A Tool/Database Interface for Multi-Level Analyses

Kurt Eberle[1], Kerstin Eckart[1], Ulrich Heid[1,2], Boris Haselbach[1]

[1]Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart,
Azenbergstraße 12, D-70174 Stuttgart,

[2]Institut für Informationswissenschaft und Sprachtechnologie, Universität Hildesheim
Lübecker Straße 3, D-31141 Hildesheim

{eberle,eckartkn,heid,haselbbs}@ims.uni-stuttgart.de, ulrich.heid@uni-hildesheim.de

Abstract

Depending on the nature of a linguistic theory, empirical investigations of its soundness may focus on corpus studies related to lexical, syntactic, semantic or other phenomena. Especially work in research networks usually comprises analyses of different levels of description, where each one must be as reliable as possible when the same sentences and texts are investigated under very different perspectives. This paper describes an infrastructure that interfaces an analysis tool for multi-level annotation with a generic relational database. It supports three dimensions of analysis-handling and thereby builds an integrated environment for quality assurance in corpus based linguistic analysis: a vertical dimension relating analysis components in a pipeline, a horizontal dimension taking alternative results of the same analysis level into account and a temporal dimension to follow up cases where analyses for the same input have been produced with different versions of a tool. As an example we give a detailed description of a typical workflow for the vertical dimension.

Keywords: multi-level analysis, corpus study, infrastructure

1. Introduction

For empirical investigations of linguistic theories, corpus studies related to lexical, syntactic, and/or semantic phenomena are conducted. Especially in research networks, such as the collaborative research centre (SFB)732[1] many different requirements emerge. Such work therefore often comprises analyses of different levels of description, where each one must be as reliable as possible. Depending on the case, the needed depth of the linguistic analysis varies, as does the needed context size which is understood as the amount of text taken into account.

In this paper we present an infrastructure that interfaces an analysis tool for multi-level annotation with a generic relational database supporting three dimensions of analysis-handling. Section 2. is devoted to the principles of our approach and Sections 3. and 4. to the database and the analysis tool, respectively. In Section 5. we give a detailed example of a typical workflow.

2. Handling analyses: three dimensions

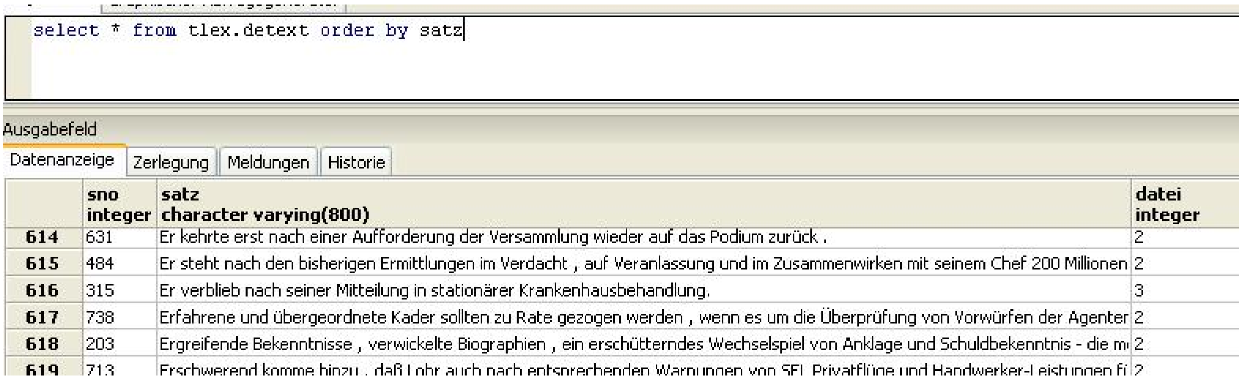
Most corpus studies require at least a morphological analysis, for example, with annotation of part-of-speech tags as produced, e.g. by *TreeTagger* (Schmid, 1994). Others require linguistic analyses of a 'higher' level, such as constituent trees or semantic representations. If a 'high-level' analysis depends on the results of lower levels, it typically can be computed more efficiently from these results, than from the underlying input sentence. Such a pipeline architecture of analysis processing is advantageous if corpora are investigated from various linguistic viewpoints with shared interest in the 'lower' levels of analysis and in corresponding reusability of the analyses of the sentences or texts. Prerequisites of this setting are that the analysis tool supports the pipeline architecture and that the analyses are stored and administrated for later reuse. We call the relations between such analyses of different depth *vertical relations*. Of course, the assignment of some 'higher' vertical relation is not necessarily to be carried out automatically. Thus, in SFB732 annotation of information status[2] labels as in (Riester et al., 2010) is carried out manually on the basis of constituent trees of the sentences considered (Eckart et al., 2012).

