The American National Corpus
Overall goals and the first release

Randi Reppen and Nancy Ide

Northern Arizona University and Vassar College

The American National Corpus

Many of you reading this article are familiar with the British National Corpus (BNC), a 100 million word corpus of both spoken and written language across a variety of registers (formal speeches, informal conversations, newspapers, etc.) (Aston & Burnard 1998). The BNC has been a tremendous resource for both English language researchers and English language teachers since 1994. Prior to the release of the BNC, there were no large corpora (i.e., over one million words) widely available to language researchers that were deliberately designed to represent a range of registers .. By providing access to a 100 million word corpus of English across a range of registers, the BNC enabled corpus linguists, computational linguists, and English language teachers to investigate a wide range of questions. Corpus linguists and teachers can explore questions about grammar, language use, and tease out patterns of variation across different contexts of use. Computational linguists can develop language models to guide language processing software, fine tunemethods of automatic and semi-automatic analysis, and explore the ways in which texts from different registers present different issues for text processing.

In 1998 a plan to build an American National Corpus (ANC) comparable to the BNC was proposed (Fillmore, Ide, Jurafsky, Macleod, 1998). Like the BNC, the ANC project is undertaken in cooperation with a consortium of publishers, organizations, and academic institutions in the US (for a complete listing of consortium members please go to the ANC website: http://AmericanNationalCorpus.org). The process of creating a representative corpus that is widely available demands a ragne of activities (e.g., text collection, format issues, legal issues related to distribution, etc.) and necessitates close cooperation and coordination across many the many involved entities. Since the project's inception in 1998 there have been various glitches and bumps, but in October of 2003 the first 11.5 million words of the ANC were released. This article provides an overview of the ANC project's goals and a detailed look at the first 11.5 million words.

Overall goals of the ANC

Like any project of this magnitude, the ANC is intended to serve a range of goals. The primary goal of the ANC project is to create a carefully designed corpus of 100 million words of American written and spoken language that generally follows the framework of the BNC. Although the ANC is following the general design of the BNC, there are some differences in the categories and texts included in the ANC. The ANC will only contain

texts from 1990 on, while the BNC contains texts from 1960 – 1993. Due to the timeframe during which the BNC was collected, it does not include electronic texts such as e-mail or webpages. The ANC, however, will contain electronic texts such as e-mail, webpages, and e-talk from chat rooms.

55% Books
Non-fiction 41%
Natural Sciences (e.g., Biology Chemistry Geology Math, Physics)
Applied Sciences (e.g., Communication, Energy, Engineering)
Social Sciences (e.g., Anthropology, Geography, Psychology Sociology)
World Affairs (e.g., Economics, Government, History, Politics)
Commerce (e.g., Business, Finance, Industry, Economics)
Arts (e.g., Media, Performing Arts, Visual Arts)
Beliefs (e.g., Mythology, Astrology, Philosophy Religion)
Leisure (Antiques, Gardening, Hobbies, Travel)
Biographies
Fiction 14%
General fiction
Historical fiction
Science fiction
Romantic fiction
Mystery
Adventure
Poetry
Drama
Humor
20% Newspapers, Magazines and Journals
Magazines – General & Specialized
Newspapers – National and Local (Daily and Weekly)
Journals
10% Spoken
Face-to-face
Meetings
Phone conversations
Planned speeches
10% Electronic
Web pages; E-mail; Chat rooms
5% Miscellaneous (published & unpublished material)
Solicitation letters; Brochures; Memos

Figure 1: Target Design of the 100 million word American National Corpus (ANC).

A secondary goal of the ANC project is to provide access to specialized corpora. In addition to the core ANC, or BNC-like portion, the ANC will also include a "satellite" corpus. During the text collection process several groups and organizations (e.g., Project MORE, ICE and LAWS) have agreed to include their corpora, which were otherwise unavailable or have only limited access, in the ANC. In some cases, portions of these corpora will be used to fill out categories in the core ANC (e.g., conversation, letters). Complete versions of these specialized corpora will also be included in the "satellite" portion of the ANC. Other corpora that do not fall into any of the BNC-like categories of the ANC, will be included in the "satellite" portion of the ANC. Access to these

"satellite" corpora will provide a valuable resource to those interested in specialized corpora, and especially text types that are not represented in existing corpora of English.

A final goal of the ANC project is to provide both a standard format for text encoding and a format that allows for different types of corpus annotation (e.g., parts of speech, rhetorical features, etc.) as well as different versions of the same type of annotation (e.g. multiple part of speech taggings). The ANC is encoded in XML and is conformant to the XML Corpus Encoding Standard (XCES) schemas for primary data and annotations (Ide 1998a & b, Ide, Bonhomme, & Romary, 2000), which is compliant with emerging standards for representing data such as various W3C standards (e.g., XPointer for inter-document linking). Following XCES recommendations, the ANC is encoded using a "stand-off" annotation strategy, meaning that linguistic annotations are contained in separate XML documents linked to the original rather than being interspersed with the original data in a single XML document. Because few processors are able to handle stand-off annotation at this time, a "merged" version of the First Release of the ANC has been provided.

