# ANC2Go:
# A Web Application for Customized Corpus Creation

**Nancy Ide, Keith Suderman, Brian Simms**

Department of Computer Science, Vassar College
Poughkeepsie, New York 12604 USA
{ide, suderman, brsimms}@cs.vassar.edu

### Abstract

We describe a web application called "ANC2Go" that enables the user to select data from the Open American National Corpus (OANC) and the Manually Annotated Sub-corpus (MASC) together with some or all of the annotations available. The user also may select from among a variety of options for output format, or may receive the selected portions of the corpus and annotations in their original GrAF XML standoff format.. The request is processed by merging the annotations selected and rendering them in the desired output format, then bundling the results and making it available for download. Thus, users can create a customized corpus with data and annotations of their choosing, delivered in the format that is most convenient for their use. ANC2Go will be released as a web service in the near future. Both the OANC and MASC are freely available for any use from the American National Corpus website and may be accessed through the ANC2Go application, or they may downloaded in their entirety.

## 1. Introduction

The Open American National Corpus (OANC)[1] is a 15 million word corpus of contemporary American English covering a range of genres comparable to those in the British National Corpus (BNC) together with newer genres such as blogs, email, rap lyrics, etc. The entire OANC is automatically annotated for a variety of linguistic phenomena, and all annotations are available as standoff files. The Manually Annotated Sub-Corpus (MASC) (Ide et al., 2008) is a balanced subset of half a million words of written texts and transcribed speech drawn from the OANC with validated or manually-produced annotations for a wide variety of linguistic phenomena. The corpus is intended to serve as the base for a community-wide annotation effort and provide an infrastructure that enables the incorporation of contributed annotations into a single, usable format that can then be analyzed as it is or transduced to any of a variety of other formats. Like the OANC, the MASC project is committed to a fully open model of distribution for all data and annotations, whether produced by the project or contributed by other annotation projects.

Both the OANC and MASC are freely available from the American National Corpus website[2] and through the Linguistic Data Consortium (LDC)[3]. In addition to enabling download of these corpora as a whole, we provide a web application that allows users to select some or all parts of the corpus and choose among the available annotations. The application merges the chosen annotations and produces them in any of several output formats of the user's choosing. The entire bundle is then made available to the user for download.

This paper describes the "ANC2Go" corpus creation service. By way of illustration we first describe the MASC data and annotations, which have recently been made available on the ANC website.

## 2. MASC

The Manually Annotated Sub-Corpus (MASC) (Ide et al., 2008) is a balanced subset of half a million words of written texts and transcribed speech drawn from the OANC. To date, validated or manually produced annotations for a wide variety of linguistic phenomena at several linguistic levels have been made available for about half of MASC's 500K words. In addition, 1000 occurrences of each of 100 words have been manually sense-tagged with Wordnet senses, and of these, 100 sentences have been annotated for FrameNet frames. At present, MASC contains sixteen different types of annotation, most produced at different sites using specialized annotation software. The MASC annotations are shown in Table 1. The contents of the 220K of the corpus for which annotations have been provided so far are shown in Table 2. All annotations, whether contributed or produced in-house, are transduced to the Graph Annotation Format (GrAF) (Ide and Suderman, 2007) defined by ISO TC37 SC4's Linguistic Annotation Framework (LAF) (Ide and Romary, 2004), which is the format in which they are distributed in both MASC and the OANC.

The layering of annotations over MASC texts dictates the use of a stand-off annotation representation format, where each annotation is contained in a separate document linked to the primary data. All annotations, whether contributed or produced in-house, are transduced to the Graph Annotation Format (GrAF) (Ide and Suderman, 2007) defined by ISO TC37 SC4's Linguistic Annotation Framework (LAF) (Ide and Romary, 2004). GrAF is an XML serialization of the LAF abstract model of annotations, which consists of a directed graph decorated with feature structures providing the annotation content. GrAF is intended to be used as a "pivot" format for annotations that may have been originally represented in different formats that will allow merging or direct comparison. The GrAF representation also enables transducing annotations to a variety of other formats, as described below in section 3..

