The Potsdam Commentary Corpus 2.1 in ANNIS3

Peter Bourgonje, Manfred Stede

Applied Computational Linguistics
University of Potsdam, Germany

`bourgonje|stede@uni-potsdam.de`

ABSTRACT
We present a new version of the Potsdam Commentary Corpus; a German corpus of news commentary articles annotated on several different layers. This new release includes additional annotation layers for dependency trees and information-structural aboutness topics as well as some bug fixes. In addition to discussing the additional layers, we demonstrate the added value of loading the corpus in ANNIS3, a tool to merge different annotation layers on the same corpus and allow for queries combining information from different annotation layers. Using several cross-layer example queries we demonstrate its suitability to corpus analysis for various different areas.

KEYWORDS: treebanks, information structure, cross-layer analysis.

1 Introduction

The Potsdam Commentary Corpus (PCC) was introduced by Stede (2004) as a collection of 176 newspaper editorials (comprising over 34k words in over 2100 sentences) from a German regional newspaper, which had been manually annotated on different layers: sentence syntax, coreference, and rhetorical structure. More recently, an updated version (PCC 2.0) was presented by Stede and Neumann (2014); besides major revisions on the rhetorical structure and coreference layers, it included an additional layer of connectives and their arguments, similar in spirit to the annotations in the Penn Discourse Treebank (Prasad et al., 2008).

In this paper, we present the new release PCC 2.1, which offers three new features:

- A new layer of manually-annotated information-structural aboutness topics (cf. (Stede and Mamprin, 2016))
- A new layer of automatically-produced dependency parses
- Availability of the 'manual' layers in the ANNIS3 linguistic database (Krause and Zeldes, 2016)

The integration in ANNIS3 allows for qualitative studies on the interactions among the various layers of annotation, and visualization of search results.

In the following, we provide background information on the corpus and the ANNIS3 system, briefly discuss the technical conversion from annotation tool files to ANNIS3, and then discuss some sample queries in order to illustrate the potential for cross-layer analyses. The paper concludes with an outlook on our next steps.

2 Corpus and Database

2.1 PCC

The PCC was deliberately built as a genre-specific corpus, mainly intended for studying the textual means for expressing opinion and argumentation in the German language. In order to support these high-level goals, a number of lower-level (i.e., closer to the linguistic surface) phenomena have been annotated manually, so that a gold standard is available for evaluating automatic experiments, and also for manually studying the interactions of phenomena. Thus, sentence syntax was annotated (in the early 00's) by the Potsdam team of the TIGER project (Brants et al., 2002), and (nominal) coreference annotation built on the proposals of the PoCoS annotation scheme suggested by Krasavina and Chiarcos (2007). In recent years, all annotations (except for syntax) have been revisited and sometimes changed, so that they now reflect the annotation guidelines, which are collectively available in the volume (Stede, 2016).

Text structure, in the spirit of Rhetorical Structure Theory (Mann and Thompson, 1988), has been at the center of interest in the PCC from the beginning. The main postulate of RST is that a text can be segmented into so-called *elementary discourse units* (sentences, certain types of clauses), and that *coherence relations* connect adjacent text spans – which can be either elementary units, or recursively-built larger spans. A set of some 20 relations is suggested, with various causal, temporal, contrastive and additive relations among them. For the majority of the relations, the connected units have different statuses of prominence: A *nucleus* is more important for the author's purposes, and the *satellite* unit supports that purpose but is overall less important.

(Matthiessen and Thompson (1988) discuss the relationship between the nucleus/satellite dichotomy and syntactic subordination.) By this analysis, a text is ultimately represented as a single tree structure that spans the entire text. See (Mann and Thompson, 1988) for a full explanation. The RST annotations in the PCC served as train/test data for the first RST parser developed for German (Reitter, 2003). Until today, PCC is the largest German-language RST resource. The ANNIS team built a specific visualization module for these discourse trees, which reflect the recursive application of coherence relations.

A similar layer of annotations holds the lexical connectives (coordinating and subordinating conjunctions, various adverbials, a few prepositions)[1] and their arguments. However, following the spirit of the Penn Discourse Treebank (Prasad et al., 2008), they do not combine into any description of a text structure; instead, they are annotated individually and without taking other connective/argument constellations into consideration. This, in turn, allows for posthoc studying the correlations between connectives/arguments on the one hand, and rhetorical structure on the other. Notice that, in contrast to the PDTB, *implicit* coherence relations have not been annotated in the connective layer of the PCC, as this would effectively duplicate parts of the RST annotation task.

