Importing MASC into the ANNIS linguistic database: A case study of mapping GrAF

Arne Neumann
EB Cognitive Science and SFB 632
University of Potsdam
neumana@uni-potsdam.de

Nancy Ide
Department of Computer Science
Vassar College
ide@cs.vassar.edu

Manfred Stede
EB Cognitive Science and SFB 632
University of Potsdam
stede@uni-potsdam.de

Abstract

This paper describes the importation of Manually Annotated Sub-Corpus (MASC) data and annotations into the linguistic database ANNIS, which allows users to visualize and query linguistically-annotated corpora. We outline the process of mapping MASC's GrAF representation to ANNIS's internal format relANNIS and demonstrate how the system provides access to multiple annotation layers in the corpus. This access provides information about inter-layer relations and dependencies that have been previously difficult to explore, and which are highly valuable for continued development of language processing applications.

1 Introduction

Over the past decade, corpora with multiple layers of linguistic annotation have been developed in order to extend the range of empirically-based linguistic research and enable study of inter-layer interactions. Recently created corpora include OntoNotes (Pradhan et al., 2007), the Groningen Meaning Bank (Basile et al., 2012), and the Manually Annotated Sub-Corpus (MASC)[1] (Ide et al., 2010). Typically, such corpora are represented in idiosyncratic in-house formats, and developers provide special software to access and query the annotations (for example, the OntoNotes "db tool" and Groningen's GMB Explorer). Access without the use of developer-supplied software often requires significant programming expertise, and as a result, it is not easy–or even possible–for others to add to or modify data and annotations in the resource.

This paper describes the importation of MASC data and annotations into the linguistic database ANNIS[2] (Chiarcos et al., 2008; Zeldes et al., 2009), which was designed to visualize and query linguistically-annotated corpora. Unlike most other corpora with multi-layer annotations, no special software has been developed for access to MASC. Instead, all MASC data and annotations are represented in GrAF (Ide and Suderman, 2007), the XML serialization of the abstract model for annotations defined by ISO TC37 SC4's Linguistic Annotation Framework (ISO/LAF) (Ide and Suderman, In press). GrAF is intended to serve as a generic "pivot" format that is isomorphic to annotation schemes conforming to the abstract model and therefore readily mappable to schemes used in available systems. We outline the process of mapping GrAF to ANNIS's internal format relANNIS and demonstrate how the system provides access to multiple annotation layers in MASC.

2 The ANNIS Infrastructure

The ANNIS system is a linguistic database geared toward the requirements of querying multi-layer annotated corpora, and providing various visualization means for layers with different structural properties. In particular, the annotation types supported are spans, DAGs with labelled edges, and pointing relations between terminals or non-terminals. For illustration, Figure 1 shows a screenshot where various parallel annotations of the same data are provided: dependency trees, constituent trees (here with "secondary edges" in dotted lines), and a grid view for annotations that assign labels to token spans. In addition, ANNIS offers a "discourse view" giving the complete text with coreference relations indicated by color and underlining. In the top of the screenshot, it can be noted that the system also stored video (and au-

[1] www.anc.org/MASC

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

Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse, pages 98–102,
Sofia, Bulgaria, August 8-9, 2013.