MASC is distributed with a corpus header compliant with

[1] http://www.anc.org

[2] http://www.anc.org

[3] http://www.ldc.upenn.edu

| 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 |

Table 1: Current MASC Annotations

| 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 |

Table 2: MASC Composition (first 220K)

the LAF specifications, which provides information about the corpus and its source, as well as machine-processable information concerning metadata, annotation types, media types, defined annotation layers/tiers, and the corpus file structure. Each text in the corpus is also associated with a header document that provides appropriate metadata together with machine-processable information about associated annotations. This information is accessed by ANC2Go in order to dynamically configure its interface, as described in section 3..

Each text in the corpus is provided in a separate file in UTF-8 encoding, which contains no annotation or markup of any kind. Each of the text files is associated with a set of standoff files, one for each annotation type, containing the annotations for that text in GrAF format. In addition to the annotation types listed in Table 1, documents containing annotation for minimal segmentation and logical structure (titles, headings, paragraphs, etc. down to the level of paragraph) are included.

Standoff annotations may reference regions in the text directly, by indexing the start and end position of the text segment to which they apply. Indexes are conformant to the LAF specification, which requires them to refer to virtual nodes between each character in the text file, starting at 0. Some annotations, including shallow parse (noun chunks and verb chunks), Penn Treebank syntax, and FrameNet frame elements, reference tokens rather than the text itself. Token boundaries referencing minimal segmentation units in the segmentation file are provided in a separate document. MASC includes three different tokenization files that accommodate variations used by different higher-level annotations produced at different sites: Penn Treebank, FrameNet, and tokenizations produced by the GATE system.

The MASC project is also annotating 1000 occurrences of 100 words selected by the WordNet and FrameNet teams with WordNet senses to provide corpus evidence for an effort to harmonize sense distinctions in WordNet and FrameNet (Baker and Fellbaum, 2009). One hundred of the 1000 sentences for each word are also annotated for FrameNet frame elements.[4] Because of its small size, MASC typically contain only a handful of the 1000 occurrences of a given word; the remaining occurrences are drawn from the 15 million words of the OANC. For convenience, the annotated sentences are provided as a standalone corpus, with the WordNet and FrameNet annotations represented in standoff files. Each sentence in this corpus is linked to its occurrence in the original text in the MASC or OANC, so that the context and other annotations associated with the sentence may be accessed.

## 3. ANC2Go

At present few software systems handle stand-off annotation, and those that do often demand computational expertise beyond what many MASC users have access to. To solve this problem, the ANC project developed the "ANC Tool" for merging data and its annotations in GrAF and generating output in various formats. To use the ANC Tool, it was necessary to download and install the ANC Tool first, which was often problematic, especially for non-programmers. However, the ANC project now provides a web application named "ANC2Go"[5], which allows users to choose data and annotations together with a desired output format, and send the request via a web interface. The application later returns a URL from which the user can download the requested corpus. Thus, ANC2Go allows users to "create" a personalized corpus consisting of data and annotations of their choosing, in the format most useful to them. ANC2Go can be applied to MASC data and annotations as well as all or portions of the OANC. As a result, the MASC user need never deal directly with or see the underlying representation of the corpus and its stand-off annotations, but gains all the advantages that representation offers.

ANC2Go is based on the ANC Tool, which is in turn built on the XCES parser. The XCES parser is a SAX-like parser that combines selected annotations with primary

[4] Note that several MASC texts are fully annotated for FrameNet frame elements, in addition to the WordNet-tagged sentences.

[5] http://www.anc.org/ANC2Go

data. The ANC Tool uses multiple implementations of the org.xml.sax.DocumentHandler interface, one for each output format, which the XCES parser uses to generate the desired output. Additional output formats can be easily generated by implementing additional interfaces. The parser, which is freely available from the ANC website, can also 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 MASC annotations without first merging annotations and primary data. The parser also provides several options for representing overlapping hierarchies.

The ANC tool currently generates the user's choice of texts and annotations in any of the following output formats[6]:

1. XML in-line, suitable for use with the BNCs XAIRA search and access interface and other XML-aware software;

2. tokens with part of speech tags appended to each word and separated by an character of the user's choice, suitable for input to general-purpose concordance software including MonoConc[7] and WordSmith[8];

3. token / part of speech input for the Natural Language Toolkit (NLTK)[9];

4. CONLL IOB format, used in the Conference on Natural Language Learning shared tasks.

Figures 1, 2, and 3 show the user interface for ANC2Go. As the figures show, after entering the email address to which the link to the generated output will be sent, the user chooses among one three possible corpus options: the OANC, MASC, and the corpus of WordNet sense annotations, each of which has its particular set of annotations. Once the corpus is identified, the user can choose to process the entire corpus, including all available annotations, or he can browse the corpus file hierarchy and limit the texts to be included to one or more sub-directories of the corpus. In the latter case the chosen sub-directories are displayed.