The latest layer of annotation concerns *aboutness topics*, with annotation guidelines being inspired by the characterization given by Jacobs (2001). In line with earlier work, we regard the aboutness topic as the syntactic constituent referring to the entity about which information is communicated by the central predication of the sentence. According to Jacobs, a ‘prototypical’ aboutness topic fulfils three criteria:

- Informational separation: The topic precedes the predication, and semantic processing of the sentence is correspondingly done in two subsequent steps.
- Predication: The topic fills a position in the valency frame of the predicate. (It is not an adjunct.)
- Addressation: The topic refers to an entity that serves as the ‘address’ for storing information in the common ground of speaker and hearer.

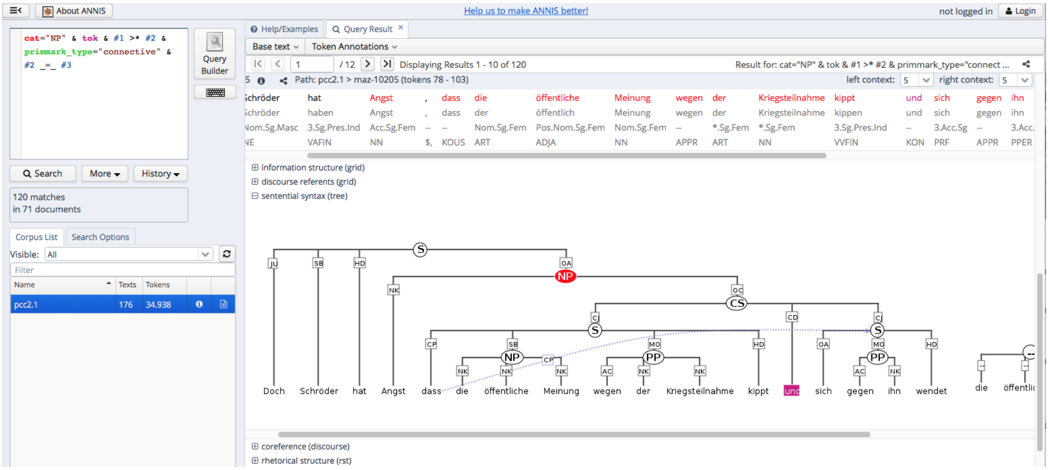
Reliably identifying topics in authentic text as opposed to single-sentence “laboratory examples” can be difficult, though (Cook and Bildhauer, 2013). Therefore, our guidelines had to make a range of additional commitments, concerning primarily the partitioning of complex sentences into clauses that should (or should not) receive a topic annotation, and the handling of incomplete sentences, i.e., fragmentary material. The annotation effort is documented in (Stede and Mamprin, 2016); annotator agreement is κ 0.6 for the theticity question (should a discourse unit be assigned a topic or not), and κ 0.71 for selecting the aboutness topic.

Finally, with release 2.1 we now began to also include automatic annotations for the purposes of providing training/test data for further automatic analysis tasks. Specifically, we added dependency parses produced by the ParZu system (Sennrich et al., 2009).[2]

In addition to these extra layers, the 2.1 release includes several minor bug fixes on different layers of the manual annotation.

[1] For discussion, see (Danlos et al., 2018), and for lists of connectives in various languages http://www.connective-lex.info

[2] Thanks to Don Tuggener for providing us with this data.



```
cat="NP" & tok & #1 >* #2 &
primmark_type="connective" &
#2 _=_ #3
```

120 matches
in 71 documents

Name	Texts	Tokens
pcc2.1	176	34.938

1 / 12 Displaying Results 1 - 10 of 120

Result for: cat="NP" & tok & #1 >* #2 & primmark_type="connect ...