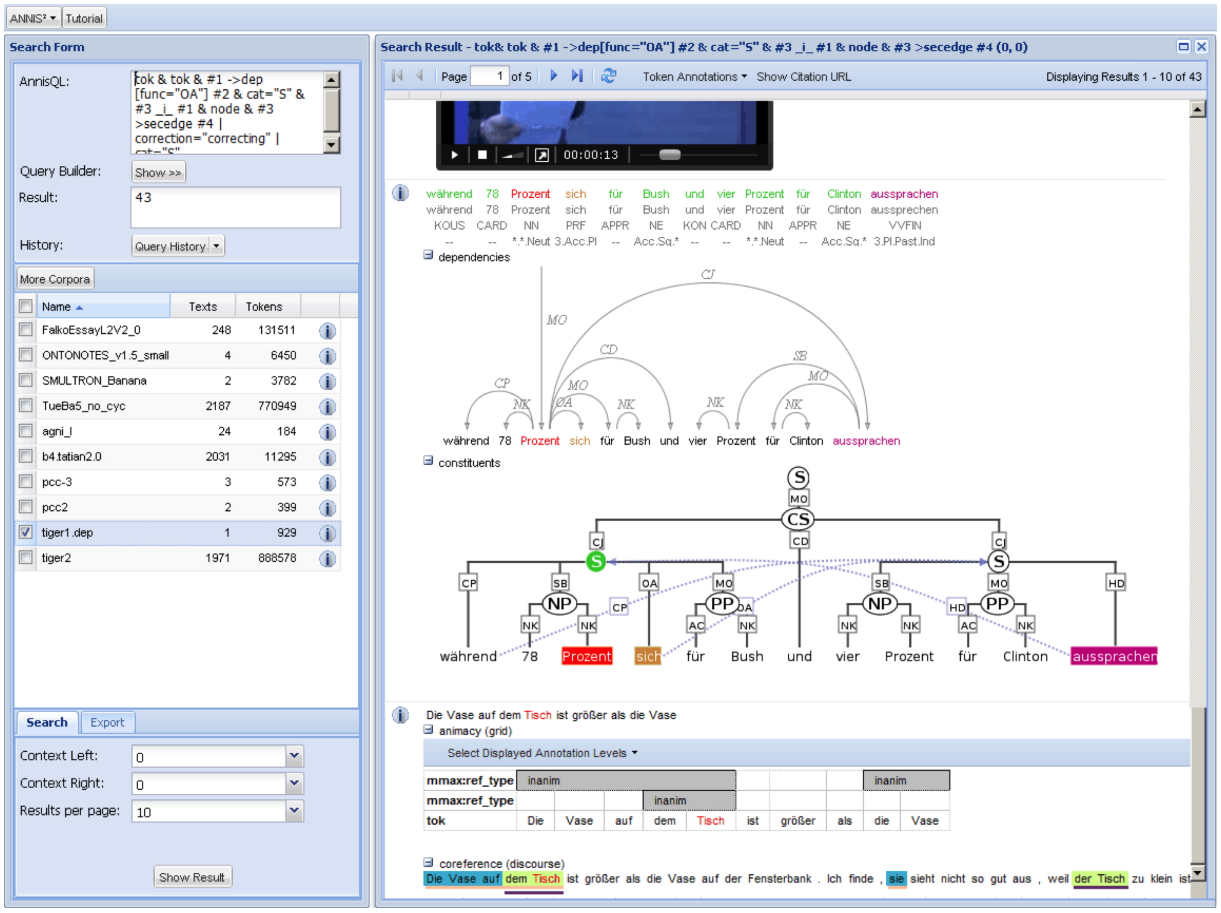

Figure 1: Screenshot of ANNIS2

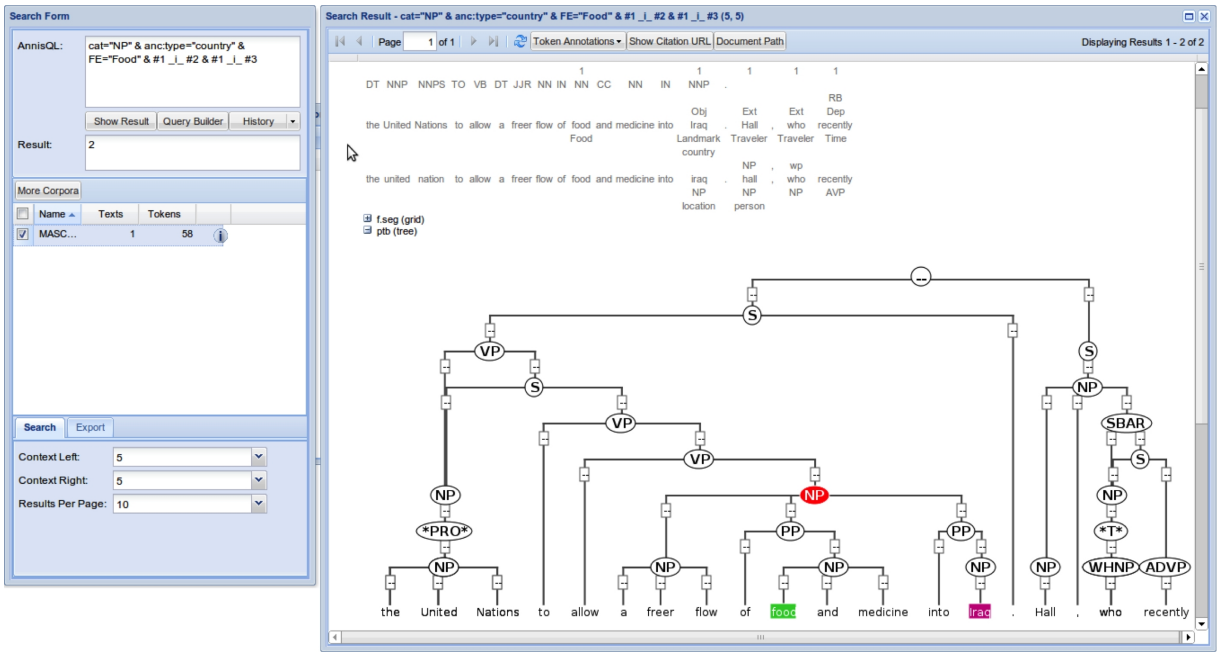

Figure 2: Querying MASC in ANNIS2 for an NP that includes both a *food* frame element and a *location* named entity

dio) data, but that aspect shall not concern us in this paper.

