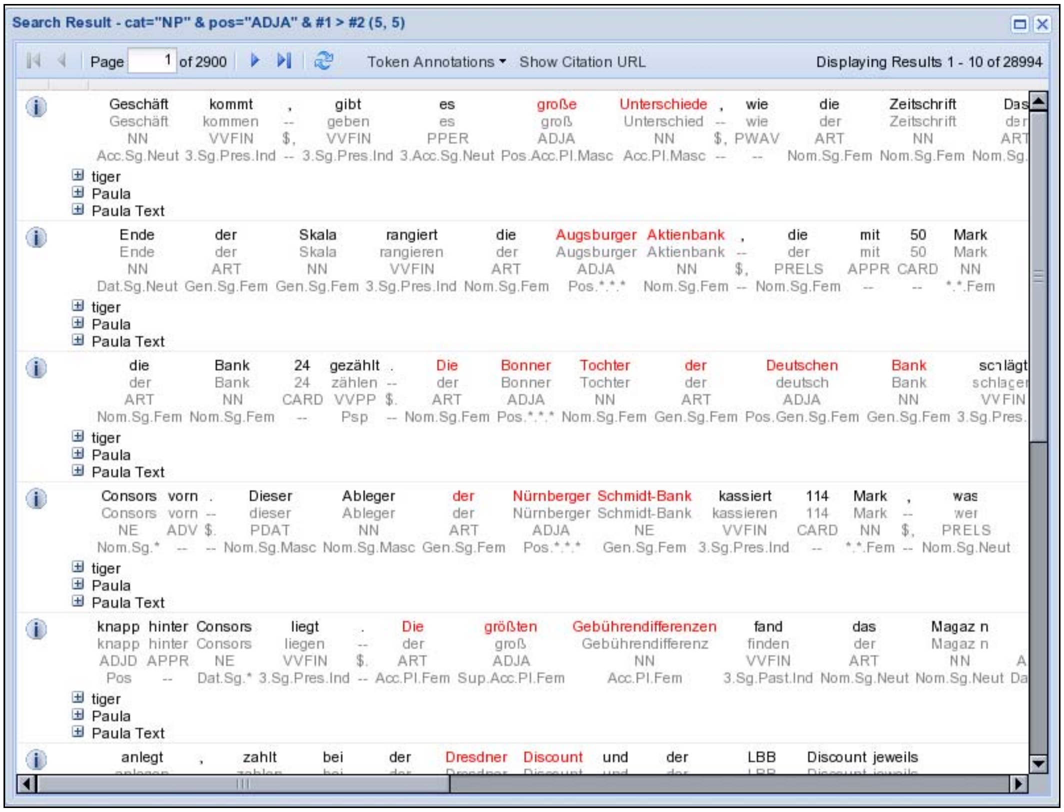
Extensions to operators include:

- '>n,m' (dominance over *n* to *m* generations, A >1,3 B means there is a sequence of 1 to 3 edges between A and B)
- '>edgetype' (edge type, e.g. A >secedge B means there is a secondary edge (a TIGER specific edge type, though arbitrary edge types may be specified in a corpus) between A and B. Leaving the edge type unspecified results in '>' matching any edge type (in the case of TIGER, 'edge' and 'secedge'). Similarly to >* it is possible to search for a path of a certain edge types with >edgetype*)
- '>[label]' (edge label, A >[feature="value"] B means there is an edge from A to B with the edge label feature="value". This may be combined with an edge type, e.g. A >secedge[func="MO"] for an edge of type secedge bearing the label func="MO")

Query (8) is an instance of a query for direct dominance. This query searches for a noun phrase (cat="NP", i.e. category: noun phrase) and an adjective (pos="ADJA", i.e. pos tag: adjective, attributive), where the noun phrase (node #1) directly dominates ('>' operator) the adjective (node #2).

(8) `cat="NP" & pos="ADJA" & #1 > #2`

The query has 28,994 hits, the first few of which are shown in Figure 6. The graph view in Figure 7 can then be opened by clicking on the '+ tiger' symbol just below each match.



Search Result - cat="NP" & pos="ADJA" & #1 > #2 (5, 5)

Page 1 of 2900 Token Annotations ▾ Show Citation URL Displaying Results 1 - 10 of 28994

Geschäft	kommt	,	gibt	es	große	Unterschiede	,	wie	die	Zeitschrift	Das
Geschäft	kommen	--	geben	es	groß	Unterschied	--	wie	der	Zeitschrift	der
NN	VVFIN	$,	VVFIN	PPER	ADJA	NN	$,	PWAV	ART	NN	ART
Acc.Sg.Neut	3.Sg.Pres.Ind	--	3.Sg.Pres.Ind	3.Acc.Sg.Neut	Pos.Acc.Pl.Masc	Acc.Pl.Masc	--	--	Nom.Sg.Fem	Nom.Sg.Fem	Nom.Sg.

⊞ tiger
⊞ Paula
⊞ Paula Text

Ende	der	Skala	rangiert	die	Augsburger	Aktienbank	,	die	mit	50	Mark
Ende	der	Skala	rangieren	der	Augsburger	Aktienbank	--	der	mit	50	Mark
NN	ART	NN	VVFIN	ART	ADJA	NN	$,	PRELS	APPR	CARD	NN
Dat.Sg.Neut	Gen.Sg.Fem	Gen.Sg.Fem	3.Sg.Pres.Ind	Nom.Sg.Fem	Pos.*.*.*	Nom.Sg.Fem	--	Nom.Sg.Fem	--	--	*.*.Fem

⊞ tiger
⊞ Paula
⊞ Paula Text

die	Bank	24	gezählt	.	Die	Bonner	Tochter	der	Deutschen	Bank	schlägt
der	Bank	24	zählen	--	der	Bonner	Tochter	der	deutsch	Bank	schlagen
ART	NN	CARD	VVPP	$.	ART	ADJA	NN	ART	ADJA	NN	VVFIN
Nom.Sg.Fem	Nom.Sg.Fem	--	Psp	--	Nom.Sg.Fem	Pos.*.*.*	Nom.Sg.Fem	Gen.Sg.Fem	Pos.Gen.Sg.Fem	Gen.Sg.Fem	3.Sg.Pres.