When a corpus is chosen, the list of available annotations in the lower pane of the interface is dynamically updated to include only those annotations available for that corpus. Annotation options may be further limited when the user chooses an output format, as shown in Figure 2, typically because of limitations of the processing software. For example, MonoConc and Wordsmith formats are available for token/POS only, since this is the only information used in these systems. Note that ANC2Go also generates only token/POS for NLTK although it processes other annotation types. This is a matter of convenience for the simple case where minimal annotation is needed as input to NLTK; for other GrAF annotations, we provide an NLTK corpus reader that is used from within the NLTK system to load data and annotations. When XML is the chosen output format, the user is provided with a set of alternative means to handle overlapping hierarchies among different annotation types, as shown in Figure 1.

The OANC and MASC each include two or more alternative annotations for tokenization and part of speech. Currently, only the XML and CONLL formats allow including more than one of these alternatives.

When generating the output, ANC2Go consults the corpus and text headers to determine dependencies that dictate inclusion of additional annotations that may not have been chosen by the user. Dependencies occur when an annotation references another annotation rather the primary data itself. For example, the Penn Treebank syntactic annotations do not point directly into the primary data, but rather reference the Penn Treebank token/POS annotations (which in MASC are also referenced by other annotations). When the user chooses to include Penn Treebank annotations in the customized corpus, as in Figure 1, ANC2Go automatically includes the Treebank token annotations as well.

ANC2Go will be implemented as a RESTful web service in the near future.

In addition to generating annotations in the output formats supported by the ANC Tool, the ANC project provides modules that can be used within the general-purpose annotation and analysis tools GATE[10] and UIMA[11], which enable GrAF annotation files to be read directly into and used by either of these systems. We also provide a GATE module that outputs annotations produced within that system into GrAF.[12] This allows not only using these systems independently with GrAF annotations, but also using them interchangeably (see (Ide and Suderman, 2009)).

The ANC Tool generates NLTK-compliant input for part of speech annotation only. For other GrAF annotations, we provide an NLTK corpus reader. The ANC Project also provides a Java API for GrAF that includes a *GraphVizRenderer* which can be used to generate input to the GraphViz graph visualization application[13] for any subgraph (starting from a specified node) in a GrAF file. This enables users to display a graphic rendering of all annotations for a given sentence, paragraph, etc.

## 4. Conclusion

ANC2Go allows users to design a customized corpus, including annotations of their choice and in a format of their choice, that they then freely download from the ANC website. The ANC2Go web application, together with additional tools to use GrAF encoded data and annotations with various general-purpose NLP systems, makes MASC and the OANC the most interoperable and usable corpus in existence. The entire configuration, which relies on freely available data and annotations, the LAF concept of a general purpose pivot format for representing annotations and

[6] A handler to produce MASC annotations in RDF/OWL format is forthcoming.

