The American National Corpus: Then, Now, and Tomorrow

Nancy Ide
Vassar College

1. The Beginning: ANC as BNC Clone

The ANC was motivated by developers of major linguistic resources such as FrameNet[1] and Nomlex,[2] who had been extracting usage examples from the 100 million-word British National Corpus (BNC), the largest corpus of English across several genres that was available at the time. These examples, which served as the basis for developing templates for the description of semantic arguments and the like, were often unusable or misrepresentative due to significant syntactic differences between British and American English. As a result, in 1998 a group of computational linguists proposed the creation of an American counterpart to the BNC, in order to provide examples of contemporary American English usage for computational linguistics research and resource development (Fillmore, Ide, Jurafsky, & Macleod, 1998). With that proposal, the ANC project was born.

The ANC project was originally conceived as a near-identical twin to its British cousin: The ANC would include the same amount of data (100 million words), balanced over the same range of genres and including 10% spoken transcripts just like the BNC. As for the BNC, funding for the ANC would be sought from publishers who needed American language data for the development of major dictionaries, thesauri, language learning textbooks, et cetera. However, beyond these similarities, the ANC was planned from the outset to differ from the BNC in a few significant ways. First, additional genres would be included, especially those that did not exist when the BNC was published in 1994, such as (we)blogs, chats, and web data in general. The ANC would also include, in addition to the core 100 million words, a 'varied' component of data, which would effectively consist of any additional data we could obtain, in any genre, and of any size. In addition, the ANC would include texts produced only after 1990 so as to reflect contemporary American English usage, and would systematically add a layer of approximately 10 million words of newly produced data every five years.

Another major difference between the two corpora would be the representation of the data and its annotations. The BNC exists as a single enormous SGML (now, XML) document, with hand-validated part of speech annotations included in the internal markup. By the time the ANC was under development, the use of large corpora for computational linguistics research had sky-rocketed, and several preferred representation methods had emerged—in particular, *stand-off* representations for annotations of linguistic data, which were stored separately and pointed to the spans in a text to which they referred, were favored over annotations that were interspersed within the text. The ANC annotations would therefore be represented in stand-off form, so as to allow, for example, multiple annotations of the same type (e.g., part of speech annotations produced by several different systems). Finally, the ANC would include several types of linguistic annotation beyond the part-of-speech annotations in the BNC, including (to begin) automatically produced shallow syntax and named entities.

1.1. Funding and Data Sources

The BNC was substantially funded by the British government, together with a group of publishers who provided both financial support and contributed a majority of the data that would appear in the corpus. Based on this model, the ANC looked to similar sources, but gained the support of only a very

[1] http://www.icsi.berkeley.edu/~framenet
[2] http://nlp.cs.nyu.edu/nomlex/index.html

few U.S. publishers. The majority of the fifteen or so publishers who did contribute funding to the ANC included several Japanese publishers of texts on English as a second language, and a subset of the same British publishers who had supported the BNC. These publishers, together with a handful of major software developers, provided a base of financial support for the project over a 3-year period, but nothing like the support that had been provided to the BNC. After a time, the ANC project also secured a small grant from the National Science Foundation to produce a 'gold standard' subcorpus of 10 million words. All in all, the ANC secured about $400,000 to support its first 4 years, much less funding than that which supported development of the BNC.

British publishers provided the bulk of the data in the 100 million-word BNC. The plan for the ANC was that the members of the ANC consortium, which included both publishers and software vendors, would do the same for the ANC. However, only a very few of the ANC consortium members eventually contributed data to the corpus.[3] As a result, it was necessary to attempt to find data from other sources, including existing corpora such as the Indiana Center for Intercultural Communication (ICIC) Corpus of Philanthropic Fundraising Discourse and the Charlotte Narrative and Conversation Collection (CNCC), as well as government and other public domain documents on the web.

1.2. Software Development and Distribution

Development of the BNC included the production of a software system for searching the corpus, generating concordances, etc. (XIARA). It was clear from the outset that without the same funding resources, the ANC project would be unable to produce software of its own. The alternative was to represent the corpus and its annotations in such a way that it could be used with existing software, including XIARA as well as widely used commercial concordance software (e.g., MonoConc, WordSmith) and text engineering systems such as GATE.[4] This meant that the ANC and its annotations had to be represented in a format that could be straightforwardly transduced to virtually any other input format required by text analytic software—a non-trivial requirement. The representation format of the ANC is described in Section 4.