⊞ tiger
⊞ Paula
⊞ Paula Text

Consors	vorn	.	Dieser	Ableger	der	Nürnberger	Schmidt-Bank	kassiert	114	Mark	,	was
Consors	vorn	--	dieser	Ableger	der	Nürnberger	Schmidt-Bank	kassieren	114	Mark	--	wer
NE	ADV	$.	PDAT	NN	ART	ADJA	NE	VVFIN	CARD	NN	$,	PRELS
Nom.Sg.*	--	--	Nom.Sg.Masc	Nom.Sg.Masc	Gen.Sg.Fem	Pos.*.*.*	Gen.Sg.Fem	3.Sg.Pres.Ind	--	*.*.Fem	--	Nom.Sg.Neut

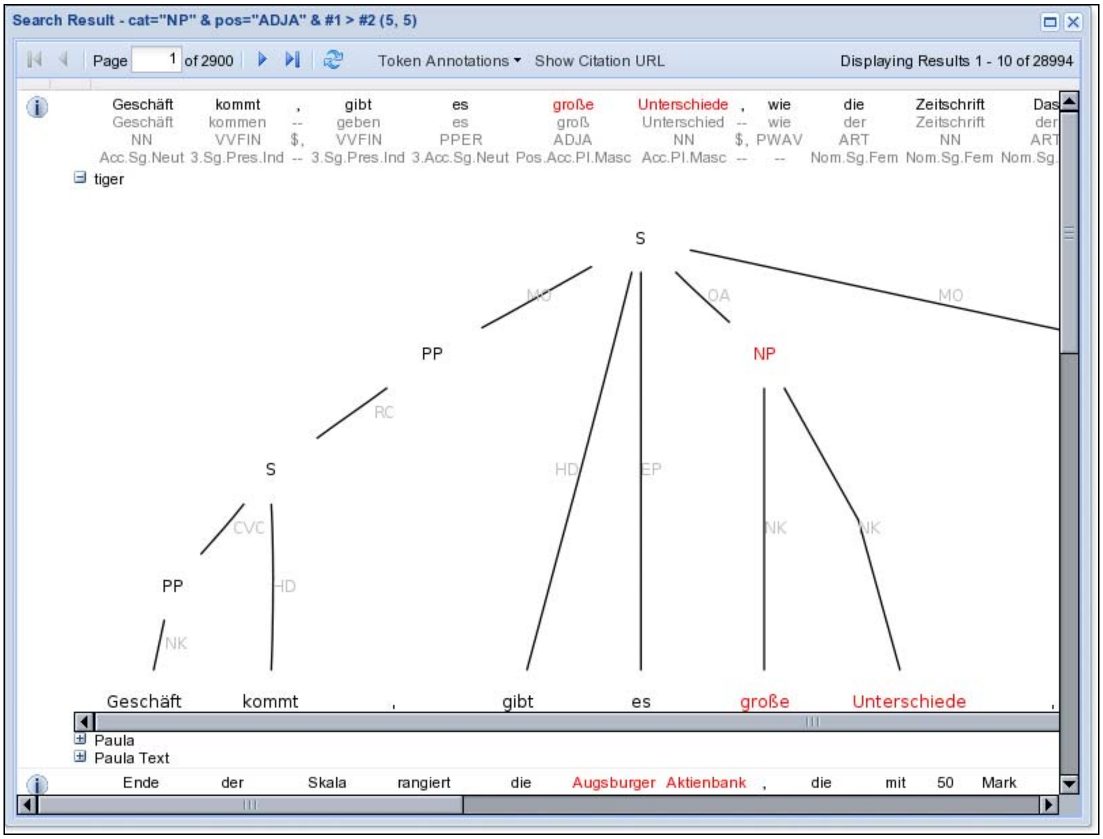
⊞ tiger
⊞ Paula
⊞ Paula Text

knapp	hinter	Consors	liegt	.	Die	größten	Gebührendifferenzen	fand	das	Magazin	
knapp	hinter	Consors	liegen	--	der	groß	Gebührendifferenz	finden	der	Magazin	
ADJD	APPR	NE	VVFIN	$.	ART	ADJA	NN	VVFIN	ART	NN	A
Pos	--	Dat.Sg.*	3.Sg.Pres.Ind	--	Acc.Pl.Fem	Sup.Acc.Pl.Fem	Acc.Pl.Fem	3.Sg.Past.Ind	Nom.Sg.Neut	Nom.Sg.Neut	Da

⊞ tiger
⊞ Paula
⊞ Paula Text

anlegt , zahlt bei der Dresdner Discount und der LBB Discount jeweils

Fig. 6: Result window for query (8), first page of matches in KWiC view (context settings: 5 tokens to either side.)



Search Result - cat="NP" & pos="ADJA" & #1 > #2 (5, 5)

Geschäft	kommt	,	gibt	es	große	Unterschiede	,	wie	die	Zeitschrift	Das
Geschäft	kommen	--	geben	es	groß	Unterschied	--	wie	der	Zeitschrift	der
NN	VVFIN	$,	VVFIN	PPER	ADJA	NN	$,	PWAV	ART	NN	ART
Acc.Sg.Neut	3.Sg.Pres.Ind	--	3.Sg.Pres.Ind	3.Acc.Sg.Neut	Pos.Acc.Pl.Masc	Acc.Pl.Masc	--	--	Nom.Sg.Fem	Nom.Sg.Fem	Nom.Sg.