The system is web-based; the user interface is written in Java and ExtJS. The backend is PostgreSQL[3]. In general, all components are open source under the Apache License 2.0, and you can download ANNIS from the above-mentioned URL. We offer two versions: A server version, and the more lightweight "ANNIS kickstarter", which can be installed locally, e.g., on laptops.

ANNIS is complemented by SaltNPepper, a framework for converting annotations stemming from various popular annotation tools (MMAX, EXMARaLDA, annotate/Synpathy, RSTTool) – see Section 4.

3 MASC and GrAF

MASC is a fully open, half-million word corpus covering nineteen diverse genres of American English drawn from the Open American National Corpus (OANC)[4]. The corpus includes manually produced or hand-validated annotations for multiple linguistic layers, including morphosyntax (two different annotations), shallow parse (noun and verb chunks), Penn Treebank syntax, and named entities. Portions of the corpus are also annotated for FrameNet frames, opinion, PropBank predicate-arguments, and WordNet 3.1 word senses. Discourse-level annotation, including coreference, clauses, and discourse markers, will be available in fall, 2013.

Like the OANC, all MASC annotations are rendered in standoff form using GrAF, the graph-based format developed as a part of the ISO Linguistic Annotation Framework (ISO/LAF)(ISO 24612, 2012). GrAF is an XML serialization of the LAF abstract model for annotations, a formalization of models used across multiple applications for associating (linking) information, including not only directed-acyclic graphs (DAGs) but also ER diagrams, the Universal Modeling Language (UML), semantic and neural networks, RDF/OWL, and, more generally, hyperlinked data on the World Wide Web. The model is sufficiently general to represent any type of linguistic annotation; any serialization of the model can therefore serve as a *pivot* or intermediary among diverse annotation formats that conform to the abstract model. Thus, any sufficiently well-formed annotation scheme should be isomorphic to a GrAF representation of the same information. Problems arise only when a scheme does not specify information explicitly but rather embeds the interpretation in processing software rather than in the representation itself; for transduction to GrAF, this information must be made explicit in the representation.

Funding for MASC did not allow for extensive software development; the expectation is that by rendering the corpus in the ISO standard GrAF format, access could rely on GrAF-aware software developed by others, or transduction from GrAF to appropriate alternative formats would be trivial. We have already developed and deployed means to import linguistic data represented in GrAF into UIMA, GATE, and NLTK, and we provide transducers from GrAF to inline XML and the CoNLL IOB format.[5] Additionally, a GrAF-to-RDF transducer is near completion, which will enable inclusion of MASC in the Linguistic Linked Open Data (LLOD) cloud[6]. The incorporation of a GRAF transducer for ANNIS provides another example of the flexibility afforded via the GrAF representation.

4 Mapping GrAF to ANNIS via SaltNPepper

Pepper is a software framework that converts linguistic data among various formats, e.g. CoNLL, EXMARaLDA, PAULA, TigerXML, RSTTool and TreeTagger (Zipser et al., 2011). It is built upon the graph-based Salt meta model (Zipser and Romary, 2010), which is in turn based on the LAF abstract model for linguistic annotation. Mapping GrAF to Salt extends the range of formats into which annotations represented in GrAF can be automatically transduced to those to which Salt has been mapped, including ANNIS's relational database format relANNIS.

The following steps were taken to import the MASC corpus into ANNIS: first, the MASC corpus data was extracted with the GrAF API[7]. Second, a mapping between GrAF and Salt data structures was created. Most of the conversion is straightforward, since both models are graph-based. The only added processing is to provide

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

[4] www.anc.org/OANC

[5] Available from http://www.anc.org/MASC.

[6] http://linguistics.okfn.org/resources/llod/

[7] http://sourceforge.net/projects/iso-graf/

explicit edge labels in the Salt representation for ordered constiuents: in GrAF, directed edges from one to several other nodes by default represent sets of ordered constituents and need not be explicitly labeled as such, whereas in Salt, the role of all edges must be specified explicitly. Explicit labels in ANNIS are required in order to generate the appropriate visualizations automatically (e.g. trees for syntactic hierarchies and arc diagrams for syntactic dependencies).

Finally, the code was structured as a plug-in for Pepper and parameterized to make it usable for GrAF-formatted corpora other than MASC. It will be included in the next SaltNPepper release. The code is currently available from our software repository[8].

