Integrating Linguistic Resources:
The American National Corpus Model

Nancy Ide, Keith Suderman

Department of Computer Science
Vassar College
Poughkeepsie, New York 12604-0520 USA
{ide,suderman}@cs.vassar.edu

Abstract

This paper describes the architecture of the American National Corpus and the design decisions we have made in order to make the corpus easy to use with a variety of existing tools with varying functionality, and to allow for layering multiple annotations over the data. The overall goal of the ANC project is to provide an "open linguistic infrastructure" for American English, consisting of as many self-generated or contributed annotations of the data as possible together with derived. The availability of a wide variety of annotations for the same data and in a common format should significantly simplify the processing required to extract annotations from different sources and enable use of the ANC and its annotations with off-the-shelf software.

1. Introduction

The American National Corpus (ANC) project[1] (Ide and Macleod, 2001; Ide and Suderman, 2004) has released over 20 million words of spoken and written American English, available from the Linguistic Data Consortium. The ANC 2nd release consists of fiction, non-fiction, newspapers, technical reports, magazine and journal articles, a substantial amount of spoken data, data from blogs and other unedited web sources, travel guides, technical manuals, and other genres. All texts are annotated for sentence boundaries; token boundaries, lemma, and part of speech produced by two different taggers[2]; and noun and verb chunks. For a complete description of the ANC 2nd release and its contents, see http://AmericanNationalCorpus.org.

The ANC annotations are automatically generated[3] using a wide range of software, either freely available or contributed. Unfortunately, despite multiple and often redundant efforts to develop means to integrate linguistic resources, including corpora and their annotations as well as lexicons, treebanks, propbanks, etc., such resources continue to be produced in a variety of representation formats. To enable merging annotations from separate stand-off documents, we require that the annotation information be represented in a common format. This format must be both powerful and generic enough to enable mapping annotations in any representation (e.g. LISP structures, XML) and with any internal structure (e.g., tree, graph) to it without information loss.

This paper describes the architecture of the ANC and the design decisions we have made in order to make the corpus easy to use with a variety of existing tools with varying functionality, and to allow for layering multiple annotations over the data. The overall goal of the ANC project is to provide an "open linguistic infrastructure" for American English, consisting of as many self-generated or contributed annotations of the data as possible together with derived resources such as bigram and trigram data, etc. The availability of a wide variety of annotations for the same data and in a common format should significantly simplify the processing required to extract annotations from different sources and enable use of the ANC and its annotations with off-the-shelf software.

2. ANC Architecture

The ANC is represented as a set of graphs over XML documents. Each set includes the header for an individual text and the primary data with no internal markup, together with annotation documents designating segment boundaries for logical structure (LS) down to the level of paragraph, token boundaries (TB), sentence boundaries (SB), and the annotation for a particular linguistic feature. The header file contains genre/domain and other bibliographical information and points to the primary data and each annotation document; annotation documents are linked to the primary data. Figure 1 shows the overall architecture for the ANC 2nd release data, which in addition to segmentation information includes annotations for the Biber part of speech tags (BT), the Penn part of speech tags (PT), noun chunks (NC), and verb chunks (VC).

An ANC primary document and its annotations form a directed graph capable of referencing *n*-dimensional regions of primary data as well as other annotations. The nodes of the graph are virtual, located between each character in the primary data. Edges defined over the nodes in the graph are labeled with feature structures containing annotation information. Thus, an annotation document typically contains sets of edges defining regions of the primary data, each of which is associated with a feature structure. However, it is also possible to define edges over other annotations (creating a second-order "edge graph") since each annotation document can itself

[1] http://AmericanNationalCorpus.org

[2] The 2nd release data includes POS annotation using the Biber tagset and the Penn tagset. The 1st release is also tagged with the C5 and C7 versions of the CLAWS tagset used to tag the BNC.

[3] A 10 million word "gold standard" ANC sub-corpus, balanced for genre and hand-validated for paragraph, sentence and word boundaries as well as lemma and part of speech annotation from the Biber Tagger (Biber, 1988, 1995) is under construction, with plans to add hand-validated syntactic annotation and to annotate and validate portions of sub-corpus for WordNet senses and FrameNet frames.

be treated as primary data over which edges can be defined.