Fig. 7: Graph view for example hit in query (8), the NP: *große Unterschiede*, 'big/ADJA differences/NN'.

The matching nodes and corresponding text portions are highlighted in red (see Figures 6 and 7). Items written in grey font (Figure 7) are edge labels. In the TIGER Corpus, the grammatical function of a node is represented as such a label, attached to the ingoing edge from its parent node. Consequently, to find sentences containing an accusative object, for instance, one would formulate a query like (9), stating that there is a node #1 labeled cat="S" (category 'S' for sentence) and a node #2 labeled cat="NP", and an edge from node #1 to node #2 labeled func="OA" (grammatical function: object, accusative). Subsequently, to find sentences where the accusative object precedes the predicate, we formulate query (10), adding to (9) the following constraints: there is a node #3 with a part of speech (pos) tag of the pattern V.FIN, a regular expression. With the period ('.') representing any character, this pattern will match any finite verb form, including auxiliaries (VAFIN), modals (VMFIN), and main verbs (VVFIN). Node #1 (the sentence node) dominates node #3 (the predicate), and node #2 (the object NP) directly precedes node #3 (the predicate).

```
(9)  cat="S" & cat="NP" & #1 >[func="OA"] #2
(10) cat="S" & cat="NP" & #1 >[func="OA"] #2 &
     pos=/V.FIN/ & #1 > #3 & #2 . #3
```

An example hit is shown in Figure 8. This particular match incidentally shows how the graph view can also handle crossing edges, in this case caused by an extraposed relative clause.

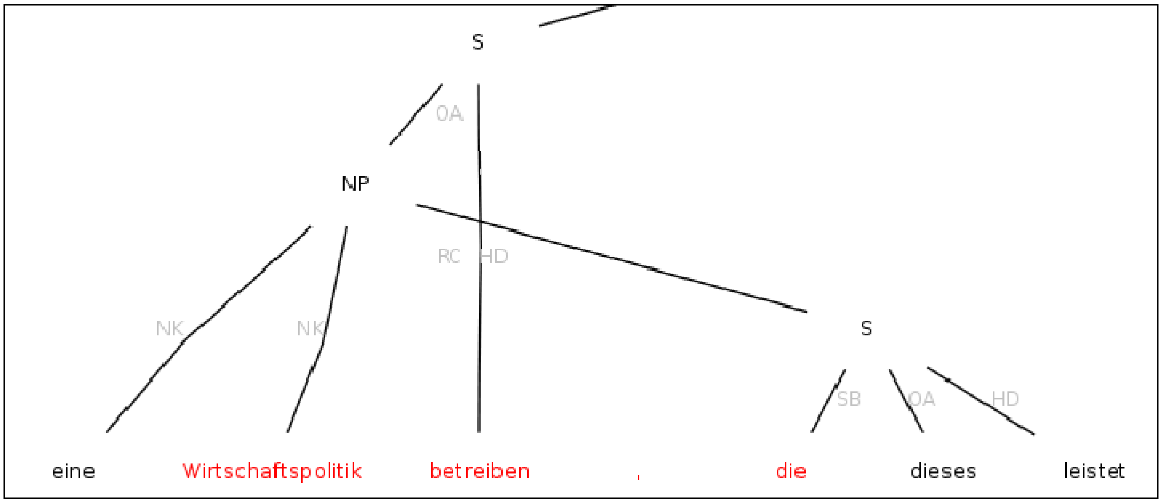


Fig. 8: Crossing edges in the graph view for a result of query (10): '...pursue a financial policy which accomplishes this'.

Finally it should be noted that queries can of course combine constraints for dominance and overlap, as introduced in Section 3.1. Query (11), for example, matches a preposition (pos="APPR") and common noun (pos="NN") where the preposition and noun have the same ancestor ('$' operator) and the noun is in the dative (morph=/.*Dat.*/).

(11) `pos="APPR" & pos="NN" & #1 $ #2 & morph=/.*Dat.*/ & #2 _=_ #3`

By using the operator '_=_', the area designated as a noun (pos="NN") is said to be identical with the area bearing the annotation for dative case.

3.3 ONTONOTES – coreference, named entities, and PTB syntax

In this section, we will show how constraints of both the span and dominance graph types introduced so far can be combined with further types of edges, in this case used to express coreference. We will also introduce the discourse visualization, which allows users to review query matches within a larger context while inspecting discourse referents. For demonstration purposes, we use the English portion of ONTONOTES 1.0 (Hovy et al. 2006), a corpus of American English newswire (Wall Street Journal, nearly 400,000 tokens). It is annotated, among other layers, for part of speech and syntax according to the PennTreebank scheme (Marcus et al. 1993), coreference, and named entities. To avoid having to query for individual coreferent IDs or identical coreferent ID values, we interpreted the coreference annotation as directed anaphoric relations, linking each entity to the nearest entity of the same coreferent class in its left context.