5 Path: pcc2.1 > maz-10205 (tokens 78 - 103)

left context: 5 right context: 5

Schröder	hat	Angst	,	dass	die	öffentliche	Meinung	wegen	der	Kriegsteilnahme	kippt	und	sich	gegen	ihn
Schröder	haben	Angst	,	dass	der	öffentlich	Meinung	wegen	der	Kriegsteilnahme	kippen	und	sich	gegen	ihn
Nom.Sg.Masc	3.Sg.Pres.Ind	Acc.Sg.Fem	--	--	Nom.Sg.Fem	Pos.Nom.Sg.Fem	Nom.Sg.Fem	--	*.Sg.Fem	*.Sg.Fem	3.Sg.Pres.Ind	--	3.Acc.Sg	--	3.Acc.
NE	VAFIN	NN	$,	KOUS	ART	ADJA	NN	APPR	ART	NN	VVFIN	KON	PRF	APPR	PPER

Figure 1: Screenshot of ANNIS3 Database

2.2 ANNIS

The first version of the ANNIS database (Dipper et al., 2004) was developed at Potsdam University in order to support research on information structure by providing an infrastructure for merging (manual or automatic) annotations of the same data into a single representation, and allowing queries across the various layers. Subsequently, further developments, including a complete redesign, was done largely at Humboldt University Berlin, and the scope of applications widened considerably. The latest release ANNIS3[3] offers, *inter alia*, the inclusion of audio/video data and support for many languages and character sets, and it contains modules for mapping the output of many manual annotation tools to the native database format. The system is available both as a server version and as a standalone app that can be run on laptops. See (Krause and Zeldes, 2016) for an overview.

The different visualization modules of ANNIS3 provide

- graphic representations of constituency trees and dependency trees for sentence syntax;
- graphic representations of discourse trees (inspired by the format used in RSTTool[4]);
- highlighting of co-referring elements in a text view;
- presentation of other kinds of span annotations in a "grid" view (inspired by the format used in Exmaralda[5]).

In addition, parallel aligned data can be represented, and inspected with the help of co-colouring the aligned elements.

The Annis Query Language (AQL) allows to search for patterns in the primary text (e.g., via regular expresions) and in the annotations. As for the latter, one can look for simple labels

[3] http://corpus-tools.org/annis/
[4] http://www.wagsoft.com/RSTTool/
[5] http://exmaralda.org/de/partitur-editor-de/

associated with annotation objects, direct and indirect dominance relations in trees, and for so-called 'pointing relations' as they can be used, for example, for coreference annotation. For illustration, Figure 1 shows a screenshot with a query, set of hits, and a syntax tree selected for visualization. This is one example of a query spanning across multiple layers of annotations and hence combining different linguistic phenomena; it will be discussed in Section 4.

3 Tranferring annotation files to ANNIS3

ANNIS3 comes with separate tools for importing, merging and exporting data sets; Salt[6] and Pepper[7]. Pepper, the tool responsible for importing data from several formats, in turn comes with a collection of existing importing modules, supporting many popular annotation file formats, including CoNLL, MMAX2, PTB, Tiger2 and many more[8]. For all but one annotation layer of PCC 2.1, an importer was readily available to merge the annotations of different layers into the format that is directly read by ANNIS3; the aboutness topics were annotated in the EXMARaLDA format; the sentential syntax (constituency trees) were available in the Tiger2 format; the discourse referents (coreference layer) were in MMAX2 format, and the RST trees in rs3 format (the native format of the RST Tool). The exception for which no importer was available, was the annotation layer for discourse relations, which were in a custom inline XML format, as produced by the annotation tool Connanno[9]. Although Pepper comes with documentation on how to extend the toolkit with additional importing modules, in this case it was simpler to convert this custom XML format to a supported one (MMAX2) and then to import. From there on, all annotation layers are easily merged, with the important prerequisite being that tokenisation is consistent across all annotation layers. After importing and merging all layers, the result is exported to a format that is finally loaded into the PostgreSQL database, which ANNIS3 is reading from and then enables querying and visualisation.

4 Querying the multi-layer PCC

In this section, we provide some examples to illustrate the potential of cross-layer queries in the PCC. Notice that the various layers have all been annotated independently of each other (and at quite different points in time), so that exploring correlations does not simply reproduce the complete set of choices made in one shot by a single annotator.