If different tools are available that can produce analyses of a particular level, it can be helpful to take all of the corresponding results into account in order to facilitate quality assurance of the annotations. We call such relations between analyses of the same level *horizontal relations*. This includes format conversions for compatibility, e.g. into an interchange format like GrAF (Ide and Suderman, 2007)), and inspections on the analysis results, e.g. counting occurrences of a specific annotated configuration. The analyses have to be identifiable with respect to their horizontal status which includes their respective annotation levels as well as their representation format. Therefore a type system is applied.

The third dimension refers to *temporal relations*. As analysis tools evolve over time, analyses produced for the same input but with different versions of a tool offer valuable clues to system improvement or decline. For example in cases where the knowledge base of a tool is enhanced, a comparison to earlier versions of the same analysis may give detailed information about effects and probably also side-effects of the changes. The prerequisites to exploit this information include the identification of tools and analyses with respect to their versions. On top of that the analyses have to be relatable to the tools or annotators producing them.

[1] `http://www.uni-stuttgart.de/linguistik/sfb732/`

[2] Information status (Prince, 1992) describes the givenness/novelty of referring expressions, classifying them according to whether they are anaphoric, inferable, deictic or discourse-new.



```
select * from tlex.detext order by satz
```

	sno integer	satz character varying(800)	datei integer
614	631	Er kehrte erst nach einer Aufforderung der Versammlung wieder auf das Podium zurück .	2
615	484	Er steht nach den bisherigen Ermittlungen im Verdacht , auf Veranlassung und im Zusammenwirken mit seinem Chef 200 Millionen	2
616	315	Er verblieb nach seiner Mitteilung in stationärer Krankenhausbehandlung.	3
617	738	Erfahrene und übergeordnete Kader sollten zu Rate gezogen werden , wenn es um die Überprüfung von Vorwürfen der Agenter	2
618	203	Ergreifende Bekenntnisse , verwickelte Biographien , ein erschütterndes Wechselspiel von Anklage und Schuldbekenntnis - die m	2
619	713	Erschwerend komme hinzu , daß Lohr auch nach entsprechenden Warnungen von SEL Privatflüge und Handwerker-Leistungen fi	2

Figure 1: Single sentences in the database with a sentence number (sno) and a filenumber (datei).

3. A generic relational database

The B3-database (B3DB) was created to handle different types of data that accumulate during a corpus-based project, such as (textual) primary data, information about tools and annotations as well as different annotations layers. Therefore it makes use of generic data structures, which are described in Eckart et al. (2010). The database supports the management of versioning information of tools and analyses, even if they evolve over time. To achieve this, the respective data has labels for start and end of validity; the object representation of the database is typed with labels for identification of e.g. annotation level and representation. On top of that, explicitly included typed relations indicate the processing pipeline. The B3DB is implemented as a PostgreSQL[3] database and queries can be conducted via SQL. As a result of the generic data structures, the SQL queries have to state in detail which data to select.

4. A multi-level processing tool

The B3-analysis-tool (Eberle et al., 2008) is based on a research prototype of the German parser of the lingenio[4] machine translation product *translate*, adapted for the research purposes within SFB732 and therefore to the idea of a pipeline where each annotation level can be extracted separately. The stored analyses provide the complete knowledge needed by a subsequent analysis step of the pipeline. This doesn't mean that each instance of an analysis level provides all of the information needed, but as the analyses of a sentence are connected to each other by text and sentence identifiers, all levels may contribute to a more detailed analysis. The tool comprises modules for morphological, syntactic, semantic, and text semantic/pragmatic analyses.

5. Filling the architecture

In the project B3 of SFB732 we work on task-specific disambiguation of German *ung*-nominializations by indicators extracted from the context. A particularly interesting context is a PP with the preposition *nach*, in combination with nominalizations of *verba dicendi*, e.g. *Mitteilung* ('announcement'), *Anmerkung* ('remark'), *Meldung* ('notice'). In these cases two kinds of *nach*-readings are possible: a temporal one ('after') and one that refers to a content ('according to'); and two readings of the nominalization: an event reading ('the act of making an announcement') and an object reading ('the content of the announcement'), cf. (Eberle et al., 2009). The following example shows how the needed processing steps and data can be handled via the B3-tool/database interface.