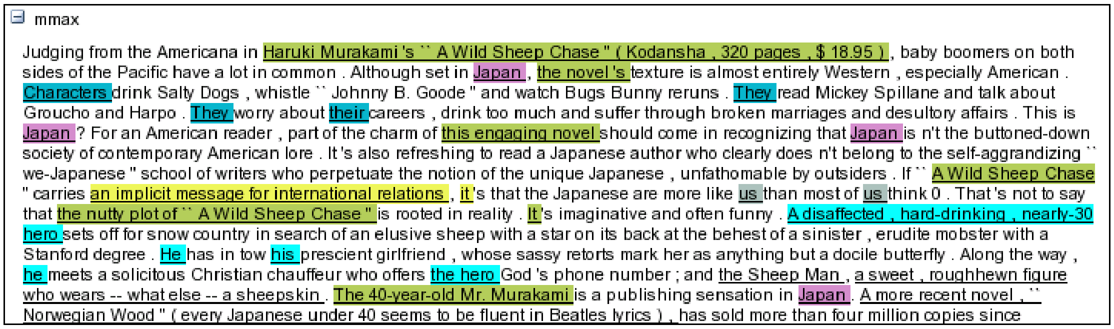
For anaphoric relations, ANNIS provides the following operators:

- '->' (directly linked, A->B means there is a pointing relation from node A to node B),
- '->relation type' (type of pointing relation, e.g. A->IDENT_relation B specifies that a pointing relation of the type IDENT_relation leads from A to B)
- '->relation_type*' (indirectly linked, A -> IDENT_relation* B means there is a path (i.e. sequence of pointing relations) of the type IDENT_relation from A to B)

Query (12), for example, matches a pair of nodes linked by an anaphoric relation, i.e. an entity and its direct antecedent. Query (13), by contrast, matches an entity and any antecedent in the anaphoric chain.

```
(12) node & node & #1 ->IDENT_relation #2
(13) node & node & #1 ->IDENT_relation* #2
```

Matching nodes can be viewed in the KWiC concordance for each hit, but a more convenient interpretation of coreferent entities is made possible by the discourse view portrayed in Figure 9.



Judging from the Americana in Haruki Murakami 's `` A Wild Sheep Chase " ( Kodansha , 320 pages , $ 18.95 ) , baby boomers on both sides of the Pacific have a lot in common . Although set in Japan , the novel 's texture is almost entirely Western , especially American . Characters drink Salty Dogs , whistle `` Johnny B. Goode " and watch Bugs Bunny reruns . They read Mickey Spillane and talk about Groucho and Harpo . They worry about their careers , drink too much and suffer through broken marriages and desultory affairs . This is Japan ? For an American reader , part of the charm of this engaging novel should come in recognizing that Japan is n't the buttoned-down society of contemporary American lore . It 's also refreshing to read a Japanese author who clearly does n't belong to the self-aggrandizing `` we-Japanese " school of writers who perpetuate the notion of the unique Japanese , unfathomable by outsiders . If `` A Wild Sheep Chase " carries an implicit message for international relations , it 's that the Japanese are more like us than most of us think 0 . That 's not to say that the nutty plot of `` A Wild Sheep Chase " is rooted in reality . It 's imaginative and often funny . A disaffected , hard-drinking , nearly-30 hero sets off for snow country in search of an elusive sheep with a star on its back at the behest of a sinister , erudite mobster with a Stanford degree . He has in tow his prescient girlfriend , whose sassy retorts mark her as anything but a docile butterfly . Along the way , he meets a solicitous Christian chauffeur who offers the hero God 's phone number ; and the Sheep Man , a sweet , roughhewn figure who wears -- what else -- a sheepskin . The 40-year-old Mr. Murakami is a publishing sensation in Japan . A more recent novel , `` Norwegian Wood " (every Japanese under 40 seems to be fluent in Beatles lyrics) , has sold more than four million copies since

Fig. 9: Coreference visualization in ANNIS. Entities with a coreferent are underlined, entities of same coreference class are co-highlighted upon mouse click.

The discourse visualization allows users to view an entire text to examine the larger context of query matches. Each sequence of coreferent expressions can be controlled individually by clicking on one of its members, which then highlights the entire coreference chain in the same color.

Finally, it is also possible to combine constraints on pointing relations and spans or dominance edges. For example, one can query anaphoric relations in conjunction with syntax or named entity information. Query (14) shows a search for a pronoun and its antecedent, where the pronoun is dominated indirectly by a prepositional phrase (cat="PP"). Query (15) uses the spans annotating named entities to search for a person's name and its antecedent:

```
(14) node & node & #1 ->IDENT_relation #2 & pos="PRP" & #2
     _=_ #3 & cat="PP" & #4 >* #3
(15) node & node & #1 ->IDENT_relation #2 & TYPE=”PERSON” &
     #2 _=_ #3
```

4. ANNIS in the context of related tools and frameworks

4.1 PAULA standoff XML – integrating annotation source formats

As the use cases in Section 3 have shown, ANNIS2 can be used to search and visualize very heterogeneous data, originally conceived for different types of corpora and prepared using very different annotation tools. While automatic part of speech tagging and, increasingly, syntactic parsing can achieve usable results, a variety of specialized manual annotation layers are often essential for the key interests of linguistic researchers. It is therefore imperative to incorporate automatically generated, readily available data with multiple expert annotations referring to the same text. A technical requirement in order to do this is the establishment of a generic format capable of representing annotation layers independently while maintaining a uniform reference to the raw data.

The PAULA format (Dipper, 2005; Dipper & Götze, 2005) is the generic XML format used for this purpose in ANNIS2 and other NLP pipelines (Stede et al., 2006). It uses standoff XML representations and is conceptually closely related to the formats NITE XML (Carletta et al., 2003) and GraF (Ide & Suderman, 2007). PAULA was specifically designed to support the lossless representation of different types of text-oriented annotations (layer-based/timeline annotations, hierarchical annotations, pointing relations), optimized for the annotation of multiple layers, including conflicting hierarchies and simple addition/deletion routines for annotation layers. Primary data (i.e. the running text of corpus documents) is stored in a separate file, as is each of the annotation layers (e.g. part of speech, syntactic structures etc.), providing an encapsulated data representation and avoiding any interference between concurrent annotations.

