The American National Corpus: A Standardized Resource for American English

Catherine Macleod*, Nancy Ide†, Ralph Grishman*

*Computer Science Department
New York University
New York, New York 10003-6806
{macleod, grishman}@cs.nyu.edu
†Department of Computer Science, Vassar College
Poughkeepsie, NY 12604-0520 USA
ide@cs.vassar.edu

Abstract

At the first conference on Language Resources and Evaluation, Granada 1998, Charles Fillmore, Nancy Ide, Daniel Jurafsky, and Catherine Macleod proposed creating an American National Corpus (ANC) that would compare with the British National Corpus (BNC) both in balance and in size (one hundred million words). This paper reports on the progress made over the past two years in launching the project. At present, the ANC project is well underway, with commitments for support and contribution of texts from a number of publishers world-wide.

Introduction

Linguistic research has become heavily reliant on text corpora over the past ten years. It is now widely recognized that for most applications, a sufficiently large corpus reflecting the full range of domains and usage is essential. For American English, freely available corpora that meet this requirement do not exist. The most recent balanced corpus of American English, the Brown Corpus, is not large enough to meet current needs; it contains only one million words, and, because it was created in 1960, does not reflect current usage. Although it is significantly larger (100 million words), the British National Corpus (BNC) has proved to be inadequate for language research targeted toward American English because of the substantial lexical and syntactic differences between British and American English (Algeo, 1988).

To meet the need for a corpus of American English, a proposal was put forward at the 1998 LREC conference to create a large, heterogeneous, uniformly annotated corpus of contemporary American English comparable to the BNC (Fillmore, et al., 1998). Over the past two years the project has developed, and a consortium of supporters including American, Japanese, and European dictionary publishers, as well as industry, has been formed to provide initial funding for development of the American National Corpus (ANC).

In May of 1999, representatives of several publishing houses gathered at the first ANC meeting at the University of California at Berkeley. At a subsequent meeting at New York University, which included those present at the May meeting as well as publishers from Japan and representatives from various U.S. and European software companies, the ANC consortium was formed, and the founding membership agreement was decided upon. At present, the creation of the ANC is underway, using texts contributed by some consortium members and supported by membership fees. The Linguistic Data Consortium, which will manage and distribute the corpus, is contributing manpower, software, and expertise to create a first version of the corpus, a portion of which should be ready for use by consortium members at the end of this year.

Makeup of the ANC

Our model for the ANC consortium and the format and contents of the ANC was originally based on the BNC. We intend to collect one hundred million words which will correspond closely in terms of textual balance to the BNC, in order to facilitate cross-lingual studies. However, the design of our corpus will take into consideration recent advances made in the field The design of the BNC, including both its internal format and the software to access and manipulate it, was developed over ten years ago, and significant developments in the technology of corpus annotation and representation dictate that our model differ somewhat from that of the BNC. Also, the needs of the research community together with other considerations have prompted us to plan for inclusion of a broader range of text types in the remainder of the corpus.

The main areas of application of the ANC are lexicography, both traditional and computational, and natural language processing (NLP), all of which require access to different domains and genres. For these purposes the corpus must be large (at least 100 million words), contemporary, heterogeneous, and uniformly annotated for a variety of linguistic features, and must contain only American English. The size ensures the adequate representation of infrequent words. The selection of contemporary texts is important for both lexicography and NLP, particularly in view of the significant changes in common text genres over the last few years brought about by electronic communication. Heterogeneity ensures that the range of language usage needed for the creation of "general language resources" is represented, and that one can explore a wide spectrum of language genres for NLP. Uniform annotation is paramount in any corpus; in particular, it is essential that both markup and annotation conventions follow emerging and existent standards. Annotation of basic linguistic and structural information,

including sentence and word boundaries and part of speech, is invaluable to both lexicographers and NLP researchers. Because the corpus will include only American texts, grammatical and lexical differences found in British English will not interfere with the classifying of American English.

The ANC will include primarily contemporary texts (1990 onward). The corpus will contain a *static* component and a *dynamic* component. The static component will comprise approximately 100 million words and will remain unchanged, thus providing a stable resource for comparison of research results as well as a snapshot of American English at the end of the millennium. This portion of the corpus will be comparable in balance to the BNC; although there is no set definition of "balance" in a corpus, we will follow the BNC criteria in terms of domain and medium[1] to enable cross-linguistic studies between British and American English. However, because the ANC will be comprised of contemporary texts, the ANC static corpus will overlap with only the second time period of the BNC.

A static corpus cannot keep up with current usage, and so the ANC will also include a *dynamic* component comprised of additional texts added at regular intervals. At present, we plan to add approximately ten- percent new material every five years in a layered organization, thus enabling access to all layers and the static core in chronological order. In this way, we hope to provide the advantages of both a static corpus such as the BNC and a dynamic corpus (e.g., the COBUILD corpus), while at the same time providing a resource for studies of change in American English over time.