5.1. Primary data

Sentence (1) occurs in a web article on local news.

(1) *Er verblieb* ***nach*** *seiner Mitteilung in stationärer Krankenhausbehandlung.*
He remained in stationary hospital treatment **after/according to** his announcement.

We extracted such sentences from corpora via standard corpus tools like CWB[5] (Hoffmann and Evert, 2006) and additional filtering (Haselbach et al., submitted). Figure 1 shows such a collection of sentences as an output of a database query.

5.2. Processing and storing the first step

Sometimes it is helpful to compute sentence analyses offline and to provide them for later inspection. This is the case if a specific investigation requires corresponding analysis information and if on-the-fly computation is too costly to be executed when searching the database for specific instances with criteria that relate to the corresponding analysis level.

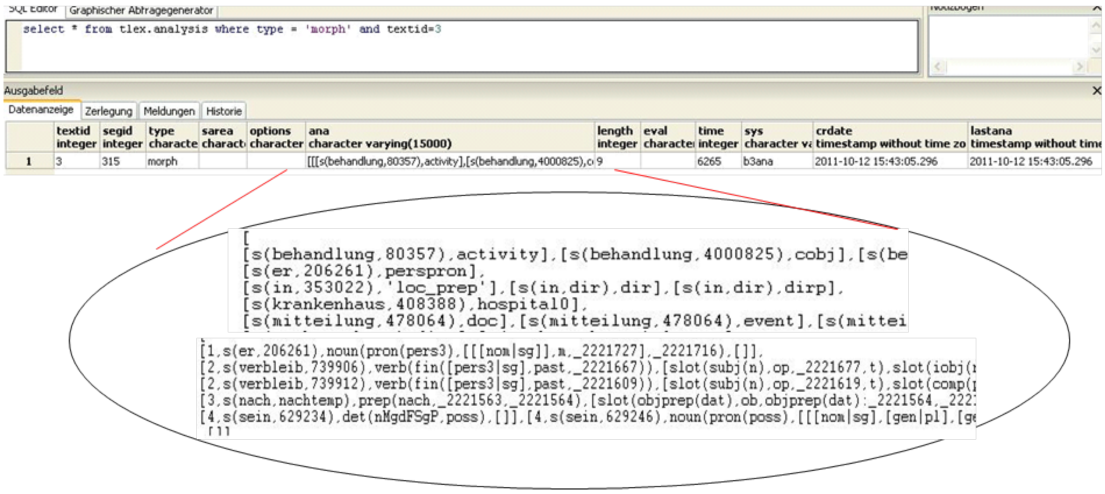
The analyses can be carried out by the *analysis frontend* of the database with respect to single sentences, texts or corpora respectively. Next to this information, the corresponding command must specify the type of input it assumes and the type of output it computes. In addition, it may specify a number of suitable parameters that may fine-tune the corresponding analysis. The general form is as follows:

(2)
```
dbanalyze(
analysis(InputID,InputAnalysisType),
Language,TypeofAnalysis,Domain,
AdditionalParameters
).
```

