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

Arne Neumann\textsuperscript{1}  Nancy Ide\textsuperscript{2}  Manfred Stede\textsuperscript{1}

\textsuperscript{1}EB Cognitive Science and SFB 632
University of Potsdam

\textsuperscript{2}Department of Computer Science
Vassar College

August 8, 2013
GrAF

- XML serialization of the Linguistic Annotation Framework (ISO/LAF) (Ide and Suderman (2007; In press))
- intended as a generic “pivot” format

ANNIS

open source, web-based search and visualization tool for corpora with diverse types of annotation (Chiarcos et al. 2008; Zeldes et al. 2009)
Introduction

GrAF
- XML serialization of the Linguistic Annotation Framework (ISO/LAF) (Ide and Suderman (2007; In press))
- intended as a generic “pivot” format

ANNIS
open source, web-based search and visualization tool for corpora with diverse types of annotation (Chiarcos et al. 2008; Zeldes et al. 2009)
Introduction

Introduction II

Goals

- use GrAF-formatted corpora in ANNIS
- simplify empirical research of inter-layer interactions in corpora

Manually Annotated Sub-Corpus (MASC)

fully open, 500k word subcorpus of the Open American National Corpus (OANC) annotated on multiple levels
Introduction

Goals

- use GrAF-formatted corpora in ANNIS
- simplify empirical research of inter-layer interactions in corpora

Manually Annotated Sub-Corpus (MASC)

fully open, 500k word subcorpus of the Open American National Corpus (OANC) annotated on multiple levels
ANNIS supports:

- Visualization of multilayered annotations: syntax trees (constituents / dependencies), morphology, RST, coreference, information structure
- Multimodal data: text, audio, video
- Conflicting data: overlapping speakers in dialogues, multiple tokenizations / segmentations
ANNIS

supports:

- visualization of multilayered annotations: syntax trees (constituents / dependencies), morphology, RST, coreference, information structure
- multimodal data: text, audio, video
- conflicting data: overlapping speakers in dialogues, multiple tokenizations / segmentations
ANNIS supports:

- visualization of multilayered annotations: syntax trees (constituents / dependencies), morphology, RST, coreference, information structure
- multimodal data: text, audio, video
- conflicting data: overlapping speakers in dialogues, multiple tokenizations / segmentations
supports:

- visualization of multilayered annotations: syntax trees (constituents / dependencies), morphology, RST, coreference, information structure
- multimodal data: text, audio, video
- conflicting data: overlapping speakers in dialogues, multiple tokenizations / segmentations
MASC

Manually Annotated Sub-Corpus (MASC)

fully open, 500k word subcorpus of the Open American National Corpus (OANC)

- manually produced or hand-validated annotations: POS, shallow parse, Penn Treebank syntax, named entities, coreference...
- partially annotated for FrameNet frames, opinion, PropBank predicate-arguments, WordNet word senses, ...
- format: GrAF
Manually Annotated Sub-Corpus (MASC)

fully open, 500k word subcorpus of the Open American National Corpus (OANC)

- manually produced or hand-validated annotations: POS, shallow parse, Penn Treebank syntax, named entities, coreference...
- partially annotated for FrameNet frames, opinion, PropBank predicate-arguments, WordNet word senses, ...
- format: GrAF
MASC

Manually Annotated Sub-Corpus (MASC)

fully open, 500k word subcorpus of the Open American National Corpus (OANC)

- manually produced or hand-validated annotations: POS, shallow parse, Penn Treebank syntax, named entities, coreference...
- partially annotated for FrameNet frames, opinion, PropBank predicate-arguments, WordNet word senses, ...
- format: GrAF
GrAF

graph-based, standoff XML format developed as a part of the ISO Linguistic Annotation Framework (ISO/LAF; ISO 24612, 2012)

LAF abstract model

- formalization of models used for associating information
  - e.g. directed-acyclic graphs, UML, ER diagrams, semantic networks, RDF…
- general enough to represent any type of linguistic annotation
GrAF

graph-based, standoff XML format developed as a part of the ISO Linguistic Annotation Framework (ISO/LAF; ISO 24612, 2012)

LAF abstract model

- formalization of models used for associating information
  - e.g. directed-acyclic graphs, UML, ER diagrams, semantic networks, RDF...
- general enough to represent any type of linguistic annotation
MASC and GrAF