Annotation layers may originate in various tools. Currently data may be converted into PAULA XML from EXMARaLDA (Schmidt 2004), TigerXML (Brants & Plaehn 2000), MMAX2 (Müller & Strube 2006), RSTTool (O'Donnell 2000), PALinkA (Orasan 2003) and Toolbox (Stuart et al. 2007). Data from ELAN (http://www.lat-mpi.eu/tools/elan/) is supported indirectly, over an intermediate conversion to EXMARaLDA, and a converter for generic XML also exists, which converts XML documents into simple hierarchical tree structures. The data from the different annotation tools is then merged to refer to the primary data by means of XLinks and XPointers.

The PAULA object model (POM) , which underlies the representation of linguistic annotations in the PAULA XML format, is formalized as a labeled graph. POM uses two kinds of elements to represent linguistic structures: nodes and edges. Nodes may be of one of three types:

- **token:** character span in the primary data that forms the basis for higher-level annotations.
- **markable:** (span of) token(s) that can be annotated with linguistic information. Markables represent flat, layer-based annotations defined with respect to the sequence of tokens.
- **struct:** hierarchical structure. A dominance relation may be established between a struct (e.g. a phrase node), and tokens, markables, or other struct nodes. Structs are used to compose trees, as in syntax or discourse structure theory.

Edges are the relational units of annotation, connecting tokens, markables and structs. They may be of one of two types:

- **dominance relation:** a directed edge between a struct and its children
- **pointing relation:** a directed edge between nodes in general (tokens, markables, structs)

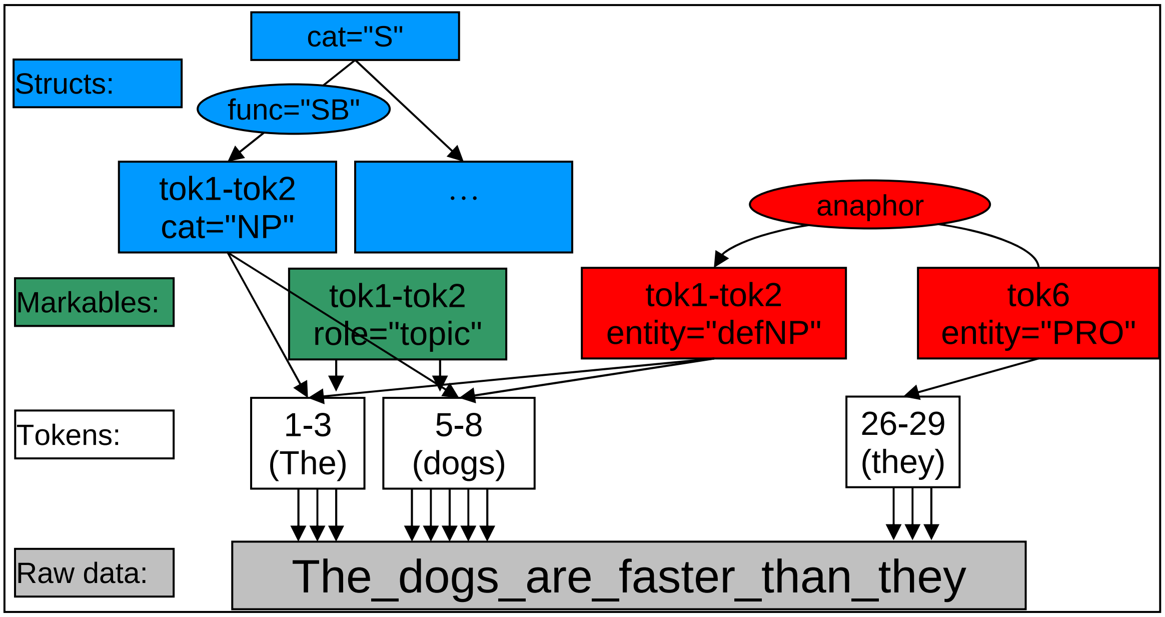


Fig. 10: Simplified PAULA representation for an annotated fragment *The dogs are faster than they*.

Figure 10 shows how these elements are used together to form an annotation graph: the token level refers to the raw data beneath it in terms of character positions (e.g. 'dogs' occupies characters 5-8). Two Different markables refer to the tokens 'the dogs' and one markable refers to the token 'they'. At the top, a struct joins 'the' and 'dogs' together, and is in turn dominated by a further struct (as marked by the directed edges between structs). Finally, the two markables on the right are joined by a pointing relation of the type 'anaphor'.

Linguistic annotations referring to these structural elements are represented by labels that can be attached to nodes and edges, e.g. role="topic" or func="SB" in Figure 10. Labels consist of key-value pairs, which are both simple strings. Optionally they may also specify a namespace, which serves to group annotations together, e.g., 'penn' in penn:pos="RB" (see also Section 2.2).

Finally, all annotation layers are bound together with the primary data file in an **annoset**. By declaring which annotation files refer to a particular stream of primary data, the annoset forms a (sub-)graph consisting of nodes and edges, which together with their labels form the complete XML representation of the linguistic annotations. Annosets can also be assigned labels like nodes and edges, which then function as metadata for the annoset, i.e. data about the annotations or the text, such as the name of the annotator, the author of the text, etc. More information on PAULA XML can be found at http://www.sfb632.uni-potsdam.de/~d1/paula/doc/. A JAVA API that allows to read, write and manipulate PAULA annosets is currently under development.

4.2 ANNIS2 and other tools

Like other tools aiming to process multi-layer corpora, ANNIS2 has been developed with the goal of providing linguists with unified access to heterogeneous linguistic annotations created using different, specialized tools for manual or automated annotation. As shown elsewhere (Dipper 2005, Dipper & Götze 2005, Chiarcos et al. 2008), the PAULA XML format described above and used as a pivot format for ANNIS2 is generic enough to represent a very wide spectrum of flat/layer-based, hierarchical and pointing-relation-based linguistic annotations. In this respect it is comparable to other standoff XML formats used for multi-layer corpora, such as GrAF (Ide & Suderman 2007), NITE XML (Carletta *et al.*, 2003), and a number of other generic formats for linguistic annotation, which are all based on a graph-theoretic object model. However by contrast to both of the above formats, the PAULA Object Model allows a semantic differentiation between dominance relations and pointing relations, and also a semantic differentiation between markables and structs. These distinctions, as expressed in the format, translate into default visualization strategies in ANNIS2, and correspond to the semantics of AQL query operators (i.e. the distinction between dominance semantics, as in the examples in Section 2.2, and span coverage or overlap in Section 2.1).