[3] http://www.postgresql.org/

[4] http://www.lingenio.de/English/Research/Cooperations/unis-ims-sfb732-b3.htm

[5] http://cwb.sourceforge.net/



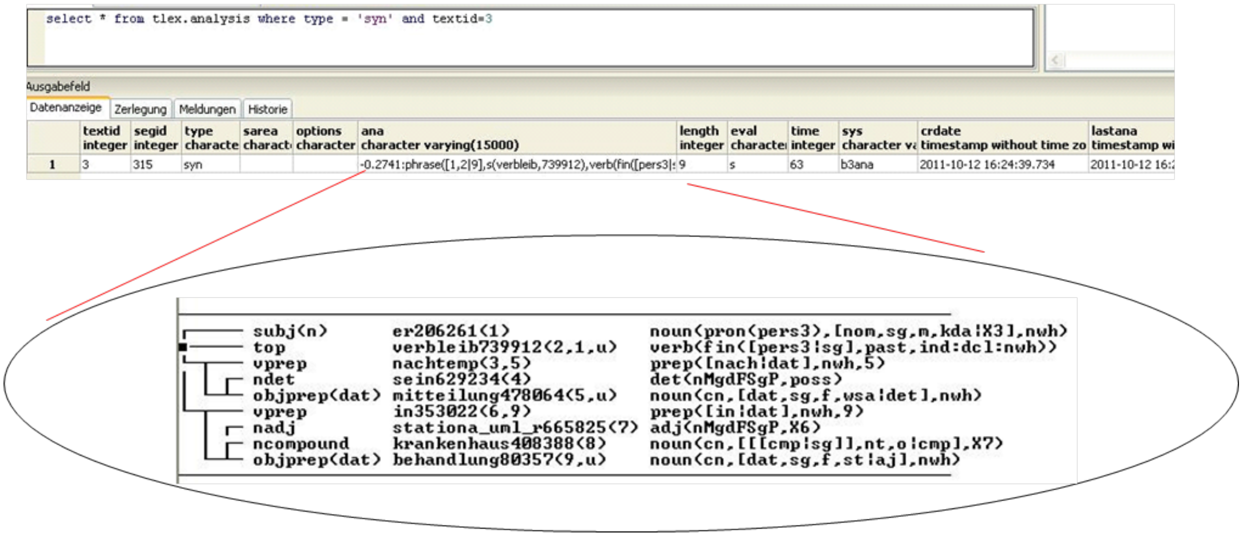
```
select * from tlex.analysis where type = 'morph' and textid=3
```

	textid integer	segid integer	type characte	sarea charact	options character	ana character varying(15000)	length integer	eval characte	time integer	sys character va	crdate timestamp without time zo	lastana timestamp without time
1	3	315	morph			[[[s(behandlung,80357),activity],[s(behandlung,4000825),c	9		6265	b3ana	2011-10-12 15:43:05.296	2011-10-12 15:43:05.296

```
[
[s(behandlung,80357),activity],[s(behandlung,4000825),cobj],[s(be
[s(er,206261),perspron],
[s(in,353022),'loc_prep'],[s(in,dir),dir],[s(in,dir),dirp],
[s(krankenhaus,408388),hospital0],
[s(mitteilung,478064),doc],[s(mitteilung,478064),event],[s(mittei
```

```
[1,s(er,206261),noun(pron(pers3),[[[nom|sg]],m,_2221727],_2221716),[]],
[2,s(verbleib,739906),verb(fin([pers3|sg],past,_2221667)),[slot(subj(n),op,_2221677,t),slot(iobj(
[2,s(verbleib,739912),verb(fin([pers3|sg],past,_2221609)),[slot(subj(n),op,_2221619,t),slot(comp(
[3,s(nach,nachtemp),prep(nach,_2221563,_2221564),[slot(objprep(dat),ob,objprep(dat):_2221564,_222
[4,s(sein,629234),det(nMgdFSgP,poss),[]],[4,s(sein,629246),noun(pron(poss),[[[nom|sg],[gen|pl],[g
[]]
```

Figure 2: Morphological analysis in the database with a text and segment id, the output string of the analysis (ana), the duration time of the analysis (time), the analysis tool (sys), the creation date (crdate) and the creation date of the last analysis conducted in the same configuration (lastana).



```
select * from tlex.analysis where type = 'syn' and textid=3
```

	textid integer	segid integer	type characte	sarea charact	options character	ana character varying(15000)	length integer	eval characte	time integer	sys character va	crdate timestamp without time zo	lastana timestamp wi
1	3	315	syn			-0.2741:phrase([1,2|9],s(verbleib,739912),verb(fin([pers3|s	9	s	63	b3ana	2011-10-12 16:24:39.734	2011-10-12 16:

```
┌──── subj(n)         er206261(1)               noun(pron(pers3),[nom,sg,m,kda|X3],nwh)
■──── top             verbleib739912(2,1,u)     verb(fin([pers3|sg],past,ind:dcl:nwh))
└──── vprep           nachtemp(3,5)             prep([nach|dat],nwh,5)
│  ┌─ ndet            sein629234(4)             det(nMgdFSgP,poss)
│  └─ objprep(dat)    mitteilung478064(5,u)     noun(cn,[dat,sg,f,wsa|det],nwh)
└──── vprep           in353022(6,9)             prep([in|dat],nwh,9)
   ┌─ nadj            stationa_uml_r665825(7)   adj(nMgdFSgP,X6)
   ┌─ ncompound       krankenhaus408388(8)      noun(cn,[[[cmp|sg]],nt,o|cmp],X7)
   └─ objprep(dat)    behandlung80357(9,u)      noun(cn,[dat,sg,f,st|aj],nwh)
```

Figure 3: Dependency tree produced by the syntactic component of the B3-tool.