The ANC is distributed through the Linguistic Data Consortium (LDC) for a reproduction fee of $75.00 for non-members who will use it for research purposes only. The BNC is distributed by the Oxford University Computer Services; it was originally available only to researchers within the European Union for a price of several hundred British pounds, and later available worldwide at a similar price. In recent years, it has become available to all users for research purposes for a fee of 75 British pounds.

2. ANC 1998-2008

In 2003, the ANC produced its first release of 11 million words of data, which included a wide range of genres of both spoken and written data.[5] Annotations included word and sentence boundaries and part-of-speech annotation produced by two different taggers: the 'Hepple tagger', which uses the Penn part-of-speech tags, and the 'Biber tagger', which uses a superset of the CLAWS part-of-speech tags used to tag the BNC. The annotations in this release were represented in stand-off form—that is, they were not included inline with the text but rather provided as separate files with links into the data. To our knowledge, the ANC First Release was the first large, publicly available corpus to be published with stand-off annotations. Because of the lack of software for handling stand-off annotations, a version of the ANC First Release with inline annotations was also included in the LDC distribution.

In 2005, the ANC released an additional 11 million words, bringing the size of the ANC to 22 million words. The Second Release includes data from additional genres, most notably a sizable subcorpus of (we)blog data, biomedical and technical reports, and the 9/11 Report issued by the U.S. Government. The Second Release was issued with stand-off annotations only for the same phenomena as in the First Release, as well as annotations for shallow parse (noun chunks and verb chunks) and two additional part-of-speech annotations using the CLAWS 5 and 7 tags, to enable comparison with BNC

[3] The consortium members who contributed texts to the ANC are Oxford University Press, Cambridge University Press, Langenscheidt Publishers, and the Microsoft Corporation.
[4] General Architecture for Text Engineering; http://gate.ac.uk
[5] The contents of the ANC First Release are described at http://www.anc.org/FirstRelease/

data. To facilitate use of the ANC, the release included the first version of the ANC Tool, which generates parts or all of the corpus with inline annotations chosen from among those available, in formats usable by XAIRA, MonoConcPro, and WordSmith, as well as a generic inline XML format.

After 2005, the ANC project had no more funding, and production of additional data came to a halt. In 2006, the project made 15 million of the ANC's 22 million words, called 'the Open ANC' (OANC), available from the ANC website. This subset of the corpus is free of licensing restrictions, and therefore can be made available to anyone for any purpose, research or commercial. The OANC distribution model of completely open access is a step beyond the more typical GNU Public License, which requires redistribution under the same license, which is prohibitive for commercial users. The fully open distribution model pioneered by the ANC has now been adopted for all future releases of ANC data and annotations.

In 2007 the ANC received a substantial grant from the U.S. National Science Foundation[6] to produce a Manually Annotated Sub-Corpus (MASC) of the ANC. The grant provided no funding to add to the existing ANC, but rather provided funds to validate automatically produced annotations for part of speech, shallow parse, and named entities, and to manually add annotations for WordNet senses and FrameNet frames to portions of the corpus. The subcorpus is also annotated with validated Penn TreeBank syntactic analysis. In the context of this grant it has been possible to process additional data for the OANC that had been collected earlier, and therefore the OANC will grow in 2009 to at least double its current size.

3. The Data Acquisition Problem: Why Not the Web?

Without substantial contributions of data from publishers and other sources, the major issue for development of the ANC has been data acquisition. Over the past several years, computational linguists have turned to the web as a source of language data, and several years ago the proponents of the web-as-corpus predicted that development of corpora like the ANC was a thing of the past. The most common counter-argument in favor of a resource like the ANC is that a web corpus is not representative of general language use; for example, one study showed that web language is highly skewed toward dense, information-packed prose (Ide, Reppen, & Suderman, 2002), and another recently expounded some of the shortcomings of unedited web data for NLP research (Kilgarriff, 2007). However, the most significant argument against the web-as-corpus is that studies involving web data are not replicable, since the 'corpus' and any accompanying annotations cannot be redistributed for use by others. Copyright law, at least in the U.S., specifies that all web data are copyrighted unless explicitly indicated to be in the public domain or licensed to be redistributable through a mechanism such as Creative Commons.[7] Contrary to popular opinion, this includes all of the data in Wikipedia, which has been heavily used in NLP research in recent years.