[7] http://www.athel.com/mono.html

[8] http://www.lexically.net/wordsmith/index.html

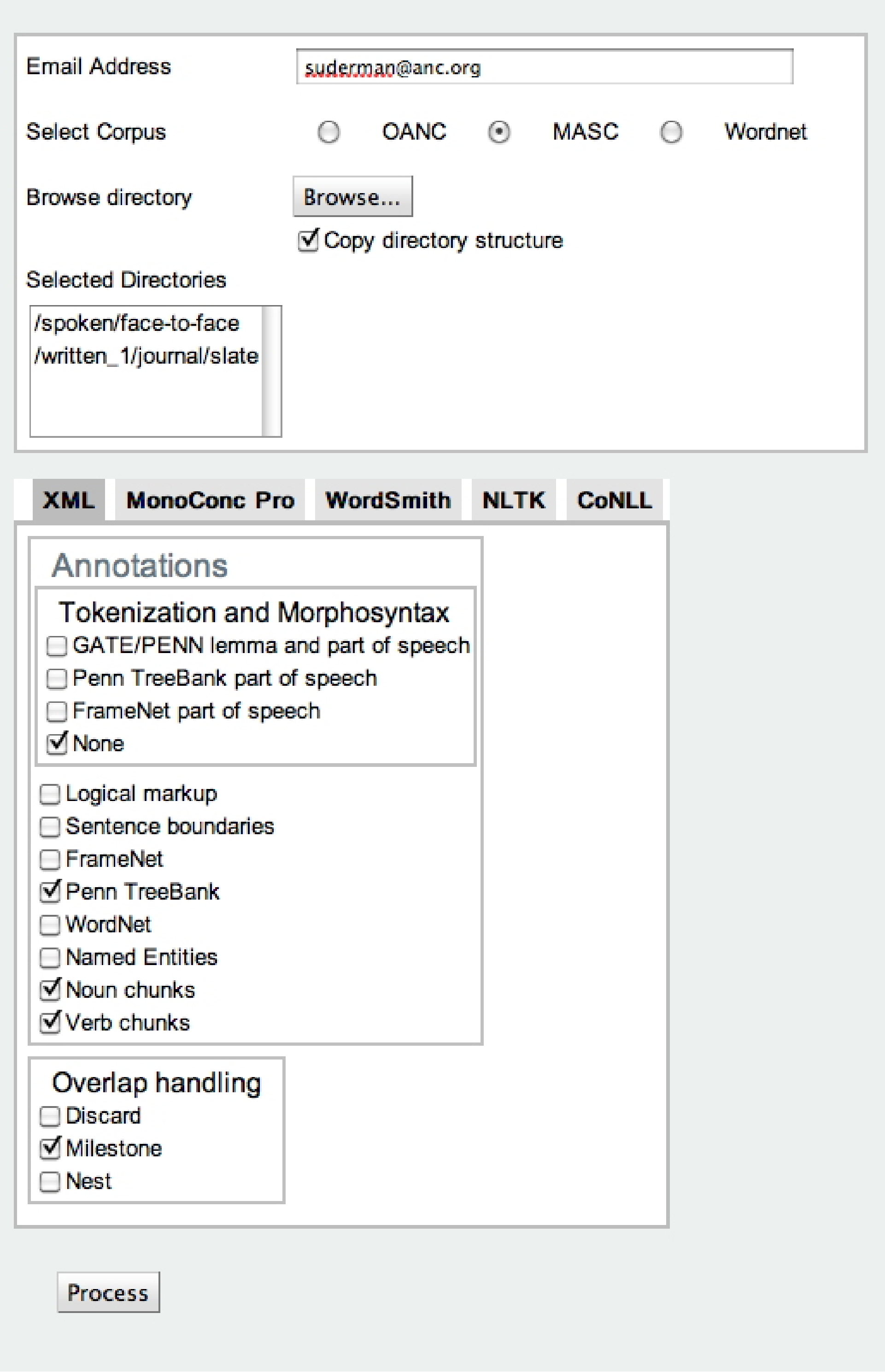
[9] www.nltk.org

[10] http://gate.ac.uk

[11] http://www.oasis-open.org/committees/uima

[12] A group affiliated with the UIMA project has been funded to develop a UIMA module to produce GrAF output from UIMA-internal representations.

[13] http://www.graphviz.org/



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

Annotations

Tokenization and Morphosyntax
- □ GATE/PENN lemma and part of speech
- □ Penn TreeBank part of speech
- □ FrameNet part of speech
- ☑ None

