The American National Corpus:

A standardized resource for American English

Nancy Ide* and Catherine Macleod†

*Department of Computer Science
Vassar College
Poughkeepsie, NY 12604-0520 USA
ide@cs.vassar.edu

†Computer Science Department
New York University
New York, New York 10003-6806 USA
macleod@cs.nyu.edu

1 Introduction

Linguistic research has become heavily reliant on text corpora over the past ten years. Such resources are becoming increasingly available through efforts such as the Linguistic Data Consortium (LDC) in the US and the European Language Resources Association (ELRA) in Europe. However, in the main the corpora that are gathered and distributed through these and other mechanisms consist of texts which can be easily acquired and are available for re-distribution without undue problems of copyright, etc. This practice has resulted in a vast over-representation among available corpora of certain genres, in particular newspaper samples, which comprise the greatest percentage of texts currently available from, for example, the LDC, and which also dominate the training data available for speech recognition purposes. Other available corpora typically consist of technical reports, transcriptions of parliamentary and other proceedings, short telephone conversations, and the like. The upshot of this is that corpus-based natural language processing has relied heavily on language samples representative of usage in a handful of limited and linguistically specialized domains.

A corpus is intended to be "a collection of naturally occurring language text, chosen to characterize a state or variety of a language" (Sinclair, 1991). As such, very few of the so-called corpora used in current natural language processing and speech recognition work deserve the name. For English, the only true corpora that are widely available are the Brown Corpus (Kucera and Francis, 1967) and the British National Corpus (BNC) (Leech, 1994). Although it has been extensively used for natural language processing work, the million words of the Brown Corpus are not sufficient for today's large-scale applications. For example, for tasks such as word sense disambiguation, many word senses are not represented, or they are represented so sparsely that meaningful statistics cannot be compiled. Similarly, many syntactic structures occur too infrequently to be significant. The Brown Corpus is also far too small to be used for computing the bigram and trigram probabilities that are necessary for training language models used in a variety of applications such as speech recognition. Furthermore, the Brown corpus, while balanced for different written genres, contains no spoken English data. The 100 million words of the BNC provide a large-scale resource and include spoken language data; however, this corpus is not representative of American English and is so far available only within Europe for purposes of research. As a result, there is no adequately large corpus of American English available to North American researchers for use in natural language and speech recognition work.

To meet the need for a corpus of American English, a proposal was put forward at the 1998 Language Resources and Evaluation Conference (LREC) to create a large, heterogeneous, uniformly annotated corpus of contemporary American English comparable to the BNC (Fillmore, *et al.*, 1998). Over the past two and a half 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).

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, and 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.

2 Why we need a corpus of American English

There is a need for a corpus of American English that cannot be met by the data in the British National Corpus, due to the significant lexical and syntactic differences between British and American English. Well-known variations are: "at the weekend" (Br.) vs. "on the weekend" (U.S.), "fight (or protest) against <something>" (Br.) vs. "fight (or protest) <something>" (U.S.), "in hospital" (Br.) vs. "in the hospital (U.S.), "Smith, aged 36,…" (Br.) vs. "Smith, age 36…" (U.S.), "Monday to Wednesday inclusive" (Br.) vs. "Monday through Wednesday" (U.S.), "one hundred and one" (Br.) vs. "one hundred one" (U.S.), etc. Also, in British English, collective nouns like "committee", "party", and "police" have either singular or plural agreement of verb, pronouns, and possessives, which is not true of American English.

British English often makes use of a to-infinitive complement where American English does not. In the following examples from the BNC “assay”, “engage”, “omit” and “endure” appear with a to-infinitive complement, there were no examples found in a small corpus (comprised of selections from the Brown Corpus, the Wall Street Journal, the San Jose Mercury News, Associated Press, and the Penn Treebank) of this construction, although the verbs themselves did appear. For the first two verbs, one can argue that there is not an equivalent verbal meaning in American English, but, for the last two, the meaning can be paraphrased in American English by the gerund, as shown below. (Note that the British English examples are from the BNC and the American English examples are paraphrases.)

Verb	Eng.	Example sentences
assay	B.E.	Jerome crept to the foot of the steps, and there halted, baulked, rather, like a startled horse, drew hard breath and ASSAYED TO MOUNT, and then suddenly threw up his arms to cover his face, fell on his knees with a lamentable, choking cry, and bowed himself against the stone of the steps.
engage	B.E.	A magnate would ENGAGE TO SERVE with a specified number of men for a particular time in return for wages which were agreed in advance and paid by the Exchequer.
omit	B.E.	“What did you OMIT TO TELL your priest?”
	A.E.	“What did you OMIT TELLING your priest?’
endure	B.E.	But Carteret's wife, who frequented health spas, could not ENDURE TO LIVE with him or he with her: there were no children.
	A.E.	But Carteret's wife, who frequented health spas, could not ENDURE LIVING with him or he with her: there were no children.