Although the fact that web data is implicitly copyrighted provides some justification for development of a resource like the ANC, this fact also presents the greatest obstacle to data acquisition. Data on the web—including PDF and other documents that are not typically included in web corpora—are the most likely source of material for inclusion in the ANC; however, the vast majority of web data in the public domain is at least 50 years old because of copyright expiration, and the ANC requires data produced since 1990. The search for more recent web documents that are explicitly in the public domain or licensed for reuse is therefore not only time-consuming, but also yields relatively meagre results. As a result, the ANC has had to rely primarily on government sites for public domain documents, as well as web archives of technical documents such as Biomed[8] and the Public Library of Science.[9] To attempt to gather data from other sources, the ANC project has put up a web interface[10] to enable contributions of texts from donors such as college students, who are asked to contribute the

[6] NSF CRI 0708952
[7] http://creativecommons.org/
[8] http://www.biomedcentral.com/
[9] http://www.plos.org
[10] http://www/anc.org/contribute.html

essays, fiction, et cetera they have written for classes; the ANC Facebook page is one mechanism for reaching out to this audience.[11]

4. The ANC Processing Pipeline

Data to be included in the ANC come to us in many forms, including plain text, HTML, Word doc and rtf format, PDF, and various publishing software formats such as Quark Express. Depending on the original format, a more or less complex series of processing steps is required to transform the document into the XML format defined by the XML Corpus Encoding Standard (XCES) (Ide, Bonhomme & Romary, 2000), which is the intermediate form used in the ANC processing pipeline. XCES documents can be loaded directly into GATE, which renders all XML formatting information into stand-off form for internal processing. We utilize many of GATE's built-in processors (some of which we have customized for our use) to annotate documents for token and sentence boundaries, part-of-speech (Penn tags), noun and verb chunks, and named entities. We have developed plugins to GATE that export all of these annotations in their final GrAF format for inclusion in the ANC (see Section 5). The full ANC processing pipeline, including potential transduction to specific formats, is shown in Figure 1.

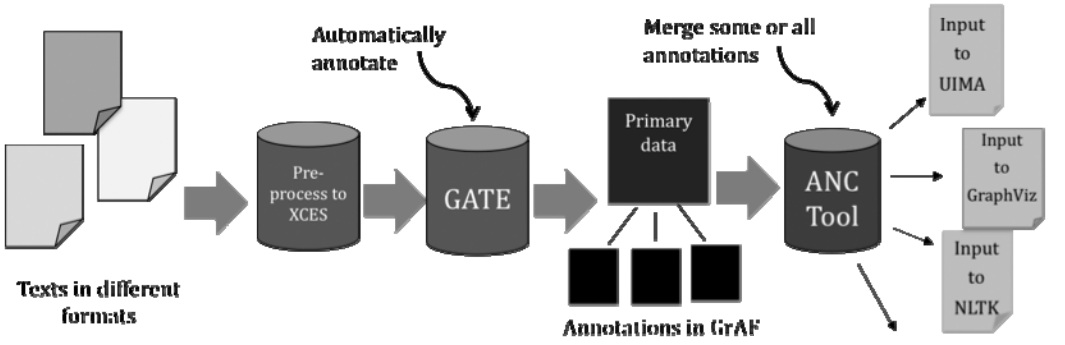

Figure 1. ANC processing pipeline.

Unexpectedly, the easiest format to preprocess is Word doc and rtf; after a document in one of these formats is opened in Open Office,[12] it can be exported in XCES XML using an XSLT style sheet, which enables retaining the formatting information. Thus, for a Word document, the full processing pipeline is a push-button operation. However, other formats pose more difficulties. Text cannot be straightforwardly extracted from PDF documents without requiring some semiautomatic post-editing to eliminate page numbers and running heads, et cetera. We have so far discovered no method to extract text from multi-column PDF that does not require extensive post-editing.[13] Formats such as Quark Express, which are used to represent print-ready documents for publishers, present problems with special characters such as ligatures, initial capitals, et cetera. HTML poses its own set of well-known problems, which are documented in detail in proceedings of the CLEANEVAL exercises.[14] Plain text is easy to process, but lacks formatting information beyond the identification of paragraph boundaries.