Part of speech annotation of the ANC has been done using the Biber tagger (Biber 1988; Biber, Conrad and Reppen 1998). The ANC is also being tagged with the C5 and C7 versions of the CLAWS tagger (see Garside, Leech, and Sampson 1987 for a description of an earlier version of CLAWS). Both taggers identify parts of speech (POS) and some morpho-syntactic features. In the future, the ANC will also be tagged with a version of the Penn Tags (Marcus, Santorini, and Marcinkiewicz, 1993) r Full descriptions of all the tag sets used by the ANC are available at the ANC website (http://americannationalcorpus.org).

In addition to the different POS taggings, it is hoped that as researchers explore different aspects of language (e.g., cohesion, rhetorical moves, etc.), and annotate texts in the ANC for those features, they will contribute their annotations for distribution by the ANC project. This multi-layering of annotations can provide rich linguistic descriptions from different perspectives, and also enable combining annotations at different linguistic levels for more comprehensive description.

The ANC received a grant from the National Science Foundation to hand edit 10% of the corpus for both encoding and linguistic annotation. This information will be used to fine tune the tools used to encode and annotate the ANC and to develop statistical language models to guide parsers and other language processing tools.

Challenges

Some of the major challenges of creating the ANC are selection and acquisition of texts; legal issues related to copyright and use of the texts; and trransduction of the texts into a common format. In order to build a representative corpus of American English it is necessary to use texts that are created for and/or by the general public. In some ways this may seem an easy task, but imagine that you were asked to take a photo that was characteristic of the US. At once you will be confronted with the geological, geographical, and cultural diversity of the US and any decision both includes and excludes certain aspects. Many of these same issues are faced when building a corpus. Since we were following the general framework of the BNC, decisions concerning the

categories to include were already made; the task of selecting which individual books to include within each category still demanded many decisions.

After titles were selected, lists of desired texts were submitted to the participating publishers, who then determined if those texts were available for inclusion in the ANC. The task of acquiring texts once they have been identified is still a challenge. Often sources are not able to contribute all of texts requested and then the cycle of selection and request begins again.

Those contributing texts, particularly publishers and other major organizations, needed to be certain that the copyright laws were going to be honored. At the beginning of the project we naively thought this task would take about three to six months; however, in reality it has taken much longer. Without the help of the Linguistic Data Consortium (LDC), which is handling the licensing issues involved in the ANC, the task of creating legal agreements for text contributors and creating licensing agreements for both text contributors and users of the ANC would not have been possible.

Aside from the legal issues, the other major challenge has been that of text formats. Because the ANC relies on texts being contributed from a variety of sources, the formats used by these sources to create and or store the texts vary greatly. Even different versions of the same software (e.g., Quark) can render files in formats that are time consuming to process. Keith Suderman, the programmer for the ANC, has wrestled with the different formats and has been able transduce a myriad of formats into a common encoding format. In addition to dealing with the transformation of text formats, Suderman is also responsible for the creation of tools to allow users to interface with the ANC and carry out various types of queries.

The First Release

The First Release of the ANC has several purposes. First and foremost, in addition to providing an immediate resource to corpus linguists, computational linguists and English language educators, the First Release demonstrates that the ANC project is making progress. The First Release, although by no means a balanced or representative corpus, does provide wide access to more American speech than was previously available, along with several other registers. The First Release also provides an opportunity to identify bugs and user issues that can be addressed prior to the release of the entire 100 million words.

The CD containing the First Release of the ANC can be ordered from the LDC (LDC membership is not necessary) through the LDC website.. The First Release is a sampler of texts that have been automatically processed, and which therefore contain many inaccuracies (e.g., sentence boundary errors) that we hope to correct in the final ANC. However, the First Release provides an idea of the general format of texts the completed ANC, and we invite user comment concerning its format that can serve as input to development of the complete ANC. The composition of the First Release is given in Table 1 below. For full descriptions of the texts please see the ANC website.

Text type	Text name	No. of texts	No. of words	Contributor
Spoken	Callhome	24	50,494	LDC
Spoken	Switchboard	2320	3,056,062	LDC
Spoken	Charlotte Narrative	95	117,832	Project MORE
			TOTAL SPOKEN	3,224,388
Written	New York Times	4148	3,207,272	LDC
Written	Berlitz Travel Guides	101	514,021	Langensheidt Publishers
Written	Slate Magazine	4694	4,338,498	Microsoft
Written	Various non-fiction	27	224,037	Oxford University Press
			TOTAL WRITTEN	8,283,828
			TOTAL CORPUS	11,508,216

Table 1. Composition of the ANC First Release

A closer look

The ANC First Release CD contains two main folders: DOCS and DATA. The DOCS folder has a file listing of all the texts in the First Release and also a readme file with instructions for extracting the files and installing them on the Hard Drive.