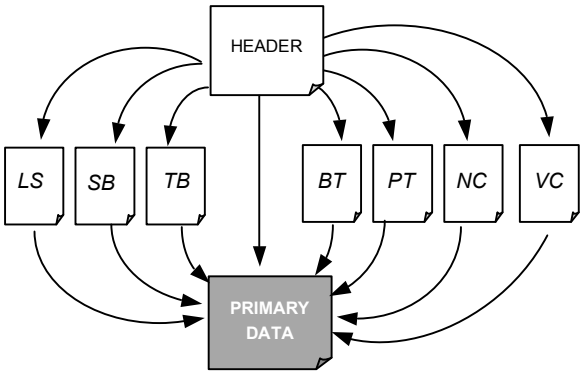

Figure 1. ANC architecture overview

The ANC architecture is an implementation of the International Standards Organization TC37 SC4 (Language Resources)[4] specifications for representing linguistic data and its annotations.

2.1. Representation

We call our implementation of the stand-off approach "extreme stand-off": a text is contained in one UTF-16 encoded document as plain text, without any internal markup. Annotations are represented in a generic XML format for specifying edges and the associated feature structures; as such, the XML elements provide the structure of the annotation document but do not include any information concerning annotation content. This strategy of separating annotation structure from its content allows us to use the same XML format to represent any kind of linguistic annotation. The actual content of the annotation is provided in the attribute/value pairs within the feature structure.[5]

Our choice of representation makes the data extremely flexible: the primary text can be used with no markup or annotations if desired (which is commonly the case for concordance generation, etc.), or the user can choose to deal with a particular annotation set independent of the text (e.g to generate statistics for POS taggers or parsers). Furthermore, we can apply annotations of many different types, or several versions of a single annotation type (e.g., multiple part of speech taggings) without encountering the problems of incompatibility (in particular, the famous "overlapping hierarchy" problem that arises when different systems assign different boundaries to words or other elements in text). The stand-off approach also enables us to distribute annotations independent of the text via download from the ANC site; although for copyright reasons we cannot make the corpus itself freely downloadable from the web without licensing, we can distribute annotations that contain links to the original data so that any user who has obtained the ANC from the LDC can use the annotations with the corpus. The format also has processing advantages: for example, with a corpus the size of the ANC, it can take hours or even days to re-process the entire 20 million words if an error is found in one of the annotation algorithms. However, since the data never change, when an error is found in a particular module—e.g., the sentence splitter—we need only re-run that module, which significantly reduces the processing time required to re-do the annotation.

2.2. Processing and Tools

Processing of the ANC data has been accomplished primarily with the General Architecture for Text Engineering (GATE) system[6] developed by the University of Sheffield. GATE implements a pipeline architecture for annotating corpora by allowing for the application of a series of software components. GATE provides Java software "plugins" for a variety of corpus annotation tasks such as part of speech tagging, several kinds of syntactic analysis, named entity recognition, and co-reference resolution, as well as a machine learning module and sophisticated mechanisms for ontology development and use. The feature of primary value to the ANC project is the ability to add or replace Java modules in the pipeline for processing specific to our needs.

We have developed several GATE plugins for ANC-specific processing and a Java-based scripting language that enables us to pipeline texts in any format (Word, PDF, HTML, Quark Express, etc.) to XML, through a series of annotation tools for sentence splitting, tokenization, lemmatization, part of speech annotation, noun and verb phrase chunking, and output the primary and stand-off documents in the final ANC format.

The layering of annotations over the ANC dictates the use of a stand-off annotation representation format, in which each annotation is contained in a separate document linked to the primary data. At present few software systems handle stand-off annotation, and those that do often demand computational expertise beyond what many ANC users--who include linguists, teachers of English as a second language, etc.--have access to. Therefore, we have developed an easy-to-use tool and user interface to merge the stand-off annotations with the primary data to form a single, well-formed XML document from the user's choice of annotations, which is distributed with the corpus. As a result, the ANC 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.

The ANC merging tool is built upon a SAX-like parser that combines selected annotations with primary data by firing the appropriate events from the SAX2 ContentHandler interface to construct an XML document with in-line annotations. 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 ANC data without first merging annotations and primary data. In the current version of the parser, when overlapping annotations are encountered they are "truncated"; for example:

```
<s>Sentence <em>one.</s><s>Sentence</em>
two.</s>
```

becomes

[4] http://www.tc37sc4.org

[5] A full description of the XML feature structure representation is available at http://www.tc37sc4.org/doc1/iso_tc37-4_n033_rev.11_feature_structures.pdf.

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

```
<s>Sentence <em>one.</em></s><s>Sentence
two.</s>
```

Work is underway to provide the option of generating milestones in CLIX/HORSE (DeRose, 2004) format to represent overlapping hierarchies.

The ANC merging tool provides a graphical user interface that enables users to specify their choice of annotations to be included. Currently, the tool generates the following output formats:

- XML XCES format, suitable for use with the BNC's XAIRA[7] search and access interface;
- Text with part of speech tags appended to each word and separated by an underscore;
- WordSmith/MonoConc Pro format.