In addition to the formats mentioned above, the ANC often receives data rendered in an arbitrary XML format that provides some kind of annotation. Although this would seem to be ideal since XSLT could be used to transform it to XCES, it should never be assumed that one person's XML is mappable to another's. For example, the ICSI[15] Meeting Corpus, consisting of spoken transcripts of multi-participant meetings, was contributed to the ANC in an XML format that encloses every distinct fragment of the transcript within a `<Segment>` element, including not only spans of speech, but also 'events' such as microphone noise, laughing, et cetera., and added information such as comments by

[11] To date, we have collected over a half a million words of college essays an fiction contributed by college students.
[12] http://www.openoffice.org
[13] For this reason, we have been unable to include a million words of contributed data from the ACL Anthology in the ANC.
[14] http://cleaneval.sigwac.org.uk/
[15] International Computer Science Institute, University of California at Berkeley

the transcribers. Because there is no embedding of Segment elements in the transcripts, extensive processing is required to rejoin parts of a speaker turn that are separated by a segment indicating an interruption (noise, etc.) or transcriber comment. Because these interruptions frequently occur in mid-sentence, the separation poses problems for subsequent part-of-speech and syntactic analysis. It is, however, often cost-prohibitive to render contributed annotations in an optimal form, and in such cases the data and annotations are released 'as is'.

5. Representation of ANC Data and Annotations

The ANC is the poster child for the specifications under development by the International Standards Organization (ISO) TC37 SC4 (Language Resource Management) for representing language data and annotations. The ISO Linguistic Annotation Framework (LAF) provides a general framework for representing annotations (Ide & Romary, 2004, 2006), and the Graph Annotation Format (GrAF) (Ide & Suderman, 2007) is the XML serialization of the LAF abstract data model. GrAF is intended to function in much the same way as an interlingua in machine translation, that is, as a 'pivot' representation into and out of which user- and tool-specific formats are transduced, so that a transduction of any specific format into and out of GrAF accomplishes the transduction between it and any number of other GrAF-conformant formats.

The development of LAF and the ANC has been symbiotic: The ANC has effectively served as a testing ground for LAF, which has in turn evolved based on the experience gained in representing the ANC. This evolution has meant that the representation format for the ANC has changed from release to release, but now that LAF is stable, the ANC provides the first example of the state-of-the-art in language resource representation. In GrAF, annotations are represented with feature structures, and each feature structure is associated with a node in a graph over primary data. The leaf nodes of the graph are *n*-dimensional regions in the primary data to which an annotation may apply. Annotations can also be linked via the edges between nodes to other annotations, thus allowing consideration of multiple annotations as a single graph. This feature differentiates a GrAF representation from other representation formats such as annotation graphs (Bird & Liberman, 2001).

The rendering of ANC data and annotations in GrAF satisfies a primary criterion of the ANC design: the ability to transduce ANC data and annotations into formats required by various software systems. The original ANC Tool has now evolved into a GrAF API that provides transduction from GrAF to an increasing number of formats (including UIMA CAS, for use with the Unstructured Information Management Architecture,[16] and formats compatible with NLTK[17]), as well as easy means to develop transducers to other formats. However, in addition to being readily transduced to other formats, the GrAF format is useful in itself: One of the most salient features of the graph representation for linguistic annotations is the ability to exploit the wealth of graph-analytic algorithms for information extraction and analysis. For example, it is trivial to merge independently produced annotations of the same data in GrAF form, as well as to apply algorithms to find common subgraphs that reflect relations among different annotations. The exploitation of GrAF representations of multiple annotation types, as well as multiple annotations of the same type, is ripe for future linguistic research.

6. What Now?

As noted in Section 2, the ANC currently has support from the U.S. National Science Foundation to prepare a subcorpus of the OANC with validated annotations. We expect to produce up to 500,000 words of data with validated annotations by the end of the project in 2011. All of the data and annotations will be freely available for download from the ANC website.