Beyond the 100 million words comparable to the BNC, the ANC will also include additional texts from a wide range of styles and domains that will be varied rather than balanced; i.e., it will include smaller samples of a greater variety of texts rather than differing percentages of texts according to their representative importance in the language. To some extent the contents of this portion of the corpus will be dictated by availability: we hope to take advantage of the availability of large quantities of contemporary texts such as email, rap music lyrics, etc., as well as to add historically significant novels and other writings. Up to now, much NLP research has been focused on newspaper or newswire text, reflecting the availability of common corpora and annotated corpora in these areas. However, other genres are becoming not only common but also available in massive quantities, including unedited electronic data, email, web announcements and discussion groups, technical writing in computer manuals, help files and telegraphic reports. These genres differ in vocabulary, names and "named entity" structures (e.g., formulas, addresses, currencies, etc), syntax, lexical semantics, and discourse structure. It has been shown that adapting to genre-specific language can significantly improve analysis performance for syntactic structures and preferences (Sekine 1997) and for semantic or selectional preferences. A standard multi-genre corpus can foster research on genre adaptation, where some experiments can be conducted on raw text data and others can be effective with small amounts of syntactically-annotated data. A standard corpus will encourage common annotation and, we hope, attract funding for selective annotation.

[1] See the BNC User's Reference Guide (Burnard, 1995) for details of the criteria for balance in the BNC.

Encoding and Annotation of the ANC

The ANC will be encoded according to the specifications of the eXtensible Markup Language (XML) version of the Corpus Encoding Standard (XCES)[2] (Ide, Bonhomme, & Romary, 2000), which specifies a flexible document structure that is suitable for delivery on the World Wide Web, is easy to modify or add to, and allows for "layering" annotation and related documents that can be added incrementally at later stages. XML is the emerging standard for data representation and exchange on the World Wide Web (Bray, Paoli, & Sperberg-McQueen, 1998). Although at its most basic level XML is a document markup language directly derived from SGML (i.e., allowing element nesting and element references), various features and extensions of XML make it a far more powerful tool for data representation and access. For example, the eXtensible Style Language (XSL) provides a powerful transformation language (XSLT)[3] (Clark, 1999) that can be used to convert any XML document(s) into another XML document by selecting, rearranging, and adding information to it, in order to serve any application that relies on part or all of its contents.

The overall plan for development of the ANC is in two broad stages. In the first stage, a "base level" encoding (conformant to a Level 0 encoding as specified by the Corpus Encoding Standard (Ide, 1998a,b)) of the data will be provided, by automatically transducing original printer codes to XCES markup for gross logical structure (title, paragraph, etc.). Header information regarding target audience, text type, etc. will be inserted manually; at this stage, only minimal header information will be provided, based on the headers used in the BNC. This will allow us to test the applicability of the basic BNC header to our corpus at an early stage and give us the opportunity to tune it to the needs of the ANC in the fully annotated version. The base level annotation will be performed by the Linguistic Data Consortium at the University of Pennsylvania (UPenn).

Automatic part-of-speech tagging will also be performed on the base corpus, using the part-of-speech tags of the Penn TreeBank. At this stage, only spot checking of the data will be done; the object of this step is to harmonize the data to the extent possible using only automated means, thus avoiding the time and cost of hand-work. The resulting base level corpus should be sufficient for many needs (in particular, those of dictionary publishers), such as concordance generation. Software for viewing and analyzing the corpus data in this format will be made available to consortium members along with the data, although the data will also be available separately.

The second stage of development will be undertaken in parallel with the first, but is partly dependent on funding. In this stage, the corpus will be produced in its "final" form, with the goals of (1) marking as much

[2] See http://www.cs.vassar.edu/XCES.

[3] A description of and specifications for the current version of the XSL transformation language is available at http://metalab.unc.edu/xml/books/bible/updates/14.html.

information in the document as possible while providing for maximal search and retrieval capability, and (2) providing a "gold standard" corpus, consisting of some portion (possibly 10%) of the entire ANC, for use in natural language processing work for training, etc. To this end, at least the following tasks will be undertaken:

- Validation and refinement of existing markup, e.g., changing paragraph markers to more precise tags such as list, quote, etc., marking highlighted words for function (e.g., foreign word, emphasis, etc.);
- Provision of a full XCES-compliant header, including a full description of provenance and all encoding formats utilized in the document. In this phase we will correct, where necessary, categories and other header information drawn from the BNC in the first phase, and substantially add to it;
- Insertion of additional markup for sub-paragraph elements, such as tokens, names, dates, numbers, etc. Identification of these elements will, to the extent possible, be done automatically;
- Hand validation of markup for sub-paragraph elements, including sentence, token, names, dates, etc., in the "gold standard" portion of the corpus;
- Transduction of part of speech markup to XCES specifications, and possible transduction of annotation categories to a standard scheme such as the EAGLES morpho-syntactic categories (Monachini & Calzolari, 1996);
- Hand-validation of part-of-speech tags in the "gold standard" portion of the corpus;
- Implementation of the layered data architecture for annotations;
- Adaptation and/or development of search and retrieval software, together with development of XSLT scripts for common tasks such as concordance generation, etc.