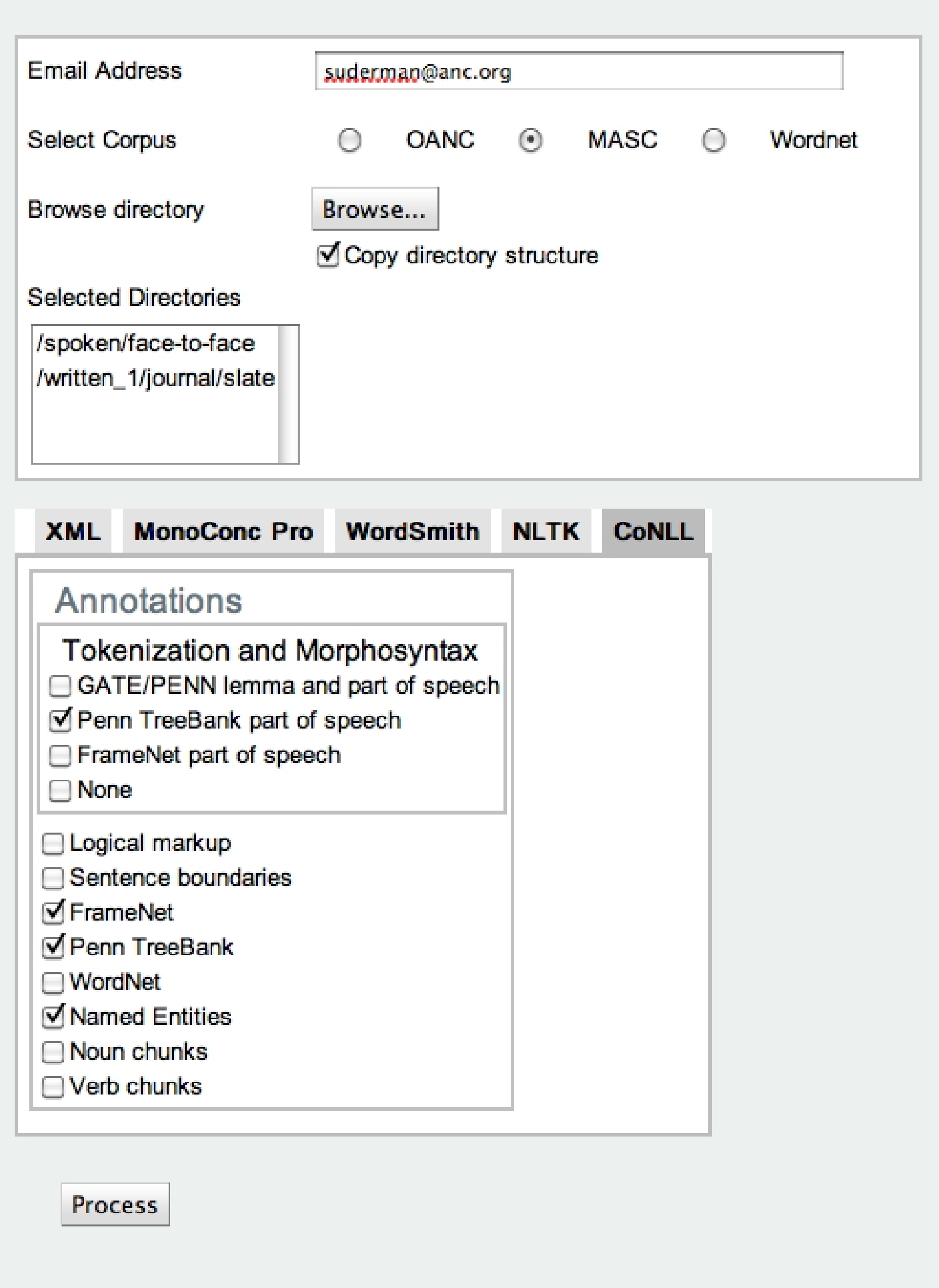
- □ Logical markup
- □ Sentence boundaries
- □ FrameNet
- ☑ Penn TreeBank
- □ WordNet
- □ Named Entities
- ☑ Noun chunks
- ☑ Verb chunks

Overlap handling
- □ Discard
- ☑ Milestone
- □ Nest

Figure 1: ANC2Go interface with XML output chosen

its GrAF XML implementation, and easy development of modules to render GrAF into multiple formats, comprises a very different approach to resource distribution from current common practice. Whether or not this approach or some variant reliant on web services becomes a standard method for resource distribution in the future, ANC2Go provides a proof-of-concept of a viable alternative.

Because ANC2Go is a new type of application for the community of corpus users, we feel it is of considerable interest to monitor its success and use to see if this approach to providing language resources is both viable and valuable to the community. Although ANC2Go has not been available long enough to collect significant statistics, we hope that in the near future, it will serve as a means to assess the viability of the approach and serve as a model for similar services for language resource distribution.



Email Address suderman@anc.org

Select Corpus ○ OANC ● MASC ○ Wordnet

Browse directory

☑ Copy directory structure

Selected Directories

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

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

Figure 2: ANC2Go interface with CONLL output chosen

## Acknowledgments

This work was supported by National Science Foundation grant CRI-0708952.