Based on our experience and a growing interest in the research community in fully free and open access to resources, the ANC project is committed to producing only data and annotations that can be distributed without restriction from our website. Therefore, all future releases will become a part of the OANC and be accessible from the ANC website; for convenience, the Linguistic Data Consortium will continue to distribute all ANC data and annotations as well. Although we have very little funding for

[16] http://www.oasis-open.org/committees/uima
[17] http://www.nltk.org

production of additional data for the corpus, the project intends to continue releasing data periodically, especially since in many cases the processing pipeline is in place and very little effort is required once we acquire the data itself.

The major obstacle to continued growth of the OANC is data availability, especially since we are now committed to including only data that can be redistributed without restriction. This requirement has ramifications for the balance of the corpus across genres; at this point, we are more interested in obtaining large amounts of data and less interested in creating a corpus whose balance reflects that of the BNC. Balance still remains a concern, and we actively seek the hard-to-obtain genres such as fiction. However, although the OANC is not (yet) a model of representativeness like the BNC, it nonetheless contains the widest variety of genres of any large, redistributable corpus of contemporary English in existence.

Corpus creation is an extremely time- and cost-intensive effort. The situation is made worse by the lack of funding that can plague such endeavors if substantial government and private support is not provided up front, as it has been for the BNC. Despite the difficulties, though, several new 'national corpus' efforts are proposed or underway, in recognition of the need for these resources to further linguistic research. We hope that the experience we have gained and, in particular, the methods and tools we have developed, can be used in such projects and lessen the time and effort involved to develop these important language resources. More to the point, we urge proposed efforts—and especially the Australian National Corpus project—to represent its data and annotations in ways that, if not identical to, are interoperable with the representation of the OANC so that the resources can be used together, either for comparison or in conjunction. This requirement is important for all resource development, but especially for the national corpora of the various brands of English.

References

Bird, Steven, & Mark Liberman. (2001). A formal framework for linguistic annotation. *Speech Communication, 33* (1-2), 23-60.

Fillmore, Charles, Nancy Ide, Dan Jurafsky, & Catherine Macleod. (1998). An American national corpus: A proposal. *Proceedings of the First International Language Resources and Evaluation Conference* (LREC), Granada, Spain, 965-70.

Ide, Nancy, Randi Reppen, & Keith Suderman. (2002). The American National Corpus: More than the web can provide. *Proceedings of the Third Language Resources and Evaluation Conference* (LREC), Las Palmas, Canary Islands, Spain, 839-44.

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

Ide, Nancy, & Laurent Romary. (2006). Representing linguistic corpora and their annotations. *Proceedings of the Fifth Language Resources and Evaluation Conference* (LREC), Genoa, Italy.

Ide, Nancy, & Keith Suderman. (2007). GrAF: A graph-based format for linguistic annotations. *Proceedings of the Linguistic Annotation Workshop*, held in conjunction with ACL 2007, Prague, June 28-29, 1-8.

Ide, Nancy, Patrice Bonhomme, & Laurent Romary. (2000). XCES: An XML-based standard for linguistic corpora. *Proceedings of the Second Language Resources and Evaluation Conference* (LREC), Athens, Greece, 825-30.

Kilgarriff, Adam. (2007). Googleology is bad science. *Computational Linguistics, 33* (1), 147-151.

Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages

edited by Michael Haugh, Kate Burridge, Jean Mulder, and Pam Peters

Cascadilla Proceedings Project Somerville, MA 2009

Copyright information

ISBN 978-1-57473-435-5 library binding

Ordering information

Orders for the library binding edition are handled by Cascadilla Press.
To place an order, go to www.lingref.com or contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, fax: 1-617-776-2271, e-mail: sales@cascadilla.com

Web access and citation information

This entire proceedings can also be viewed on the web at www.lingref.com. Each paper has a unique document # which can be added to citations to facilitate access. The document # should not replace the full citation.

This paper can be cited as:

Ide, Nancy. 2009. The American National Corpus: Then, Now, and Tomorrow. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus*, ed. Michael Haugh et al., 108-113. Somerville, MA: Cascadilla Proceedings Project.

or:

Ide, Nancy. 2009. The American National Corpus: Then, Now, and Tomorrow. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus*, ed. Michael Haugh et al., 108-113. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #2293.