Interesting uses of connectives. The German word "da" has a number of readings, one of them being that of a subordinate conjuntion, where it routinely plays the role of a causal (or argumentative) connective. To begin, the simple query `"da" | "Da"` yields 56 hits, and 8 of these are labelled as connectives. Zooming in on subordinate clauses following the main clause (i.e., on the non-capitalied "da"), we get 4 connectives. Of these, one turns out to be somewhat non-standard because the syntax layer does not tag it as subordinate conjunction. The corresponding query is:

```
tok = "da" & primmark_type = "connective" & pos != "KOUS"
& #1 _=_ #2 & #2 _=_ #3
```

And the hit is *jetzt, da das gesamte Paket überschaubar wird* ('now, as the complete package becomes visible'). And indeed, the annotated sense of this connective is not, as usual, causal but temporal.

[6] http://corpus-tools.org/salt/

[7] http://corpus-tools.org/pepper/

[8] See http://corpus-tools.org/pepper/knownModules.html for the exhaustive list.

[9] http://angcl.ling.uni-potsdam.de/resources/connanno.html

Embedded discourse units. One point of debate in discourse structure theories (e.g., (Hoek et al., 2018)) is the proper handling of 'elemenary discourse units' that are syntactically embedded. Using the AQL operator for transitive domination in trees, we can for instance look for connectives that are embedded in an NP:

```
cat="NP" & tok & #1 >* #2 & primmark_type="connective" & #2 _=_ #3
```

This query, which is shown above in Figure 1, yields 120 hits. The screenshot shows the relatively common case of a noun modified by a complement clause ('the fear that public mood will change because of entering the war *and* turn against him'). With queries like this, such "isolated" connective annotations can be set in correspondence with the syntactic configurations, and then the consequences for an approach to representing complete discourse structures, for example in RST, can be pondered.

Aboutness topic and grammatical role. The "typical" aboutness topic in a sentence is the grammatical subject (cf., e,g., (Jacobs, 2001)). Using ANNIS queries, we can compute the set of topics that are *not* subjects and then determine the reasons. The PCC contains 1.417 topics, 1.184 of which have a grammatical role annotation (the others merely overlap with a syntactic unit with a role annotation). Of these, in turn, 347 have a role other than subject, which amounts to 24% of all aboutness topics. A qualitative investigation of 90 instances shows that the largest group is due to the subjects being impersonal (e.g., *man*/'one') or expletive pronouns, which prompted annotators to associate the topic label with a different constituent. In the group of more interesting cases, we find subjects that are non-specific NPs, or they are discourse-new yet hearer-old, which makes other candidates in the sentence more suitable topics, given Jacobs' three criteria quoted above in Section 2.1. See Bourgonje and Stede (to appear) for details.

5 Summary and Outlook

The PCC 2.1 release is now available both as raw corpus with the original annotation tool output files[10], and ready for inspection in the ANNIS3 interface[11]. Its multi-layer architecture makes it a suitable resource for corpus analyses of different types, easily combining information from different annotation layers through the ANNIS3 query engine; we gave three examples of queries in the AQL language. Another use case for annotated corpora is to serve as train/test data for specific applications, such as described in (Reitter, 2003) for RST parsing. Recently, Tuggener (2016) employed the PCC for testing his German coreference resolver. As a final example, in (Bourgonje and Stede, 2018), the connective layer has been exploited to train an automatic connective classifier. To establish the impact of parsing errors, the manually-annotated syntax trees (originating from a different annotation layer than the connective one) have been used.

While more annotation layers are not imminent right now, one step we foresee is to add information on sentence semantics and pragmatics, in order to build a further bridge between sentential syntax and the discourse-level annotations. This may concern 'semantic entity types' (as annotated by Becker et al. (2016) on a different German corpus) as well as expressions of different kinds of subjectivity.

One potential extension of an existing layer addresses the discourse connectives. As mentioned in Section 2, the connective layer now contains explicit relations only. To expand this to also cover implicit relations (and potentially also alternative lexicalizations, entity relations and 'no

[10] http://angcl.ling.uni-potsdam.de/resources/potsdam-commentary-corpus-2.1.zip

[11] https://korpling.org/annis3/#_c=cGNjMi4x

relations', as annotated in the PDTB), currently the RST layer can be exploited: For two adjacent text spans for which no explicit relation has been annotated, an RST relation may exist from which an implicit relation at the shallow level can (semi-)automatically be derived. However, RST and PDTB use different sets of relations, and – more importantly – it can be interesting to actually compare annotator's assignments of coherence relations (i) with and (ii) without the requirement of an overall spanning structural representation. Hence, we plan to work toward a complete, PDTB-style layer of annotation.