The 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.

3. The Open Linguistic Infrastructure

As the ANC has developed, its potential to provide not only a representative corpus of American English, but also a wide variety of invaluable linguistic materials and data derived from it, has become apparent. We now envision the creation of a comprehensive "open linguistic infrastructure" (OLI) for American English, including not only basic resources like frequency wordlists etc., but also multiple annotations for part of speech and syntax using different tagsets, annotation for co-reference and named entities, semantic annotation using categories that link the ANC data to databases such as WordNet and FrameNet, a dependency database reflecting co-occurrences on the basis of semantic role, categories and ontologies describing and linking linguistic information in the ANC, and comparative data for British and American English such as syntactic and lexical variants, etc.

The idea behind the OLI is to use existing and often freely available tools to automatically generate annotations for the ANC data. In particular, we generate multiple versions of annotations of the same type, all provided in stand-off documents downloadable from the ANC web site. For example, we generate syntactic analyses using several freely available parsers, including Minipar[8], Charniak's statistical parser[9], and parsers downloadable from the University of Pennsylvania and Carnegie-Mellon University[10], together with several POS taggings, including the Biber tags, the version of the Penn Tagset generated by the Hepple tagger in GATE[11]. Although unvalidated, multiple alternative annotations for the ANC data not only provide annotations suited to different schemes and linguistic theories, but also enable the comparison and merging of these annotations that could lead to methods for disambiguating automatically-produced tags without the prohibitive cost of hand-validation. For example, combining the results of multiple part of speech taggers has been previously shown as a viable means to produce a more accurate tagging (Brill and Wu, 1998; Tufis, 2000; Sjöbergh, 2003). Also, since part of speech tagsets are designed according to varying criteria (granularity, more or less semantic information, etc.), the availability of a massive corpus annotated with different tagsets can provide valuable information for comparison of linguistic theories. The same applies to combining alternative syntactic and semantic annotations, a strategy that has received less attention.

In addition to generating annotations ourselves, we anticipate that members of the computational linguistics research community will annotate some or all of the ANC data for any of a variety of linguistic features, and will contribute these annotations to the OLI. This expectation is not unreasonable, first of all because the ANC has already received unsolicited annotations for the 1st release data (CLAWS tagging from University of Lancaster, co-reference annotation from University of Alberta; sense tagging from CL Research). More to the point, available corpora such as the *Wall Street Journal* and the MUC corpora have been annotated for different linguistic features (in addition to the co-reference and named entity annotations produced by the MUC evaluation exercises) by different groups, many of which are freely downloadable.[12] Given the availability of a relatively large and "clean" corpus spanning a variety of genres, it can be expected that researchers will annotate the data and willingly contribute the annotations to the OLI.

The ANC project obtains annotations, whether automatically generated in-house or contributed by colleagues, in formats particular to the systems that generate them. At the *representation* level, annotations may be produced in XML, LISP-like formats, or virtually any format meaningful to the producer. XML formats can differ as much from each other as any other type of representation, since an "XML format" uses XML tags but the tags can be used in a dozen different ways—even when supposedly instantiating a particular architecture.[13] At the *content* level, discrepancies may be even more radical, since not only may one scheme label some linguistic phenomenon differently than another, but the content categories themselves may not be mappable due to conceptual variations.

All OLI annotations are transduced into the stand-off format described in section 2.1 and provided on the web in this format; this is possible because of the generality of the representation scheme we have adopted, which provides only a "structural skeleton" to which actual annotation content is attached. It is important to note that content categories in the original annotation are left untouched; however, the availability of a wide and varied set of annotations represented in a single structural format significantly simplifies the processing required to use, merge, and compare them, if only because it reduces the

[7] http://sourceforge.net/projects/xaira

[8] http://www.cs.ualberta.ca/~lindek/minipar.htm

[9] http://www.cs.brown.edu/~ec/#software

[10] CMU Link parser, available at http://www.link.cs.cmu.edu/link/index.html

[11] http://americannationalcorpus.org/FirstRelease/gatetags.txt

[12] For example, metonymy annotation of MUC from the Mascara project: http://homepages.inf.ed.ac.uk/mnissim/mascara/

[13] For example, "annotation graphs" can be represented using XML pointers, an offset scheme given as element attributes or as content of special elements, or, if no overlapping hierarchies exist, XML elements to delimit start and end points of an annotation.

number of schemes from which to transduce from many to one.

