Potsdam Commentary Corpus 2.0: Annotation for Discourse Research

Manfred Stede, Arne Neumann
Applied Computational Linguistics
EB Cognitive Science
Universität Potsdam / Germany
stede@uni-potsdam.de, arne.neumann@uni-potsdam.de

Abstract

We present a revised and extended version of the Potsdam Commentary Corpus, a collection of 175 German newspaper commentaries (op-ed pieces) that has been annotated with syntax trees and three layers of discourse-level information: nominal coreference, connectives and their arguments (similar to the PDTB, (Prasad et al., 2008)), and trees reflecting discourse structure according to Rhetorical Structure Theory (Mann and Thompson, 1988). Connectives have been annotated with the help of a semi-automatic tool (*Conano,* (Stede and Heintze, 2004)) that identifies most connectives and suggests arguments based on their syntactic category. The other layers have been created manually with dedicated annotation tools. The corpus is made available on the one hand as a set of original XML files produced with the annotation tools, based on identical tokenization. On the other hand, it will be distributed together with the open-source linguistic database ANNIS3 (Chiarcos et al., 2008; Zeldes et al., 2009), which provides multi-layer search functionality and layer-specific visualization modules. This allows for comfortable qualitative evaluation of the correlations between annotation layers.

Keywords: Discourse annotation, Multi-layer annotation, Argumentative text

1. Introduction

Multi-layer discourse annotation (see, e.g., (Stede, 2008a)) aims at advancing discourse research by providing corpora that are annotated for a variety of linguistic phenomena and can be queried for combinations of features on different layers, so that correlations or dependencies between different levels of description can be studied both qualitatively and quantitatively.

In this vein, the *Potsdam Commentary Corpus* (PCC) was first released ten years ago (Stede, 2004). Its idea is to provide a genre-specific corpus in German that is relatively small but features rich manual annotation geared towards research on various discourse phenomena. The genre of newspaper commentary was chosen because the study of subjectivity and argumentation is a major goal of the work surrounding this corpus.

In the first collection effort, we gathered 1268 commentaries from *Märkische Allgemeine Zeitung*, a German local newspaper. This choice was made because we were looking for relatively short texts, simple language, and simple argumentation structure, so that the modelling tasks and an automatic analysis do not have to deal right away with all the difficulties of elaborate opinionated texts as they are published in larger newspapers. The raw texts have been used for a variety of statistical tasks (finding distributions of connectives, building models for text genre identification, etc.).

One major research interest in our research group is the study of argumentation. For this reason, over the past five years we also collected texts from the *pro and contra* page of the Berlin-based newspaper *Tagesspiegel am Sonntag.* The page always deals with one current "hot topic" of local, national, or international relevance; it provides a background article plus two opposing views on the issue. In comparison to the MAZ texts, the pro and contra pieces reliably feature crisp argumentation that clearly leads to a conclusion (an answer to a yes/no question). However, it is difficult to reach a distribution agreement with the publisher, so that the pro and contra texts are so far not part of the released version of the PCC.

In the following, we first describe the annotations that were originally made for the first version of PCC. Then, Section 3. gives an overview of the improvements and extensions made for the new release 2.0. A major change is the addition of a new annotation layer — connectives and their arguments —, which we describe in Section 4. Another novelty is that the corpus will be made available via our web-based linguistic database ANNIS3, which allows for multi-layer querying and visualization. This is dicussed in Section 5. Finally, Section 6. gives an outlook on additional layers of annotation that we are planning to build.

2. Background: PCC Version 1.0

The subset used for annotation in PCC 1.0 (and 2.0) consists of 175 texts. Their typical length is 8 to 10 sentences, with 15.8 words on average and 1.8 verbs per sentence; the total number of tokens in the subset is roughly 32000.

2.1. Syntax

All texts were semi-automatically annotated with sentence syntax following the TIGER scheme (Brants et al., 2004), which is also the basis of the largest available syntactically-annotated German corpus.[1] (Our PCC syntax annotators were part of the group that worked on the actual TIGER corpus at the time.) The annotation tool *annotate*[2] suggests a parse tree to the annotator, who can inspect it and revise where necessary. The TIGER scheme aims at integrating the advantages of constituency and dependency

[1] http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html

[2] http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/annotate.html

analysis; trees are relatively flat, and they employ so-called *secondary edges* to represent long-distance dependencies, where secondary edges are allowed to cross the primary edges.