## 5. References

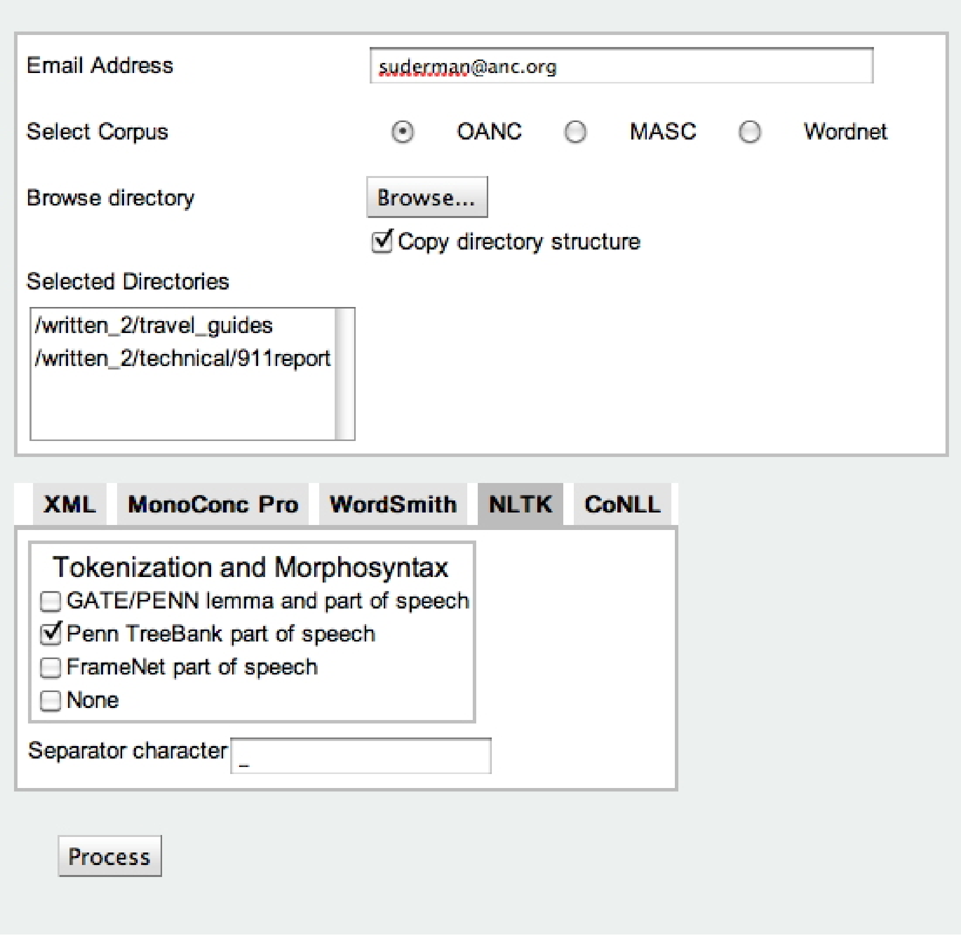
Collin F. Baker and Christiane Fellbaum. 2009. WordNet and FrameNet as complementary resources for annotation. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 125–129, Suntec, Singapore, August. Association for Computational Linguistics.

Nancy Ide, Collin Baker, Christiane Fellbaum, Charles Fillmore, and Rebecca Passonneau. 2008. MASC: The manually annotated sub-corpus of American English. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.

Nancy Ide and Laurent Romary. 2004. International standard for a linguistic annotation framework. *Journal of Natural Language Engineering*, 10(3–4):211–225.

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. 2009. Bridging the gaps: interoperability for GrAF, GATE, and UIMA. In *ACL-IJCNLP '09: Proceedings of the Third Linguistic Annotation Workshop*, pages 27–34, Morristown, NJ, USA. Association for Computational Linguistics.



Email Address suderman@anc.org

Select Corpus ● OANC ○ MASC ○ Wordnet

Browse directory

☑ Copy directory structure

Selected Directories

/written_2/travel_guides
/written_2/technical/911report

Tokenization and Morphosyntax

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

Separator character _

Figure 3: ANC2Go interface with NLTK output chosen