The PAULA-ANNIS approach to merging corpora from multiple sources is also comparable to such multi-tool NLP pipelines as UIMA (Ferrucci & Lally 2004), general-purpose annotation frameworks such as GATE (Cunningham, 2002), and libraries that support annotations on the basis of a generic data format (e.g., LAF, Ide & Romary 2004; NXT, Carletta *et al.* 2005). The main outstanding feature of ANNIS2 as comparaed to these frameworks is the focus on the development of an extensible, versatile and performant database implementation of the graph model, capable of expressing different structures in a modular way using multiple visulizations, such as the grid, tree and discourse views presented in Section 3. Based on the structure of the data, a default visualization is determined for every namespace. If only non-hierarchical information is to be visualized, the grid view is chosen, if dominance relations are to be visualized, the tree view, and if pointing relations are to be visualized, the discourse view is currently chosen. In the future it may be possible to extend the visualizations to more special cases with specialized namespaces, depending on the semantics of the relations being visualized (e.g. dominance structures expressing rhetorical relations between sentences as opposed to constituent or dependency structures between phrases and terminals, see also Section 5). To our knowledge, this modularity of possible visualizations is a unique characteristic of ANNIS2, allowing a wide spectrum of applications – in other applications, multi-layer annotations are usually merged into one single visualization, e.g., a grid view as in EXMARaLDA and ELAN, a tree (or graph) view as in TIGERSearch (Lezius 2002), GrAF (Ide & Suderman 2007) and TrED (Pajas & Štěpánek 2008), a discourse view as in Serengeti (Diewald et al. 2008), or a costumizable view as in MMAX. In ANNIS2, different views complement each other and can be consulted in parallel, unlike different means of visualizing XML annotations described in Rehm et al. (2008).

As for querying capacity, AQL shows two main advantages. Firstly, the support for highly specific descriptions of the relationship between elements (typed labeled edges) and their position with regard to the tokenized data (coverage operators) provides an intuitive mechanism for combined constraints on hierarchical and positional annotation. Secondly, the implementation based on a relational database platform ensures high performance as compared to XML database solutions using XQuery (e.g. AnnoLab, Chiarcos et al. 2007, Rehm et al. 2008; Splicr, Rehm et al. 2008). A further benefit of searches using ANNIS2 is the lack of boundaries between sentence graphs: unlike TIGERSearch and PML Query (Pajas & Štěpánek 2009), the scope of AQL operators extends beyond sentence boundaries, allowing searches over longer stretches of text within a document.

A weakness of AQL as compared to PML is the current lack of support for negation of operators and annotations. Currently negation is only achievable using negation within RegEx values, though an expansion of negation functionality is planned. Another feature currently lacking in AQL is quantification (existence, universality), which is supported e.g. in NXT. These are therefore foci for the development of the ANNIS2 backend.

5. Summary and outlook

In this paper we have presented the approach to the search and visualization of mutli-layer corpora taken in ANNIS2. By adopting a multi-layer architecture based on underlying stand-off XML corpora in the PAULA format, ANNIS2 supports extensible corpora with collaborative specialized annotations, allowing researchers to explore richer, more deeply annotated corpora than previously possible. The AQL query mechanism, based on graphs of nodes with labeled edges, permits the formulation of complex expressions corresponding to linguistic structures at all levels of representation, not only in morphosyntax, semantics and pragmatics, but also error annotation, phonology, text structure, paleography or any other form of annotation a researcher may be interested in.

In future work on the data modeling side, we currently plan to extend ANNIS2 in two directions, by offering support for parallel corpora and subtoken units. Parallel corpora present a challenge to the PAULA/ANNIS2 model, since they contain documents contain multiple primary texts. The document concept must therefore be adjusted to allow for this option. Once this is achieved, the existing model would offer an optimal mechanism for representing parallel corpora in the use of labeled edges between aligned elements. Aligned texts, chapters, paragraphs, sentences or tokens can be annotated as markables with alignment edges between them. This would even allow for conflicting alignment, as in sentences being aligned despite the paragraphs containing them not being aligned, etc. (for edge-based models accommodating conflicting alignments, see Romary and Bonhomme 2000).

Subtokenization offers a second challenge, which involves representing and annotating units smaller than the word form, such as morphemes, phonemes, syllables or graphemes, the latter also including extra-linguistic symbols e.g. in corpora representing ancient manuscripts. Simply defining these as the tokens in a corpus may reduce usability, as adjacency of consecutive words may be interrupted by such sub-tokens, the distance between two matches in words (the operator '.*n*,*m*'), and the definition of context (e.g. 5 words to the left or right of a match) may also behave unpredictably. We therefore aim to separate the smallest atomic units of annotation (tokens proper) from the granular unit of reference for purposes of context and search (adjacency or distance in words), which is usually the word form. Corresponding adjustments to the query language may then also become necessary. Other desirable extensions to the query language in the areas of negation and quantification have already been mentioned in Section 4.2.

On the visualization side, we look to improve our current discourse view by visualizing different types of relations appropriately, and also to extend our support for new specialized visualizations, such as rhetorical annotations, or generalized views for pointing relations. Adjustments to old visualizations will also be necessary to support parallel corpora and subtokens. We hope that open-source contributions to the ANNIS architecture will make the system as flexible and diverse as the data for which it was conceived.