2.2. Nominal Coreference

The coreference annotation of PCC 1.0 followed the 'base' part of the *PoCoS* annotation scheme (Krasavina and Chiarcos, 2007). This 'base' annotation covers only nominal coreference established by the identity relation: there is no event anophara, nor indirect coreference ('bridging'). The annotation tool is MMAX2[3] (Müller and Strube, 2006), which allows for comfortable annotation of relations between words or phrases in running text (such as coreference links). Unlike the syntax part, the coreference annotation is done completely manually, without any automatic suggestions.

2.3. Rhetorical Structure

PCC was the first German text collection to be annotated according to Rhetorical Structure Theory (Mann and Thompson, 1988). For the 1.0 version ten years ago, the annotators worked solely on the basis of the directions and relation definitions given in the Mann/Thompson paper. RST posits that a coherent text can be assigned a tree structure that results from establishing discourse relations between adjacent spans of text; the same set of about 20 relations is used for joining the minimal units of analysis (roughly: clauses) and, recursively, the larger units. Relations are defined in terms of speaker intentions. Most of them adjoin two segments of different weight: the more important one is called the 'nucleus', the less important, supportive, one is the 'satellite'.

The popular annotation tool for this (completely manual) analysis is RSTTool[4] (O'Donnell, 2000), which allows for segmenting the text and then stepwise constructing the tree structure.

An early use case for PCC was the development of the first SVM-based automatic RST parser (Reitter, 2003). Other studies included investigations into the relationships between rhetorical structure and coreference (Krasavina et al., 2007). Also, the experiences with the substantial amount of manual RST analyses lead to proposals for disentangling the notion of nuclearity (Stede, 2008b).

3. PCC 2.0: Overview of additions and improvements

In 2013, PCC was extended and revised to a large extent, which lead to the release of version 2.0. While we saw no need to change the syntax annotations, the other two layers have been completely revised.

- Coreference: Annotation guidelines have been largely rewritten and various clarifications made. In particular, changes were made to the definition of markables, with respect to certain types of pronouns. Guidelines for defining markables were separated from those for establishing coreference links (see below). Then, all coreference annotations were checked and corrected in accordance with these revisions.
- Rhetorical structure: Annotation guidelines have been devised that not only state the relation definitions (with a few changes to the Mann/Thompson definitions, reflecting certain features of the genre), but also suggest a specific, stepwise annotation procedure, with hints on nuclearity assignment, attachment points, etc. In order to reflect these changes, and to remove a number of mistakes from the earlier annotations, all RST trees have been checked and revised where necessary.

A general design decision for the corpus is to establish only minimal inter-dependencies between annotation layers. In contrast to an approach such as implemented in the Prague Dependency Treebank (Hajičová et al., 2001), at present we do not aim at an integrated, theoretically-motivated linking of all the different annotations. In particular, our syntax annotation is *not* the systematic basis for the other annotation layers. Instead, most of the annotation guidelines are designed to work on un-analyzed surface text, so that correlations between independent layers can be explored afterwards. As an exception to this rule, there are currently two 'base layers' that are being re-used in other layers:

- A layer of *discourse segments* is used as elementary units for RST, and also for illocutions (currently under development). Reason: One research direction is to develop an enhanced version of RST that more systematically accounts for the pragmatic status of the elementary units.
- A layer of *nominal referring expressions* provides the units for annotating both coreference and information status (given/new/etc.; currently under development). Reason: Coreference and information status both are aspects of *information structure*, and to study this systematically, there has to be a common set of referring expressions.

Regarding the accessibility of the corpus, we negotiated with the publisher of *Märkische Allgemeine Zeitung* that the raw data can be freely distributed. Hence, the 175 texts along with the annotations described in this paper are made avaliable via the the corpus website[5]. That page also contains current information on all layers, such as annotation guidelines and results of inter-annotator agreement studies.