Applied to sentence 315 of file 3 (“datei”, see Figure 1) with stipulations as in (3) we obtain a new database entry as reproduced in Figure 2[6].

(3) `dbanalyze(sent(3,315),de,morph,[],[]).`

5.3. Creating further steps directly or indirectly

If we need syntactic analyses we may get them directly by a corresponding command as in (4) or indirectly as in (5) via applying the analysis tool to the existing morphological information as given in Figure 2, thus making use of vertical relations (cf. Section 2.) and of the pipeline architecture of the analysis system.

(4) `dbanalyze(sent(3,315),de,syn,[],[]).`

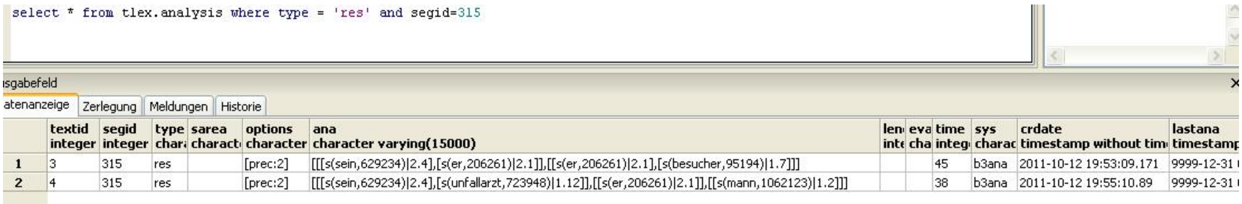
(5) `dbanalyze(`
`analysis(3,315,morph),de,syn,[],[]`
`).`

Figure 3 shows the result of applying a syntactic analysis step to the morphological description represented in Figure 2. The content of the analysis column of the corresponding entry can be represented by the analysis tool as in the ’zoom’ bubble.

5.4. Indicating readings by pronoun resolution

As indicated in Section 2., analyses are stored in order to allow quick access to references of particular phenomena investigated. For instance, if we search for a sentence with a VP that is modified by a *nach*-PP whose NP head is a verbum dicendi -*ung*-nominalization and whose determiner is a referential (possessive) pronoun, we will find the analysis of Figure 3 and the corresponding sentence.

[6] For space reasons only the first part of each entry is shown in the ’zoom’ bubble.



```
select * from tlex.analysis where type = 'res' and segid=315
```

	textid	segid	type	sarea	options	ana	len	eva	time	sys	crdate	lastana
	integer	integer	chara	characte	character	character varying(15000)	inte	cha	integ	charac	timestamp without tim	timestamp
1	3	315	res		[prec:2]	[[[s(sein,629234)|2.4],[s(er,206261)|2.1]],[[s(er,206261)|2.1],[s(besucher,95194)|1.7]]]			45	b3ana	2011-10-12 19:53:09.171	9999-12-31
2	4	315	res		[prec:2]	[[[s(sein,629234)|2.4],[s(unfallarzt,723948)|1.12]],[[s(er,206261)|2.1]],[[s(mann,1062123)|1.2]]]			38	b3ana	2011-10-12 19:55:10.89	9999-12-31

Figure 4: Here the analysis string (ana) is of type ’res’, denoting pronoun resolution. Text 3 corresponds to the context in (6), text 4 to the context in (7).

There are reasons to assume that knowing the antecedent of the possessive pronoun (and drawing inferences from this knowledge) helps to disambiguate the sentence. Compare the different possible contexts of Example (1) above: In Example (6) the (original) context resolves both *er* and *seiner* to *Herbstfestbesucher*, i.e. the person being in hospital treatment, which triggers a preference for the temporal reading of *nach*. In Example (7) *seiner* refers to the attending physician, *er* to the person being in hospital treatment and the preferred reading for *nach* here is the propositional one.

(6) *In der Nacht [...] teilte ein Herbstfestbesucher der Polizei [...] vom Wasserburger Krankenhaus aus mit, dass er auf dem Weg [...] zusammengeschlagen worden sei. Er verblieb nach seiner Mitteilung in stationärer Krankenhausbehandlung.*
In that night a visitor of the ’Herbstfest’ called the police from Wasserburg hospital and reported that he had been attacked on his way. He remained in stationary hospital treatment **after** his announcement.