The DATA folder is divided into four folders. The first folder, ANC, contains the Header for the corpus and a file listing the responsible persons for various aspects of the ANC and the publication information for the ANC First Release. The latter two files are used by the header program to supply that information when needed. The Schemas folder contains the XCES schema and supporting files.

The last two folders in the DATA folder, the Merged folder, and the Standoff folder, each contain a version of the ANC First Release. The Standoff folder contains the corpus in stand-off annotation form, where part of speech (POS) annotations are stored in documents separate from the primary data. There is the primary data file, the header file, and the POS annotated file. The Merged folder contains the corpus in merged form, with part of speech (POS) annotation included within the primary data. In each of these folders (i.e., Merged and Standoff) the corpus is organized in subdirectories with the various components of the corpus (e.g., Berlitz, NY Times) allowing researchers to easily address register specific linguistic questions. Each primary file has a header file associated with it that contains basic information such as subject, domain, ISBN numbers of books that are included, and information on the speakers.

The text sample below is an excerpt from the merged version of an article from the July 1, 2002 sports section of the NY Times (ANC file name = NYT20020701.0001). The excerpted phrase "the Rockets went into summer with seven free agents, including the greatest player in franchise history" is shown below with its POS and the lemmas. Each token is explicitly marked with <tok> tags, and part-of-speech and lemma are given as the codes of msd and base, respectively. For example, in line three in the sample below,

the original word in the text (*went*) is assigned the POS tag *vbd* identifying it as a past tense verb – msd = vbd. The lemma is indicated by the code *base*. In this case the automatic tagging correctly assigned the POS and also correctly lemmatized *went* with the base form *go*.

```
<tok msd="ati++++" base="the">the</tok>
<tok msd="nns++++" base="rocket">Rockets</tok>
<tok msd="vbd+++xvbn+" base="go">went</tok>
<tok msd="in++++" base="into">into</tok>
<tok msd="nn++++" base="summer">summer</tok>
<tok msd="in++++" base="with">with</tok>
<tok msd="cd++++" base="seven">seven</tok>
<tok msd="jj+atrb+++" base="free">free</tok>
<tok msd="nns++++" base="agent">agents</tok>
<tok base="," msd=",+clp+++">,</tok>
<tok msd="in++++" base="including">including</tok>
<tok msd="ati++++" base="the">the</tok>
<tok msd="jjt+atrb+++" base="great">greatest</tok>
<tok msd="nn++++" base="player">player</tok>
<tok msd="in++++" base="in">in</tok>
<tok msd="nn++++" base="franchise">franchise</tok>
<tok msd="nn++++" base="history">history</tok>
```

Having a Standoff and a Merged version of the corpus allows maximum flexibility for researchers. Those interested in computational issues can work with various encoded forms and also different versions. Corpus linguists are also able to search either the plain text or the linguistically annotated versions.

In conclusion, by having a corpus that is tagged and lemmatized, a range of linguistic questions can be addressed. The completed ANC will be a valuable resource for several reasons. First, it will be the first widely available large corpus of spoken and written American English containing a variety of registers. Second, the ANC will represent a synchronic slice of American English across many registers and therefore allow diachronic comparisons of changes in American English. Finally, since the ANC is following the same general framework as the BNC, it will be a rich resource for exploring differences between British and American English.

Bibliography

American National Corpus website: http://americannationalcorpus.org

Aston, Guy and Lou Burnard 1998. *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Biber, Douglas. 1988 *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas, Susan Conrad, and Randi Reppen. 1998. *Corpus Linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Garside, Roger, Geoffrey Leech and Geoffrey Sampson (Eds.) 1987. *The computational analysis of English: A corpus-based approach*. Harlow: Longman.

Fillmore, Charles, Nancy Ide, Dan Jurafsky, and Catherine Macleod. 1998. An American National Corpus: A Proposal. *Proceedings of the First Annual Conference on Language Resources and Evaluation*. Paris: European Language Resources Association, 965--969.

Ide, Nancy, 1998a. Encoding Linguistic Corpora. In *Proceedings of the Sixth Workshop on Very Large Corpora*. San Francisco: Morgan Kauffman, 9-17.

Ide, Nancy, 1998b. Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora. *Proceedings of the First International Language Resources and Evaluation Conference,* 463-70.

Ide, Nancy, Bonhomme, P., and Romary, L., 2000. XCES: An XML-based Encoding Standard for Linguistic Corpora. *Proceedings of the Second International Language Resources and Evaluation Conference*. Paris: European Language Resources Association, 825-30.

Ide, Nancy, Randi Reppen, and Keith Suderman. 2002. The American National Corpus: More than the Web can provide. In *Proceedings of the Third Language Resources and Evaluation Conference* (LREC). 839- 844. Las Palmas, Canary Islands, Spain.