Verb complementation containing prepositions often differs from British English to American English John Algeo (1988) gives a number of examples. In British English, “cater for” and “cater to” both occur, but “cater to” has a pejorative connotation and is less frequent. In American English, only “cater to” is used and is not considered pejorative. British English “claim for” contrasts with American English “claim” + NP (claim for benefits vs claim benefits), and, conversely, “agree” + NP is acceptable in British English but not in American English, which demands a preposition such as *upon, on, about,* or *to*. Algeo's example of British English “..yet he refused to agree the draw” would be “..yet he refused to agree to a draw” in American English Similarly, the bare infinitive after "insist", "demand", "require", etc. (e.g., "I insist he be here by noon.") is common in American English but rare in British English.

Adverbial usage is also different. The British English use of “immediately” in sentence initial position is not allowed in American English For example, British English “Immediately I get home, I will attend to that.” is incorrect in American English, in which one would say “As soon as I get home, I will attend to that.”

Other syntactic differences include formation of questions with the main verb “have”. In British English, one can say, “Have you a pen?” where American English speakers must use “do” (“Do you have a pen?”). Support verbs for nominalizations also differ : for example, the British English “take a decision” vs the American English “make a decision”.

There are also considerable semantic differences between the two brands of English: in addition to well-known variations such as lorry/truck, pavement/sidewalk, tap/faucet, presently (currently)/soon, autumn/fall, etc., there are numerous examples of more subtle distinctions, for example: "tuition" is not used to cover tuition fees in British English; "surgery" in British English is "doctor's office" in American English; "school" does not include higher education in British English, etc. Usage not only differs but can be misleading, for example, British English uses "sick" for the American "nauseous", whereas "sick" in American English is comparable to "ill" in British English; British "braces" are U.S. "suspenders", while "suspenders" in British English refers to something else entirely. Overall, the distribution of various semantic classes will also distort a British and an American corpus differently, for example, names of national institutions and positions (Whitehall, Parliament, Downing Street, Chancellor of the Exchequer, member of parliament, House of Lords, Royal Family, the queen, senate, president, Department of Agriculture, First Family, and heavy use of the word "state", etc.) and sports (baseball terms will be more frequent in an American corpus, whereas hockey--itself ambiguous between British and American usage--and soccer will predominate in the BNC). Idiomatic expressions also show wide variation between British and American English. Of course, spoken data between the two brands of English are not comparable at all.

The above comprise only a few examples, but it should be clear that when a uniquely British corpus is used, such examples skew the representation of lexical and syntactic phenomena. For applications that rely on frequency and distributional information, data derived from samples of British English are virtually unusable. The creation of a representative corpus of American English is critical for such applications.

3 Makeup of the ANC

Our model for the contents of the ANC includes, ideally, a balanced representation of texts and transcriptions of spoken data, as well as a large component of annotated speech data. In addition, the ANC should include representative samples for different dialects of American English (including Canadian English). Finally, samples of other major languages of North American, especially Spanish and French Canadian, should also comprise a portion of the corpus and, ideally, be aligned to parallel translations in English. However, in view of the high cost of collection and annotation of speech data, creation of a speech component of the ANC has been put off for the near future. Similarly, we are delaying attempts to provide comprehensive and balanced representation of regional dialects and collection of samples of North American Spanish and French. Our goals for the next few years of the project therefore include collection of only textual data and transcriptions of spoken data.

The ANC 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, unlike the BNC, the ANC will include primarily contemporary texts (1990 onward). The selection of contemporary texts is important for both lexicography and NLP, particularly in view of the significant changes in language usage over the last few years (e.g., changes brought about by electronic communication). Therefore, 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

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

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.

4 Encoding and annotation of the ANC

An American Natural Corpus will be most useful if it is more than just a collection of words. The corpora that have become most useful to both publishers and researchers in natural language and speech research have been those which are annotated. The paradigm example of this is the Brown Corpus, which has been the cornerstone of language-related research across disciplines in the United States, indeed in psychology as much as in natural language processing. Tagging of the Brown Corpus has played an essential role across disciplines, both in the original version (Kucera and Francis, 1967), and in the various on-line tagged versions, such as the Penn Treebank version (Marcus *et al*., 1993). For example, many modern part-of-speech taggers are trained on the Penn Treebank tagged corpus (see, for example, Brill, 1995). Part-of-speech tagged data has been used to automatically acquire sub-categorization dictionaries (Manning, 1993), in spell checking, and for applications that require partial parsing, etc. The syntactic parse trees that annotate the Brown Corpus in the Penn Treebank have played a similarly fundamental role in the training and evaluation of parsing systems.

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 final version. The base level encoding and annotation will be performed by the Linguistic Data Consortium at the University of Pennsylvania (Penn).