**LAF/GrAF**

GrAF

- graph-based, standoff XML format developed as a part of the ISO Linguistic Annotation Framework (ISO/LAF; ISO 24612, 2012)

**LAF abstract model**

- formalization of models used for associating information
  - e.g. directed-acyclic graphs, UML, ER diagrams, semantic networks, RDF...
- general enough to represent any type of linguistic annotation
GrAF as an intermediary

available converters

- importers/exporters for UIMA, GATE
- NLTK importer
- exporters for CoNLL IOB format, inline XML
- upcoming: GrAF-to-RDF → Linguistic Linked Open Data (LLOD)
Mapping GrAF to ANNIS via Salt

Pepper
- framework that converts linguistic data among various formats, e.g. CoNLL, EXMARaLDA, PAULA, TigerXML, RSTTool, MMAX, TreeTagger and relANNIS (Zipser et al. 2011)
- uses Salt as an intermediary

Salt
- graph-based meta model (Zipser and Romary 2010) based on LAF
- advantage: write one converter (GrAF to Salt), use all existing exporters
Mapping GrAF to ANNIS via Salt

Pepper
- framework that converts linguistic data among various formats, e.g. CoNLL, EXMARaLDA, PAULA, TigerXML, RSTTool, MMAX, TreeTagger and relANNIS (Zipser et al. 2011)
- uses Salt as an intermediary

Salt
- graph-based meta model (Zipser and Romary 2010) based on LAF
- advantage: write one converter (GrAF to Salt), use all existing exporters
Conversion steps

- extract MASC corpus data using the Java-based GrAF API\(^1\)
- mapping graphs is straightforward
  - **exception**: explicit edge labels needed in Salt (e.g. for ordering constituents)

The GrAF-to-Salt/relANNIS will be included in the next SaltNPepper release, but source code is already available\(^2\).

---

\(^1\) [http://sourceforge.net/projects/iso-graf/](http://sourceforge.net/projects/iso-graf/)

\(^2\) [https://korpling.german.hu-berlin.de/svn/saltnpepper/PepperModules/GrAFModules/](https://korpling.german.hu-berlin.de/svn/saltnpepper/PepperModules/GrAFModules/)
Conversion steps

- extract MASC corpus data using the Java-based GrAF API\(^1\)
- mapping graphs is straightforward
  - exception: explicit edge labels needed in Salt (e.g. for ordering constituents)

The GrAF-to-Salt/relANNIS will be included in the next SaltNPepper release, but source code is already available\(^2\).

---

\(^1\)http://sourceforge.net/projects/iso-graf/
\(^2\)https://korpling.german.hu-berlin.de/svn/saltnpepper/PepperModules/GrAFModules/
Conversion steps

- extract MASC corpus data using the Java-based GrAF API\(^1\)
- mapping graphs is straightforward
  - exception: explicit edge labels needed in Salt (e.g. for ordering constituents)

The GrAF-to-Salt/relANNIS will be included in the next SaltNPepper release, but source code is already available\(^2\).

---

\(^1\)http://sourceforge.net/projects/iso-graf/
\(^2\)https://korpling.german.hu-berlin.de/svn/saltnpepper/PepperModules/GrAFModules/
Querying MASC in ANNIS

ANNIS Query Language (AQL)

allows to search for specific token values and annotations as well as relationships between them, even across annotation level boundaries
Queries across multiple annotation levels

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

Chiarcos, Christian et al. (2008). “A Flexible Framework for Integrating Annotations from Different Tools and Tag Sets”. In: Traitement Automatique des Langues (TAL) 49.2.

Ide, Nancy and Keith Suderman (2007). “GrAF: A Graph-based Format for Linguistic Annotations”. In: Proceedings of the First Linguistic Annotation Workshop. Prague, pp. 1–8.

— (In press). “The Linguistic Annotation Framework: A Standard for Annotation Interchange and Merging”. In: Language Resources and Evaluation.

ISO 24612 (2012). Language Resource Management – Linguistic Annotation Framework. International Standards Organization, Geneva, Switzerland.
Zeldes, Amir et al. (2009). “ANNIS: A Search Tool for Multi-Layer Annotated Corpora”. In: *Proceedings of Corpus Linguistics 2009*. Liverpool, UK.

Zipser, Florian 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*. Malta, pp. 7–18.

Zipser, Florian et al. (2011). “Pepper: Handling a Multiverse of Formats”. In: *33. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft*. Göttingen.