Finally, work is ongoing to expand the text base. A currently non-public part of the PCC contains editorials from *Der Tagesspiegel*, with some of the annotation layers mentioned above. Following an agreement with the publisher we hope to be able to release a bigger corpus (albeit not with all the layers) in the not too distant future.

References

Becker, M., Palmer, A., and Frank, A. (2016). Argumentative texts and clause types. In *Proceedings of the Third Workshop on Argumentation Mining*, Berlin. Association for Computational Linguistics.

Bourgonje, P. and Stede, M. (2018). Identifying explicit discourse connectives in German. In *Proceedings of the 19th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2018)*, pages 327–331, Melbourne, Australia. Association for Computational Linguistics.

Bourgonje, P. and Stede, M. (To appear). Topics and subjects in German newspaper editorials: A corpus study.

Brants, S., Dipper, S., Hansen, S., Lezius, W., and Smith, G. (2002). The TIGER treebank. In *Proc. of the Workshop on Treebanks and Linguistic Theories*, Sozopol.

Cook, P. and Bildhauer, F. (2013). Identifying 'aboutness topics': two annotation experiments. *Dialogue and Discourse*, 4(2):118–141.

Danlos, L., Rysova, K., Rysova, M., and Stede, M. (2018). Primary and secondary discourse connectives: definitions and lexicons. *Dialogue and Discourse*, 9(1):50–78.

Dipper, S., Götze, M., Stede, M., and Wegst, T. (2004). Annis: A linguistic database for exploring information structure. In *Interdisciplinary Studies on Information Structure*, ISIS Working papers of the SFB 632 (1), pages 245–279.

Hoek, J., Evers-Vermeul, J., and Sanders, T. (2018). Segmenting discourse: Incorporating interpretation into segmentation? *Corpus Linguistics and Linguistic Theory*, 14(2):357–386.

Jacobs, J. (2001). The dimensions of Topic–Comment. *Linguistics*, 39(4):641–681.

Krasavina, O. and Chiarcos, C. (2007). PoCoS: The Potsdam Coreference Scheme. In *Proc. of the Linguistic Annotation Workshop (LAW) at ACL-07*, Prague.

Krause, T. and Zeldes, A. (2016). ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*, 31.

Mann, W. and Thompson, S. (1988). Rhetorical Structure Theory: Towards a functional theory of text organization. *TEXT*, 8:243–281.

Matthiessen, C. and Thompson, S. (1988). The structure of discourse and ‘subordination’. In Haiman, J. and Thompson, S., editors, *Clause combining in grammar and discourse*, pages 275–329. John Benjamins, Amsterdam.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse Treebank 2.0. In *Proc. of the 6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.

Reitter, D. (2003). Simple signals for complex rhetorics: On rhetorical analysis with rich-feature support vector models. *LDV Forum*, 18(1/2):38–52.

Sennrich, R., Schneider, G., Volk, M., and Warin, M. (2009). A new hybrid dependency parser for German. In Chiarcos, C., de Castilho, R. E., and Stede, M., editors, *From Text to Meaning: Processing Text Automatically. Proceedings of the Biennial GSCL Conference 2009*, pages 115–124, Tübingen. Narr.

Stede, M. (2004). The Potsdam Commentary Corpus. In *Proc. of the ACL Workshop on Discourse Annotation*, pages 96–102, Barcelona.

Stede, M., editor (2016). *Handbuch Textannotation: Potsdamer Kommentarkorpus 2.0*, volume 8 of *Potsdam Cognitive Science Series*. Universitätsverlag, Potsdam.

Stede, M. and Mamprin, S. (2016). Information structure in the Potsdam Commentary Corpus: Topics. In *Proc. of the Ninth International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia. European Language Resources Association (ELRA).

Stede, M. and Neumann, A. (2014). Potsdam Commentary Corpus 2.0: Annotation for discourse research. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 925–929, Reikjavik.

Tuggener, D. (2016). *Incremental Coreference Resolution for German*. PhD thesis, University of Zurich, Faculty of Arts.