The Consortium and Development Plan

Commercial members of the ANC consortium pay a membership fee over three years, which will be used to support the development of the base level corpus. In addition, publishers and other members are expected to provide contributions of data for inclusion in the corpus. Consortium members will receive the data as soon as it is processed and will have exclusive commercial rights to it for a period of three years.

All ANC data will be freely available to non-profit educational and research organizations from the outset (aside from a nominal fee for licensing and distribution). There will be no restrictions on obtaining the corpus based on geographical location; restrictions on the distribution of the BNC, which has so far been unavailable outside the European Union, have limited large-scale and comparative research based on the corpus. We hope to encourage comparative research by providing global access.

The Linguistic Data Consortium will obtain licenses from text providers and provide licenses to users. In general, the license will prohibit redistribution of the corpus and the publication or similar use of substantial portions of text drawn from the corpus without the permission of its original publisher. For dictionary makers, who comprise a large portion of the current consortium membership, usage of short portions of text in published dictionary examples etc. is allowed under legal definitions of "fair use".

We also plan to provide for an “open sub-corpus”, licensed to permit redistribution on the model of open-source software. The size of this corpus will be determined by the contributors.

Development of the Level 1 corpus and the "gold standard" sub-corpus will necessarily begin later than development of the base-level version, due to the need to secure substantial funding from external sources to support it. In addition, this development requires time for significant planning to ensure that the corpus is maximally usable by a broad range of potential applications and meets the needs of the research and industrial communities. We are currently soliciting input from the research community to feed this development. A meeting on the topic of annotation and encoding formats and data architectures for large corpora was held at this year's ANLP/NAACL conference in Seattle in early May; another more comprehensive workshop on the same topics was held preceding this LREC conference. By taking into account past experience, current and developing technologies, and user needs, we hope to be able to provide a state-of-the-art platform for universal access to the ANC.

Summary

The ANC, first proposed at the first LREC in 1998, is well on its way to realization. Within the year, the first data in its base level representation will be available to the NLP community and consortium members. The final corpus in its fully marked and annotated form should be available within three years.

A corpus of contemporary American English is a valuable resource not only for commercial applications and research, but also for educators, students, and the general public. It is also an important historical resource: the corpus will provide a "snapshot" of American English at the turn of the millennium, valuable for linguistic studies in the decades to come.

Acknowledgments

The first ANC meeting in Berkeley, California was funded by National Science Foundation grant ISI-9978422. We would like to thank Sue Atkins, Michael Rundell, and Rob Scriven for their support and for providing information concerning the creation of the BNC. We would also like to acknowledge the contribution of Wendalyn Nichols, Frank Abate, and Yukio Tono, who have been instrumental in obtaining the support of the publishing community.

References

Algeo, J., 1988. British and American Grammatical Differences. *International Journal of Lexicography.* 1:1-31.

Bray, T., Paoli, J., Sperberg-McQueen, C.M. (eds.), 1998. Extensible Markup Language (XML) Version 1.0. W3C Recommendation. http://www.w3.org:TR/1998/REC-xml-19980210

Burnard, L., 1995. British National Corpus: User's Reference Guide for the British National Corpus. Oxford: Oxford University Computing Service.

Clark, J., (ed.) 1999. XSL Transformations (XSLT). Version 1.0. W3C Recommendation. http://www.w3.org/TR/xslt

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

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

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

Ide, N., Bonhomme, P., Romary, L., 2000. XCES: An XML-based Standard for Linguistic Corpora. In *Proceedings of the Second Annual Conference on Language Resources and Evaluation*, this volume.

Monachini, M. & Calzolari, N., 1996. Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora: A Common Proposal and Applications to European Languages. EAGLES Report EAG-CLWG-MORPHSYN/R. http://www.ilc.pi.cnr.it/EAGLES96/morphsyn/.

Sekine, S., 1997. A new direction for sublanguage NLP. In D. Jones & H. Somers (eds.), *New Methods in Language Processing*, UCL Press, 165--177.

Sekine, S., 1997. The Domain Dependence of Parsing. In *Proceedings of the Fifth Conference on Applied Natural Language Processing.*