
A Web Service for Customized Corpus Delivery

Nancy Ide

Department of Computer Science

Vassar College

USA

Web services

- Language processing tasks are likely to be handled by web services in the future
 - Applications such as machine translation
 - Low- and mid-level NLP tasks that can be pipelined
 - Tokenization
 - POS labeling
 - Shallow syntax
 - Etc.
 - Access to resources that is otherwise undesirable or prohibited (e.g. retrieval of information from lexicons)

Context

- Open American National Corpus
 - 15 million word corpus of contemporary American English
 - Range of genres comparable to those in the British National Corpus (BNC)
 - Also newer genres such as blogs, email, rap lyrics, tweets, etc.
 - Automatically annotated for a variety of linguistic phenomena
 - All annotations available as standoff files

Context

- Manually Annotated Sub-Corpus (MASC)
 - Ide et al., 2008, LREC
 - Balanced subset of half a million words of written texts and transcribed speech
 - drawn from the OANC
 - Hand-validated or manually-produced annotations for a wide variety of linguistic phenomena
 - Also a corpus of 10,000 sentences manually annotated for WordNet senses (WN 3.1)

Context

- All data and annotations freely available for any use
- Downloadable from ANC web site
<http://www.anc.org>

Context

- Both OANC and MASC
 - many annotations of different types
 - several different annotations of the same type
 - E.g. 3-4 different tokenizations, POS annotations
 - Many different genres
- **Many users want only part of data, certain annotations**
- **Desirable to provide in a variety of formats**

MASC

- RIGHT NOW (well, Friday)
 - validated or manually produced annotations for a wide variety of linguistic phenomena at several linguistic levels available for 82K words
 - 1000 occurrences of each of 50 words manually sense-tagged with Wordnet senses
- MASC contains or shortly will contain sixteen different types of annotation
 - most produced at different sites using specialized annotation software
 - FrameNet annotation of 500 sentences from the WordNet corpus

MASC Annotations (first 220K words)

Annotation type	Method	No. texts	No. words
Token	Validated	119	220469
Sentence	Validated	119	220469
POS/lemma	Validated	119	220469
Noun chunks	Validated	119	220469
Verb chunks	Validated	119	220469
Named entities	Validated	119	220469
FrameNet frames	Manual	20	16566
WordNet senses	Manual	n/a	n/a
HSPG	Validated	90	64000
Discourse	Manual	40	30106
Penn Treebank	Validated	63	78586
PropBank	Validated	58	41578
Opinion	Manual	72	41578
Committed belief	Manual	13	2191
Event	Manual	13	2191
Coreference	Manual	2	1877

MASC Composition (first 220K words)

Genre	No. texts	Total words
Email	2	468
Essay	4	17516
Fiction	4	20413
Gov't documents	1	6064
Journal	10	25635
Letters	31	10518
Newspaper/newswire	42	17324
Non-fiction	4	17118
Spoken	11	27824
Debate transcript	2	32325
Court proceedings	1	20187
Technical	3	15417
Travel guides	4	12463
Total	119	220469

MASC Format

- The layering of annotations over MASC texts dictates the use of a stand-off annotation representation format
 - each annotation contained in a separate document linked to the primary data
- All annotations represented using the Graph Annotation Format (GrAF)
 - XML serialization of abstract model : directed graph decorated with feature structures providing the annotation content
 - Enables merging annotations originally represented in different
 - Enables transducing annotations to a variety of other formats

MASC File Structure

- Each text provided in a separate file in UTF- 8 encoding
 - contains no annotation or markup of any kind
- Each text associated with a set of stand- off files, one for each annotation type
 - Some or all of annotation types in previous slide
 - Also, minimal segmentation and logical structure (titles, headings, paragraphs, etc.)

WordNet/FrameNet Annotation

- Annotation of 1000 occurrences of 100 words selected by WordNet and FrameNet teams
 - Provide corpus evidence for an effort to harmonize sense distinctions in WordNet and FrameNet
 - Use WordNet 3.1 senses
 - WN database modified based on tagging experience in MASC
 - WN3.1 will be released in the future
- 100 of the 1000 sentences for each word also annotated for FrameNet frame elements

Sentence Corpus

- Annotated sentences provided as a stand-alone corpus
- WordNet and FrameNet annotations represented in standoff files
- Each sentence linked to its occurrence in the original text in the MASC or OANC

ANC Tool

- Few software systems handle stand-off annotation
 - or require lots of computational expertise
- To solve the problem, ANC project developed the “ANC Tool”
 - Merges data and annotations in GrAF
 - Generates output in various formats
- **Required downloading and installing the ANC Tool in order to use**

ANC Tool

- Built on XCES parser
 - SAX- like parser that combines selected annotations with primary data
 - Uses multiple implementations of the `org.xml.sax.DocumentHandler` interface
- Freely available from the ANC website
- Can be used by any application that allows the user to specify the SAX parser to be used
 - e.g., Saxon can be used to apply XSLT stylesheets to GrAF annotations without first merging annotations and primary data
- Provides several options for representing overlapping hierarchies.