References

All URLs were checked on 29 September 2009

Bański, P. and A. Przepiórkowski (2009). Stand-off TEI Annotation: the Case of the National Corpus of Polish. In *Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP 2009*. Suntec, Singapore, 64-67.

Bernsen, N.O., L. Dybkjær and M. Kolodnytsky (2002). The NITE Workbench - A Tool for Annotation of Natural Interactivity and Multimodal Data. In *Proceedings of the Third International Conference on Language Resources and Evaluation LREC 2002)*. Las Palmas, Spain.

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing* 8(4), 243-257.

Bies, A., M. Ferguson, K. Katz and R. Mac-Intyre (1995). *Bracketing Guidelines for Treebank II Style Penn Treebank Project*. Technical report, University of Pennsylvania.

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

Brants, T. and O. Plaehn (2000). Interactive Corpus Annotation. In *Proc. of LREC 2000*. Athens, Greece.

Buchholz, S. and E. Marsi (2006). CoNLL-X shared task on Multilingual Dependency Parsing. In *Proceedings of the Tenth Conf. on Computational Natural Language Learning (CoNLL)*. New York City, USA.

Carletta, J., S. Evert, U. Heid, J. Kilgour (2005). The NITE XML Toolkit: data model and query. *Language Resources and Evaluation Journal* 39(4), 313-334.

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

Chiarcos, C., I. Fiedler, M. Grubic, A. Haida, K. Hartmann, J. Ritz, A. Schwarz, A. Zeldes and M. Zimmermann (2009). Information Structure in African Languages: Corpora and Tools. In *Proceedings of the Workshop on Language Technologies for African Languages (AFLAT), 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09)*. Athens, Greece, 17-24.

Chiarcos, C., R. Eckart, G. Rehm (2007). An OWL- and XQuery-Based Mechanism for the Retrieval of Linguistic Patterns from XML-Corpora. In *Proceedings of the International Conference Recent Advances in Natural Language Processing - RANLP 2007*. Borovets, Bulgaria.

Chiarcos, C., S. Dipper, M. Götze, U. Leser, A. Lüdeling, J. Ritz and M. Stede (2008). A Flexible Framework for Integrating Annotations from Different Tools and Tag Sets. *Traitement Automatique des Langues* 49(2), 271-293.

Christ., O (1994). A modular and flexible architecture for an integrated corpus query system. In: *Proceedings of Complex 94*. Budapest, 23-32.

Cunningham, H. (2002). GATE. A General Architecture for Text Engineering. *Computers and the Humanities* 36, 223–254.

Diewald, N., M. Stührenberg, A. Garbar and D. Goecke (2008). Serengeti - Webbasierte Annotation semantischer Relationen. *Journal for Language Technology and Computational Linguistics* 24(2), 74-93.

Dipper S., L. Faulstich, U. Leser and A. Lüdeling (2004). Challenges in Modelling a Richly Annotated Diachronic Corpus of German. In *Proc. of the Workshop on XML-based Richly Annotated Corpora*. Lisbon, Portugal, 21-29.

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

Dipper, S. and M. Götze (2005). Accessing Heterogeneous Linguistic Data - Generic XML-Based Representation and Flexible Visualization. In *Proc. of the 2nd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*. Poznań, Poland, 206-210.

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

Dipper, S., T. Brants, W. Lezius, O. Plaehn, and G. Smith. (2001). The TIGER Treebank. In: *Third Workshop on Linguistically Interpreted Corpora LINC-2001*. Leuven, Belgium.

Doolittle, S. (2008). *Entwicklung und Evaluierung eines auf dem Stellungsfeldermodell basierenden syntaktischen Annotationsverfahrens für Lernerkorpora innerhalb einer Mehrebenen-Architektur mit Schwerpunkt auf schriftlichen Texten fortgeschrittener Deutschlerner*. Magisterarbeit, Humboldt-University, Berlin.

Dürscheid, C. (2007). *Syntax. Grundlagen und Theorien*. Göttingen: Vandenhoeck & Ruprecht.

Faulstich, L., U. Leser, A. Lüdeling (2005). *Storing and Querying Historical Texts in a Database*. Technical Report n° 176, Institut für Informatik, Humboldt-Universität zu Berlin.

Ferrucci, D. and A. Lally (2004). UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering* 10(3-4), 327-348.

Fitschen, A. and P. Gupta (2008). Lemmatising and morphological tagging. In A. Lüdeling and M. Kytö, (eds.), *Corpus Linguistics. An International Handbook, vol. 1*. Berlin: Mouton de Gruyter 552-564.

Granger, S. (1993). The International Corpus of Learner English. In: J. Aarts, P. de Haan and N. Oostdijk (eds.), *English Language Corpora: Design, Analysis and Exploitation*. Amsterdam: Rodopi, 57-69.

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

Grust, T., M. V. Keulen, J. Teubner (2004). Accelerating XPath Evaluation in any RDBMS. *ACM Transactions on Database Systems* 29 (1), 91-131.

Hirschmann, H. (2009). Von der Restkategorie Adverb zur korpusrelevanten syntaktischen Ausdifferenzierung. *Grammar and Corpora 3*. Mannheim, Germany, September 24.

Hovy, E., M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel (2006). OntoNotes: The 90% Solution. In *Proceedings of the Human Language Technology Conference of the NAACL*, Companion Volume: Short Papers , Association for Computational Linguistics , New York City, USA, 57-60.

Ide, N. and L. Romary (2004). International Standard for a Linguistic Annotation Framework. *Journal of Natural Language Engineering* 10(3-4), 211-225.

Ide, N. and K. Suderman (2007). GrAF: A Graph-based Format for Linguistic Annotations. In *Proceedings of the Linguistic Annotation Workshop*, held in conjunction with ACL 2007. Prague, 1-8.