4. New annotation layer: Connectives and their arguments

In the annotation of connectives and their arguments, we followed the general practice of the Penn Discourse Treebank (Prasad et al., 2008), but with a few modifications. Most importantly, only explicit connectives are being annotated; implicit relations are not covered. In addition, we do not annotate sense relations. The reason for these two deviations from the PDTB scheme is that we are interested in correlating this layer with others, in particular to syntax

[3] http://mmax2.sourceforge.net

[4] http://www.wagsoft.com/RSTTool

[5] http://www.ling.uni-potsdam.de/acl-lab/Forsch/pcc/pcc.html

and to RST, where rhetorical relations have already been annotated (irrespective of the presence of connectives).

The third difference from the PDTB scheme is that we constrain lexical signals to the closed-class set of connectives. For German, this set has been studied in depth by Pasch et al. (2003), who came up with an inventory of 350 connectives. We encoded many of these in DiMLex (Stede, 2002), our lexicon of connectives, which also provides syntactic information, associated coherence relations, and some other features of connectives. An abridged version of DiMLex in turn is used by and distributed with our semi-automatic annotation tool Conano (Stede and Heintze, 2004)[6], which in a text automatically detects words that can be connectives (but might have other readings; see below). When the user confirms this, Conano uses the syntactic features from the lexicon in order to guess the arguments (based on heuristic rules operating on the text surface). The user can then confirm or edit the scope, and proceed to the next connective candidate. However, annotators are strongly encouraged to watch out in the text for any additional connectives that are not being suggested by Conano.

Of the 175 PCC texts, 20 have been double-annotated in two different experiments for measuring inter-annotator agreement. In Experiment 1, two lightly-trained annotators used Conano for the annotation. As for the presence of a connective, both annotators agreed in 74.5% of their annotations. To a small extent, the disagreement seems to be influenced by Conano's suggestions: Annotators made different decisions on adding words that had not been highlighted by the tool as connectives. The larger problem is difficult cases of ambiguity with adverbials, which sometimes act as connectives and sometimes do not. As found by (Dipper and Stede, 2006), in German, about 40% of the connective words also have other, non-connective readings, and this includes many high-frequent words. To give just one example, German *auch* (English 'too') can be a connective if it has scope over a full clause, or a focus particle that typically has scope over just an NP. Due the free word order in German, in many cases both a narrow-scope and wide-scope reading seem intuitively plausible. A future version of the annotation guidelines will make more specific suggestions on handling such cases.

Regarding connective scope, there (naturally) is always overlap between the two annotations. If we consider items where the arguments match exactly, agreement is not very good (46.6%, or 50.0% if we allow them to differ in one token). Manual inspection revealed that scope disagreement occurs (not surprisingly) most frequently with adverbial connectors, which often leave some room for interpretation, especially with their "external" argument ('Arg1' in PDTB terminology). Another common source for disagreement were adjacent connectives with overlapping or even identical arguments (e.g. *aber doch* or *denn schliesslich*). Again, we hypothesize that Conano's suggestions of argument scopes influence annotators in different ways.

In Experiment 2, two trained annotators (one of them is a co-author of this paper) annotated the same data as in Experiment 1, but *without* using Conano. Agreement on the presence of a connective was 83.3%; an error analysis revealed a small number of oversights and again, quite a few problems with ambiguous adverbials. For the connectives annotated by both annotators, agreement on argument extension (modulo 1 token, to account for different handling of punctuation symbols) is 90.7%, i.e., much higher than in Experiment 1. For judging the difference between the experiments, It is difficult to tease apart the influence of the two factors use-of-Conano and amount-of-training, so that an in-depth study of the role of Conano's suggestions is an item for future work.

5. New access method: PCC in ANNIS3

PCC 2.0 is being distributed to interested parties in two different ways:

- Source files from the annotation tools:
  - Syntax: TIGER XML
  - RST: RS3 XML (the format of RSTTool[7])
  - Coreference: MMAX2[8] XML
  - Connectives: Conano XML (Stede and Heintze, 2004)
- Direct online access: Search and retrieval with the ANNIS3 database

ANNIS3[9] is a web-based database and search tool, which was designed to query and visualize linguistic corpora with multiple layers of annotation (Chiarcos et al., 2008; Zeldes et al., 2009). Unlike other linguistic query tools, it is not tied to a specific corpus and is able to visualize a wide range of linguistic annotations (e.g. spans, pointing relations, DAGs with labelled edges).

ANNIS3 can incorporate annotations produced by various tools (e.g., MMAX2, EXMaRaLDA, RSTTool, annotate/Synpathy) with the help of the SaltNPepper (Zipser and Romary, 2010; Zipser et al., 2011) linguistic converter framework.