ANC2Go

- Web application
 - will be implemented as a RESTful web service this summer
- Users choose data and annotations and desired output format
- Send request via web interface
- ANC2Go returns a URL from which the user can download the requested corpus
- **Users create a “personalized” corpus**
 - data and annotations of their choosing
 - format most useful to them

Additional GrAF Tools

- Modules to use GrAF annotations in general-purpose annotation and analysis tools
 - GATE
 - UIMA
 - NLTK
- Can use these systems independently or interchangeably
- Java API for GrAF
 - includes a GraphVizRenderer to generate visualizations of an annotation subgraph

ANC2Go

- Can be applied to
 - MASC data and annotations
 - Sentence Corpus (WV/FN)
 - OANC
- **MASC user need never deal directly with or see the underlying representation of the corpus and stand-off annotations**
 - But gains all the advantages that representation offers

Output formats

- XML in-line
 - Suitable for use with the BNCs XAIRA search and access interface
 - Input to any XML-aware software
- Token with part of speech tags, separated by character of the user's choice
 - input to general-purpose concordance software including MonoConc and Word-Smith
- Token/part of speech input for the Natural Language Toolkit (NLTK)
- CONLL IOB format
 - used in the Conference on Natural Language Learning shared tasks

Email Address

Select Corpus OANC MASC Wordnet

Browse directory

Copy directory structure

Selected Directories

/spoken/face-to-face
/written_1/journal/slate

XML **MonoConc Pro** **WordSmith** **NLTK** **CoNLL**

Annotations

Tokenization and Morphosyntax

- GATE/PENN lemma and part of speech
- Penn TreeBank part of speech
- FrameNet part of speech
- None

- Logical markup
- Sentence boundaries
- FrameNet
- Penn TreeBank
- WordNet
- Named Entities
- Noun chunks
- Verb chunks

Overlap handling

- Discard
- Milestone
- Nest

Email Address

suderman@anc.org

Select Corpus

OANC

MASC

Wordnet

Browse directory

Browse...

Copy directory structure

Selected Directories

/spoken/face-to-face
/written_1/journal/slate

XML

MonoConc Pro

WordSmith

NLTK

CoNLL

Annotations

Tokenization and Morphosyntax

- GATE/PENN lemma and part of speech
- Penn TreeBank part of speech
- FrameNet part of speech
- None

- Logical markup
- Sentence boundaries
- FrameNet
- Penn TreeBank
- WordNet
- Named Entities
- Noun chunks
- Verb chunks

Process

Email Address

suderman@anc.org

Select Corpus

OANC

MASC

Wordnet

Browse directory

Browse...

Copy directory structure

Selected Directories

/written_2/travel_guides

/written_2/technical/911report

XML

MonoConc Pro

WordSmith

NLTK

CoNLL

Tokenization and Morphosyntax

GATE/PENN lemma and part of speech

Penn TreeBank part of speech

FrameNet part of speech

None

Separator character

_

Process

Using ANC2Go

- Choose among 3 corpus options (OANC, MASC, WordNet sense corpus)
- Choose entire corpus and annotations, or browse corpus file hierarchy and limit the texts to be included
- When a corpus is chosen, list of available annotations dynamically updated to include only those annotations available for that corpus
- Annotation options may be further limited when the user chooses an output format
 - For example, MonoConc and Wordsmith formats available for token/POS only

Using ANC2Go

- ANC2Go generates only token/POS for NLTK, although NLTK processes other annotation types.
 - NLTK corpus reader for other annotation types
- If XML is the chosen output format, the user is provided with a set of alternative means to handle overlapping hierarchies

Using ANC2Go

- ANC2Go consults the corpus and text headers to determine dependencies among annotations
 - Occurs when annotation references another annotation rather than the primary data itself
- ANC2Go automatically includes required annotations
 - For example, the Penn Treebank syntactic annotations reference Penn Treebank token/POS annotations
 - also referenced by other annotations in MASC
 - When user chooses Penn Treebank annotations ANC2Go automatically includes the Treebank token annotations

Download

- <http://www.anc.org>
 - OANC data (15 million words) and annotations freely downloadable
 - NB Still in a previous version of GrAF—will be updated this summer
- <http://www.anc.org/MASC>
 - MASC data (220K words) and annotations (for 82K words) freely downloadable
 - Uses the new and final ISO version of GrAF
- <http://www.anc.org:8080/ANC2Go>
 - ANC2Go web application