5 MASC in ANNIS: Examples

The ANNIS Query Language (AQL) allows users to search for specific token values and annotations as well as relationships between them, even across annotation level boundaries.[9] Token values are represented as text between quotes (e.g. `"men"`), while annotations are specified as attribute-value pairs (e.g. `pos="NN"`, a part-of-speech attribute with the value `NN`). A query for an annotation will return all elements with that annotation. Where necessary, namespaces[10] can be added to any element to disambiguate, e.g., `ptb:cat="NP"` signifies all annotation attribute-value pairs (attribute: `cat`, value: `NP`) that are in the `ptb` (Penn Treebank) namespace.

Relations among elements are specified by back-referencing incremental variable 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 that combine annotations from different layers:

1. A VP that dominates a PP which contains a named person at its right border:

```
cat="VP" & cat="PP" & NER="person" &
#1>#2 & #2_r_#3
```

2. a VP of passive form in past tense that includes a mention of a FrameNet frame element:

```
cat="VP" & voice="passive" &
tense="SimPas" & FE="Event" & #1_i_#2
& #1_i_#3 & #1_i_#4
```

Figure 2 shows the results of a search for an NP that includes both a named entity of the type *country* and a FrameNet frame element of the type *Food*:

```
cat="NP" & anc:type="country" &
FE="Food" & #1 _i_ #2 & #1 _i_ #3
```

6 Summary and Outlook

We explained the mapping of the MASC multilayer corpus to the ANNIS database by interpreting the GrAF format via the Pepper framework. Both MASC and ANNIS are freely available; a portion of MASC will also be added to the online demo version of ANNIS. We are also making the Pepper converter module for GrAF available.

Version 3 of ANNIS is currently under development[11]. Besides a new front-end and a REST-based API, it offers improved tokenization support (annotation on the level of subtokens; conflicting tokenizations) and handles dialogue corpora with simultaneous speakers as well as time-aligned audio/video data.

The ability to query across multiple annotation levels opens up significant new possibilities for exploring linguistically annotated data. Most commonly, language models are developed using information from at most one or two linguistic layers; ANNIS enables user to explore interdependencies that have been previously difficult to detect. By providing tools and data that are entirely free for use by the community, the ANNIS and MASC efforts contribute to the growing trend toward transparent sharing and openness of linguistic data and tools.

[8] `https://korpling.german.hu-berlin.de/svn/saltnpepper/PepperModules/GrAFModules/`

[9] Note that ANNIS does not allow searching for arbitrary strings from the primary data, but only for pre-identified segments such as tokens, named entities, etc.

[10] A namespace groups one or more types of annotation into a logical unit, e.g all annotations produced by a specific tool or project.

[11] Early development releases can be found at `http://www.sfb632.uni-potsdam.de/annis/annis3.html`

Acknowledgments

MASC and GrAF development was supported by US NSF award CRI-0708952. The work of A.N. and M.S. was supported by Deutsche Forschungsgemeinschaft as part of the Collaborative Research Center "Information Structure" (SFB 632) at Univ. Potsdam and HU Berlin.

References

Valerio Basile, Johan Bos, Kilian Evang, and Noortje Venhuizen. 2012. Developing a large semantically annotated corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3196–3200, Istanbul, Turkey.

Christian Chiarcos, Stefanie Dipper, Michael Götze, Ulf Leser, Anke Lüdeling, Julia Ritz, and Manfred Stede. 2008. A flexible framework for integrating annotations from different tools and tag sets. *Traitement Automatique des Langues (TAL)*, 49(2).

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

Nancy Ide and Keith Suderman. In press. The Linguistic Annotation Framework: A Standard for Annotation Interchange and Merging. *Language Resources and Evaluation*.

Nancy Ide, Collin Baker, Christiane Fellbaum, and Rebecca Passonneau. 2010. The Manually Annotated Sub-Corpus : A community resource for and by the people. In *Proceedings of the The 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden.

ISO 24612. 2012. *Language Resource Management – Linguistic Annotation Framework*. International Standards Organization, Geneva, Switzerland.

Sameer S. Pradhan, Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2007. OntoNotes: A unified relational semantic representation. In *ICSC '07: Proceedings of the International Conference on Semantic Computing*, pages 517–526, Washington, DC, USA. IEEE Computer Society.

Amir Zeldes, Julia Ritz, Anke Lüdeling, and Christian Chiarcos. 2009. ANNIS: A search tool for multi-layer annotated corpora. In *Proceedings of Corpus Linguistics 2009*.

Florian Zipser and Laurent Romary. 2010. A model oriented approach 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.

Florian Zipser, Amir Zeldes, Julia Ritz, Laurent Romary, and Ulf Leser. 2011. Pepper: Handling a multiverse of formats. In *33. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft*, Göttingen.