(7) *Der Mann zog sich schwere Verletzungen zu, teilte der behandelnde Unfallarzt mit. Er verblieb nach seiner Mitteilung in stationärer Krankenhausbehandlung.*
The attending physician stated that the man was severely injured. He remained in stationary hospital treatment **according to** his announcement.

When investigating this type of ambiguity, a reasonable step is to resolve the pronouns used. Therefore, the context of the sentence must be known. The architecture of the database supports this type of knowledge processing by modularly administrating the content and structure of a new corpus or text when reading it in. As there are tables with information about adjacency of sentences in a specific text, contexts of any size can be reconstructed easily. If all of the sentences taken into account are already (syntactically) analysed, the module for pronoun resolution just has to be applied to the corresponding analyses; otherwise the missing analyses have to be computed before.
The call in (8) applies pronoun resolution to the syntactic analysis of sentence 315 of text 3 by taking into account two preceding sentences:

(8)
```
dbanalyze(
analysis(3,315,syn),de,res,[],[prec:2]
).
```

Figure 4 shows the results of resolution applied to the different contexts as in the example above.

6. Conclusion

We showed a tool/database-interface for the handling of analyses along three dimensions and we discussed an example of vertical multi-level processing in detail. Assigning creation date, expiration date and origin of the analysis to each entry also allows for comprehension of the history of an analysis and comparison of analyses provided by different tools. To extract information, the structures of the database require the user to have a detailed understanding of the mapping of the annotations to the database representation and also of the tools producing the analyses. The infrastructure therefore rather supports detailed project work but simplifies the creation of analysis levels by utilizing the analysis frontend. The design of a temporal database along with an analysis tool adapted to the idea of a pipeline architecture supports fast, reliable, and, on the same system, also reproducible analyses.

7. Acknowledgements

This work was funded by the German Research Foundation DFG, via the Collaborative Research Centre SFB 732 *Incremental Specification in Context*.

8. References

Kurt Eberle, Ulrich Heid, Manuel Kountz, and Kerstin Eckart. 2008. A tool for corpus analysis using partial disambiguation and bootstrapping of the lexicon. In Angelika Storrer et al., editor, *Text Resources and Lexical Knowledge – Selected Papers from the 9th Conference on Natural Language Processing (KONVENS 2008)*, pages 145 – 158, Berlin. Mouton de Gruyter.

Kurt Eberle, Gertrud Faaß, and Ulrich Heid. 2009. Proposition oder Temporalangabe? Disambiguierung von -ung-Nominalisierungen von verba dicendi in nach-PPs. In Christian Chiarcos et al., editor, *Von der Form zur Bedeutung: Texte automatisch verarbeiten / Proceedings of GSCL 2009*, pages 81 – 91, Tübingen. Gunter Narr Verlag.

Kerstin Eckart, Kurt Eberle, and Ulrich Heid. 2010. An Infrastructure for More Reliable Corpus Analysis. In *Proceedings of the Workshop on Web Services and Processing Pipelines in HLT (LREC’10)*, pages 8–14, Valletta, Malta, may.

Kerstin Eckart, Arndt Riester, and Katrin Schweitzer. 2012. A discourse information radio news database for linguistic analysis. In Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, editors, *Linked Data in Linguistics. Representing and Connecting Language*

Data and Language Metadata, pages 65–75. Springer, Heidelberg.

Boris Haselbach, Wolfgang Seeker, and Kerstin Eckart. submitted. German nach particle verbs in semantic theory and corpus data.

Sebastian Hoffmann and Stefan Evert. 2006. BNCweb (CQP-edition): The marriage of two corpus tools. In S. Braun et al., editor, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, volume 3 of *English Corpus Linguistics*, pages 177 – 195. Peter Lang, Frankfurt am Main.

Nancy Ide and Keith Suderman. 2007. GrAF: A graph-based format for linguistic annotations. In *Proceedings of the Linguistic Annotation Workshop, 1-8*.

Ellen F. Prince. 1992. The ZPG Letter: Subjects, Definiteness and Information Status. In W. C. Mann and S. A. Thompson, editors, *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text*, pages 295–325. Benjamins, Amsterdam.

Arndt Riester, David Lorenz, and Nina Seemann. 2010. A Recursive Annotation Scheme for Referential Information Status. In *Proceedings of the Seventh LREC*, pages 717–722, Valletta, Malta.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, pages 44 – 49, Manchester, UK.