Once corpora have been imported, the ANNIS Query Language (AQL) allows users to search for specific token values and annotations as well as relationships between them, even accross annotation layers created with different tools. Token values are represented as text between quotes (e.g. `"hat"`), while annotations are specified as attribute-value pairs (e.g. `pos="NN"`, a part-of-speech attribute with the value `NN`).

Relations among elements are indicated by back-referencing variable with incremental numbers, e.g. #1, #2, etc. Linguistically-motivated operators bind the elements together; e.g. `#1 > #2` means that the first element dominates the second in a tree. Operators can express overlap and adjacency between annotation spans, as well as recursive hierarchical relations that hold between nodes (such as elements in a syntactic tree).

The following examples show AQL queries on different layers:

[6] http://www.ling.uni-potsdam.de/acl-lab/Forsch/pcc/pcc.html

[7] http://www.wagsoft.com/RSTTool

[8] http://mmax2.sourceforge.net

[9] http://www.sfb632.uni-potsdam.de/annis/

1. a node that dominates another node via an RST *elaboration* relation (Fig. 1):

```
node & node & #1
>[relname="elaboration"] #2
```

2. a pointing (anaphoric) relation between two definite NPs (Fig. 2):

```
np_form="defnp" & np_form="defnp" & #1
->anaphor_antecedent #2
```

6. Summary and Outlook

We described the new release of the Potsdam Commentary Corpus. PCC 2.0 consists of 175 newspaper editorials annotated with syntax trees, completely-revised coreference links and rhetorical structure, plus a new layer of connectives and their arguments. Annotation guidelines have also been revised; they are accessible from the corpus website given above. The corpus is made availabe as source XML files, and readily-accessible in the web-based database ANNIS3. One additional format that we will be converting the data to is the ISO-standardized linguistic exchange format GrAF (Ide and Suderman, 2007).

The systematic multi-layer annotation will allow for tackling new research questions such as the correspondences between connectives and rhetorical structure, the patterns of information structure developing in a text (in relation to sentence syntax and discourse structure), and more.

We have been experimenting with other layers of discourse annotation, which so far have been applied to various subsets of the PCC (and to other texts): aspects of information structure, illocutionary roles, and argumentation structure. These will step-by-step be added to the full distribution of the corpus.

Acknowledgements

The work on PCC 2.0 was supported by Deutsche Forschungsgemeinschaft as part of the funding for project D1 in the Collaborative Research Center 632 'Information Structure' (Unversitt Potsdam, Humboldt Universitt Berlin, Freie Universitt Berlin)

7. References

Brants, S., Dipper, S., Eisenberg, P., Hansen, S., König, E., Lezius, W., Rohrer, C., Smith, G., and Uszkoreit, H. (2004). TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation*, 2(4):597–620.

Chiarcos, C., Dipper, S., Götze, M., Leser, U., Lüdeling, A., Ritz, J., and Stede, M. (2008). A flexible framework for integrating annotations from different tools and tag sets. *Traitement Automatique des Langues (TAL)*, 49(2).

Dipper, S. and Stede, M. (2006). Disambiguating potential connectives. In Butt, M., editor, *Proc. of KONVENS '06*, pages 167–173, Konstanz.

Hajičová, E., Hajič, J., Vidová-Hladká, B., Holub, M., Pajas, P., Kolářová-Řezníčková, V., and Sgall, P. (2001). The Current Status of the Prague Dependency Treebank. In Matoušek, V., Mautner, P., Mouček, R., and Taušer, K., editors, *Proceedings of the 5th International Conference on Text, Speech and Dialogue*, pages 11–20. Springer-Verlag Berlin Heidelberg New York.

Ide, N. and Suderman, K. (2007). GrAF: A graph-based format for linguistic annotations. In *Proc. of the Linguistic Annotation Workshop (LAW) at ACL-07*, Prague.

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

Krasavina, O., Chiarcos, C., and Zalmanov, D. (2007). Aspects of topicality in the use of demonstrative expressions in German, English and Russian. In *Proc. of the 6th Discourse Anaphora and Anaphor Resolution Colloquium DAARC-2007*, Lagos/Portugal.