All texts in the both the base and final versions of the corpus will be marked for major structural divisions, paragraphs, and sentence boundaries, as well as part of speech. The base corpus will be automatically tagged, 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 the exact time-frame of the work is dependent on funding. In this stage, the corpus will be produced in its "final" form, with the goals of (1) marking as much information in the ANC 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. In the final corpus, annotation for linguistic phenomena (part of speech, syntactic annotation, etc.) will follow *de facto* standards such as those established by EAGLES.[2]

[2] It will be necessary to use a larger tagset than the Penn set in the final corpus. The Penn set was designed to be used with a corpus that was parsed, not merely tagged, and hence eliminates information that is only recoverable from a parsed corpus, such as the distinction between prepositions and subordinating conjunctions. These were

Encoding of the ANC will also adhere to international standards. The corpus will be encoded according to the specifications of the eXtensible Markup Language (XML) (Bray, *et al.*, 1998) version of the Corpus Encoding Standard (XCES)[3] (Ide, *et al.*, 2000), part of the Guidelines developed by the Expert Advisory Group on Language Engineering Standards (EAGLES).[4] XCES was developed expressly to serve the needs of corpus-based work in language engineering applications by providing XML encoding conventions for various linguistic phenomena in text and speech, as well as several types of linguistic annotation. Because XCES is an XML application, its use for encoding the ANC guarantees that access to and use of the corpus will be supported by tools and mechanisms designed for data delivered via the World Wide Web.

The XCES specifies a flexible document structure that provides for "layering" annotation and related documents that may be added incrementally at later stages. The separation of linguistic annotation in distinct documents facilitates retrieval from different annotations (including variants of the same kind of annotation--e.g., part of speech analysis by several taggers). This strategy of “remote” or “stand-off” markup is well-suited to the XML environment, due especially to the development within the XML framework of the Extensible Style Language (XSL). XSL provides a powerful transformation language (Clark, 1999) that can be used to create new XML documents from one or several others by selecting, rearranging, modifying and adding information to it. Thus, a user of a corpus encoded following the XCES model need not be aware of the underlying document architecture, and will see only a new document containing all and only the information he or she is interested in, in any desired configuration and encoding. This will enable use of the ANC for a potentially limitless set of applications, including not only computational linguistics research but also education, etc. The XCES architecture also provides for distribution of development and enhancement, by enabling different sites to develop separate documents containing annotations for the primary ANC data, all of which are ultimately linked together and retrievable as a hyper-document.

In the second stage of ANC development, 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 and 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.

5 The ANC consortium

Founding consortium members contribute US$21,000 over 3 years in annual installments of $7,000, 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

---

combined into the single tag *IN* in the Penn tagset, since the tree-structure of the sentence disambiguated them (subordinating conjunctions always precede clauses, prepositions precede noun phrases or prepositional phrases).

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

[4] http://www.ilc.pi.cnr.it/EAGLES/home.html

members who join after March 31, 2001, contribute $40,000 in two annual installments. Consortium members receive the data as soon as it is processed and have exclusive commercial rights to it for a period of five years after the date of the first release of data, currently anticipated to begin at the end of this year. Current consortium members are:

- Pearson Education
- Random House Publishers
- Langenscheidt Publishing Group
- Harper Collins Publishers
- Cambridge University Press
- LexiQuest
- Microsoft Corporation
- Shogakukan,Inc.
- Associated Liberal Creators Press
- Taishukan Publishers
- Oxford University Press
- Kenkyusha Publishers
- International Business Machines Corporation

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 last year's ANLP/NAACL conference in Seattle in early May[5]; another more comprehensive workshop on the same topics was held preceding the LREC conference in Athens in June, 2000 (Broeder, *et al.*, 2000). 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.

6 Summary

The ANC, initially proposed at the first LREC in 1998, is now 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.

[5] Papers available at http://www.cs.vassar.edu/~ide/ANLP-NAACL2000.html.

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 CM (eds) 1998 Extensible markup language (XML) Version 1.0. W3C recommendation. http://www.w3.org:TR/1998/REC-xml-19980210.

Brill E 1995 Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21(4):543-566.

Broeder D, Cunningham H, Ide N, Roy D, Thompson H, Wittenburg P (eds) 2000 *Proceedings of the EAGLES/ISLE Workshop on Meta-Descriptions and Annotation Schemas for Multimodal/Multimedia Language Resources and Data Architectures and Software Support for Large Corpora*. Paris, European Language Resources Association.

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

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*, Granada, pp 965-969.

Ide N 1998a Encoding linguistic corpora. In *Proceedings of the Sixth Workshop on Very Large Corpora*, Montreal, pp 9-17.

Ide N 1998b Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora. *Proceedings of the First International Language Resources and Evaluation Conference*, Granada, pp 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*, Athens, pp 825-30.

Kucera H, Francis W 1967 *Computational analysis of present-day American English*. Providence, Brown University Press.

Leech G, Garside R, Bryant M 1994 CLAWS4: The tagging of the British National Corpus. *Proceedings of COLING-94*, Nantes, pp 622-628.

Manning C 1993 Automatic acquisition of a large subcategorization dictionary from corpora. *Proceedings of ACL*, Columbus, pp 235-242.

Marcus M, Santorini B, Marcinkiewicz 1993 Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2): 313-330.

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 The domain dependence of parsing. In *Proceedings of the Fifth Conference on Applied Natural Language Processing.*

Sinclair J 1991 *Corpus, concordance, and collocation.* Oxford, Oxford University Press.