The OLI is also providing semantic tagging for the ANC data. Our strategy here is similar to the one described above for part of speech and syntactic annotation: we provide multiple sense taggings (using WordNet sense tags[14]) produced by different disambiguation systems in stand-off annotation documents linked to the original data. A large number of state-of-the-art word sense disambiguation systems exist that utilize the WordNet sense tags, many of which are freely available for research purposes (e.g. Ted Pedersen's Duluth system). In addition, CL Research[15] is contributing WordNet sense tagging of nouns, verbs, adjectives, and adverbs using their software.

Although at the very best, WSD systems achieve accuracy of only about 70-80%, the existence of multiple sense annotations in a common format for a massive corpus can lead to the development of means to produce more accurate sense tagging by simple combination of results via voting or employing heuristics to exploit the strengths of individual systems. Even at 70% accuracy, a sense-tagging of the ANC will be a valuable resource in its own right.

4. Summary

The science of annotating data with linguistic information has been ongoing for over fifteen years. Since the early 1990's, several projects in Europe have worked toward development of principled, standardized annotation methods and schemes; similar activities have begun in the U.S. in the past few years. A number of "big ideas" have been accepted as *de facto* standard practice by the community, such as the use of stand-off annotation introduced in the Corpus Encoding Standard (Ide and Priest-Dorman, 1994; Ide, 1998) and the use of feature structures as a data model for the structure of annotations. However, "annotation science" remains an increasingly active area of research, and development of consistent annotation content areas (especially for semantic tagging) will require considerably more work.

It is our view that we can represent linguistic annotations of any type in the common and easily transducable format described in this paper, but that adoption of a unique scheme for describing each type of annotation *information* (i.e., content) is premature. The OLI is intended to provide resources, in the form of multiple annotations, that will accommodate different approaches to linguistic annotation while at the same time enabling exploration of ways to refine existing annotation schemes, improve automatic annotation accuracy, and enable movement toward commonly accepted practices and schemes.

The OLI is a unique project in that it relies on gathering contributed annotations, together with annotations automatically generated by available tools. No open repository of linguistic information of the OLI's kind and scope exists for corpora in any language. Certain widely available, genre-specific corpora have been annotated by different members of the research community, but these annotations are not gathered together in one place nor are they necessarily represented in the same format. Some annotations are not publicly available or are available only to members of certain research efforts. By making resources of this kind available, the ANC project will provide a much-needed resource for continued development of annotation methods and schemes. In addition, given our commitment to using leading edge representation models for linguistic data in the ANC project, these materials will comprise an instantiation of the state-of-the-art for representing linguistic data that can be integrated into the developing framework of the semantic web.

[14] The version of WordNet used for the tagging will depend in part on the software used to generate them, and it is possible we will have taggings using different WordNet versions. Mappings among different WordNet versions are available.

[15] http://www.clres.com

5. Acknowledgements

The work reported in this paper has been supported by t

he American National Corpus Consortium, National Science Foundation grant BCS 0218609, and the Linguistic Data Consortium.

6. References

Biber, D. (1988). *Variation Across Speech and Writing.* New York: Cambridge University Press.

Biber, D. (1995). *Dimension of Register Variation.* New York: Cambridge University Press.

Brill, E., Wu, J. (1998). Classifier Combination for Improved Lexical Disambiguation. In *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*. San Francisco, CA: Morgan Kaufmann Publishers, pp. 191-195.

DeRose, S. J. (2004). Markup Overlap: A Review and a Horse. In *Extreme Markup Languages 2004: Proceedings*. http://www.mulberrytech.com/Extreme/Proceedings/html/2004/DeRose01/EML2004DeRose01.html

Ide, N. (1998). Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora. In *Proceedings of the First International Language Resources and Evaluation Conference,* pp. 463-70.

Ide, N., Macleod, C. (2001). The American National Corpus: A Standardized Resource of American English. In *Proceedings of Corpus Linguistics 2001,* Lancaster UK.

Ide, N., Priest-Dorman, G. (1994). The Corpus Encoding Standard. http://www.cs.vassar.edu/CES.

Ide, N., Suderman, K. (2004). The American National Corpus First Release. In *Proceedings of the Fourth International Language Resources and Evaluation Conference* (LREC), Lisbon, Portugal, pp. 1681-1684.

Sjöbergh, J. (2003). Combining POS-taggers for improved accuracy on Swedish text, NoDaLiDa 2003, Reykjavik.

Tufis, D. (2000). Using a Large Set of EAGLES-compliant Morpho-Syntactic Descriptors as a Tagset for Probabilistic Tagging. In *Proceedings of the Second International Language Resources and Evaluation Conference,* pp. 1105-111.