Mann, W. C. and Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Müller, C. and Strube, M. (2006). Multi-level annotation of linguistic data with MMAX2. In Braun, S., Kohn, K., and Mukherjee, J., editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt.

O'Donnell, M. (2000). RSTTool 2.4 – a markup tool for Rhetorical Structure Theory. In *Proceedings of the International Natural Language Generation Conference*, pages 253–256, Mizpe Ramon/Israel.

Pasch, R., Brauße, U., Breindl, E., and Waßner, U. H. (2003). *Handbuch der deutschen Konnektoren*. Walter de Gruyter, Berlin/New York.

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). Rhetorical analysis with rich-feature support vector models. Master's thesis, University of Potsdam.

Stede, M. and Heintze, S. (2004). Machine-assisted rhetorical structure annotation. In *Proc. of the 20th International Conference on Computational Linguistics*, pages 425–431, Geneva.

Stede, M. (2002). DiMLex: A lexical approach to discourse markers. In Lenci, A. and Tomaso, V. D., editors, *Exploring the Lexicon - Theory and Computation*. Edizioni dell'Orso, Alessandria.

Stede, M. (2004). The Potsdam Commentary Corpus. In *Proceedings of the ACL Workshop on Discourse Annotation*, pages 96–102. Association for Computational Linguistics.

Stede, M. (2008a). Disambiguating rhetorical structure. *Research on Language and Computation*, 6(3):311–332.

Stede, M. (2008b). RST revisited: Disentangling nuclearity. In Fabricius-Hansen, C. and Ramm, W., editors, *'Subordination' versus 'coordination' in sentence and text*. John Benjamins, Amsterdam.

Zeldes, A., Ritz, J., Lüdeling, A., and Chiarcos, C. (2009). ANNIS: A search tool for multi-layer annotated corpora. In *Proceedings of Corpus Linguistics 2009*.