Kempen, G. and K. Harbusch (2005). The relationship between grammaticality ratings and corpus frequencies: A case study into word order variability in the midfield of German clauses. In: Kepser, Stephan and Reis, Marga (eds.), *Linguistic Evidence – Empirical, Theoretical, and Computational Perspectives*. Berlin: Mouton De Gruyter.

Krifka, M. (2007). Basic notions of information structure. In C. Fery and M. Krifka (eds.), *Interdisciplinary Studies of Information Structure* 6. Potsdam: Universität Potsdam.

Kruijff-Korbayová, I. and G.-J. M. Kruijff (2004). Discourse-level annotation for investigating information structure. In *Proceedings of the ACL Workshop on Discourse Annotation*. Barcelona, Spain, 41-48.

Kunz, K. and S. Hansen-Schirra (2003). Coreference annotation of the TIGER treebank. In *Proceedings of the Workshop Treebanks and Linguistic Theories 2003*, 221-224, Vaxjo.

Leech G. (1997). Introducing Corpus Annotation. In: R. Garside, G.. Leech and T. McEnery (eds.), *Corpus Annotation*. London, New York: Longman, 1-18.

Lezius, W. (2002). *Ein Suchwerkzeug für syntaktisch annotierte Textkorpora*. Ph.D. thesis, Stuttgart University.

Lüdeling, A. (2008). Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In: Grommes, P./Walter, M. (eds.), *Fortgeschrittene Lernervarietäten*. Niemeyer, Tübingen, 119-140.

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

Marcus, M., M. Marcinkiewicz, and B. Santorini (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics 19*.

Müller, C. and M. Strube (2006). Multi-Level Annotation of Linguistic Data with MMAX2. In S. Braun, K. Kohn and J. Mukherjee (eds.), *Corpus Technology and Language Pedagogy.* Frankfurt: Peter Lang, 197–214.

O’Donnell, M. (2000). RSTTool 2.4 - A Markup Tool for Rhetorical Structure Theory. In *Proc. International Natural Language Generation Conference (INLG'2000), 13-16 June 2000*. Mitzpe Ramon, Israel, 253–256.

Orasan, C. (2003). Palinka: A Highly Customisable Tool for Discourse Annotation. In *Proc. of the 4th SIGdial Workshop on Discourse and Dialogue*. Sapporo, Japan.

Pajas P. and J. Štěpánek (2008). Recent Advances in a Feature-Rich Framework for Treebank Annotation. In *22nd International Conference on Computational Linguistics - Proceedings of the Conference*. Manchester, 673-680.

Pajas P. and J. Štěpánek (2009). System for Querying Syntactically Annotated Corpora. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*. Suntec, Singapore, 33-36.

Petrova, S., M. Solf, J. Ritz, C. Chiarcos and A. Zeldes (to appear). Building and Using a Richly Annotated Interlinear Diachronic Corpus: The Case of Old High German Tatian. *Traitement automatique des langues*.

Rehm, G., O. Schonefeld, A. Witt, C. Chiarcos and T. Lehmberg (2008). A Web-Platform for Preserving, Exploring, Visualising and Querying Linguistic Corpora and other Resources. *Procesamiento del Lenguaje Natural* 41, 155-162.

Rehm, G., R. Eckart, C. Chiarcos and J. Dellert (2008). Ontology-Based XQuery'ing of XML-Encoded Language Resources on Multiple Annotation Layers. In: *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008.

Romary, L. and P. Bonhomme (2000). Parallel alignment of structured documents. In J. Véronis (ed.), Parallel Text Processing. Kluwer Academic Publishers, 201-217.

Schiller, A., S. Teufel, C. Stöckert, and C. Thielen (1999). *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset)*. Technical report, University of Stuttgart and University of Tübingen. http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-1999.pdf

Schmid, H. (2008). Tokenizing and part-of-speech tagging. In A. Lüdeling and M. Kytö, (eds.), *Corpus Linguistics. An International Handbook, vol. 1*. Berlin: Mouton de Gruyter 527-551.

Schmidt, T. (2004). Transcribing and Annotating Spoken Language with Exmaralda. In *Proc. of the LREC-Workshop on XML Based Richly Annotated Corpora, Lisbon 2004*. Paris: ELRA.

Schulte im Walde, S. (2006). Experiments on the Automatic Induction of German Semantic Verb Classes. *Computational Linguistics* 32(2), 159-194.

Stede, M., H. Bieler, S. Dipper and A. Suriyawongkul (2006). SUMMaR: Combining Linguistics and Statistics for Text Summarization. In *Proceedings of the 17th European Conference on Artificial Intelligence (ECAI-06)*. Riva del Garda, Italy, 827-828.

Stuart, R., G. Aumann and S. Bird. (2007). Managing Fieldwork Data with Toolbox and the Natural Language Toolkit. *Language Documentation & Conservation* 1(1), 44–57.

Vieira, R., C. Gasperin and R. Goulart (2003). From Manual to Automatic Annotation of Coreference. In *International Symposium on Reference Resolution and Its Applications to Question Answering and Sumarization*, Venice.

Witten, I. H. and E. Frank. (2005). *Data mining: Practical Machine Learning Tools and Techniques, 2nd Ed*. San Francisco: Morgan Kaufman.

Wittenburg, P. (2008). Preprocessing multimodal corpora. In A. Lüdeling and M. Kytö, (eds.), *Corpus Linguistics. An International Handbook, vol. 1*. Berlin: Mouton de Gruyter 899-919.

Wynne, M. (2008). Searching and Concordancing. In A. Lüdeling and M. Kytö, (eds.), *Corpus Linguistics. An International Handbook, vol. 1*. Berlin: Mouton de Gruyter 706-737.

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