Zipser, F. and Romary, L. (2010). A model oriented ap-

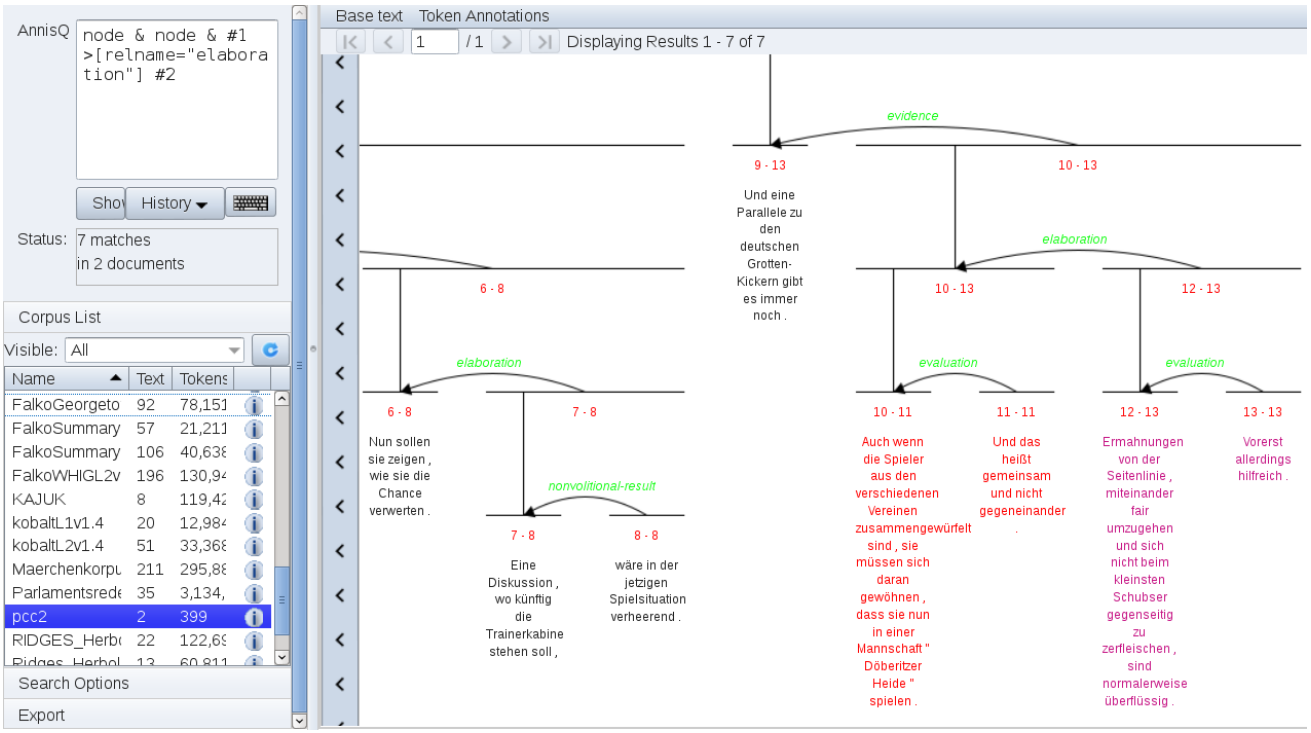

Figure 1: Querying PCC in ANNIS3 for an *elaboration* RST relation

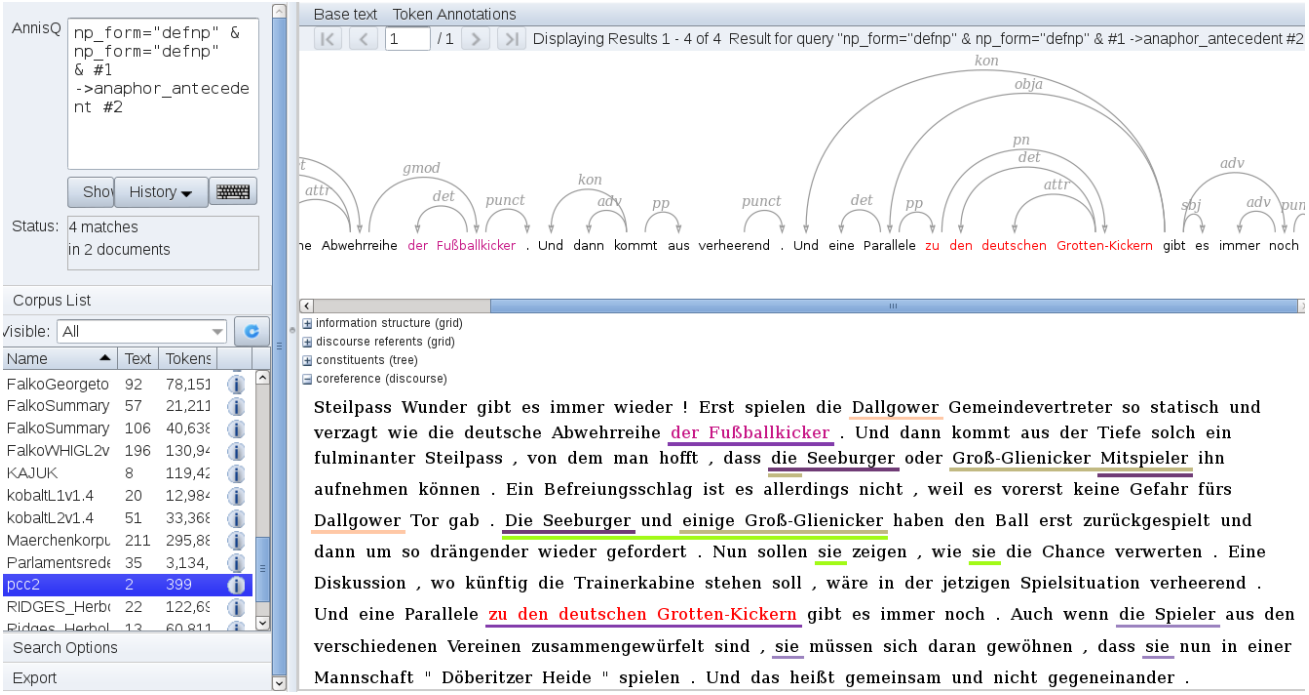

Figure 2: An anaphoric relationship between two definite NPs, indicated by color and underlining

proach to the mapping of annotation formats using standards. In *Proceedings of the Workshop on Language Resource and Language Technology Standards, LREC 2010*, pages 7–18, Malta.

Zipser, F., Zeldes, A., Ritz, J., Romary, L., and Leser, U. (2011). Pepper: Handling a multiverse of formats. In *33. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft*, Göttingen.