An American National Corpus: A Proposal

Charles J. Fillmore
International Computer Science Institute
University of California, Berkeley
Berkeley, CA 94704 USA
fillmore@icsi.berkeley.edu

Daniel Jurafsky
Department of Linguistics
University of Colorado at Boulder
Boulder, CO 80309-0295 USA
jurafsky@colorado.edu

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

Catherine Macleod
Computer Science Department
New York University
New York, NY 10003 USA.
macleod@cs.nyu.edu

Abstract

This paper proposes the development of an American National Corpus comparable to the British National Corpus. Corpus-analytic work has demonstrated that the use of the British National Corpus is inappropriate to study of American English, due to the numerous differences in the use of the language. We also propose that the corpus include a component of texts in other major North American languages, notably Spanish and French, and ideally a parallel component containing texts in these languages aligned to the English. The development of an ANC will demand a significant commitment from the funding agencies and the research community; such an effort would, however, significantly contribute to language and linguistic research as well as the U.S. National Digital Libraries Initiative and other large-scale projects.

1. Introduction

The need for large-scale corpus resources for natural language and speech research is well established. 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 (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 for speech recognition. Furthermore, the Brown corpus, while balanced for different written genres, contains no spoken English data. The 100 million words of the British National Corpus provide a large-scale resource and include spoken language data; however, this corpus is not representative of American English and is furthermore 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.

We propose the development of an American National Corpus of text and speech comparable to the British National Corpus. We suggest three criteria for such a corpus. First, it must be broad, i.e., both large and well balanced. Second, it must be deep and hence more than just a collection of words; it should, for example, be annotated with lemma and part-of-speech information and sentence-boundaries. Finally, it must be American, i.e., it must reflect not only American English but also the French and Spanish of North America.

2. The Need for Balance

In order to be representative of any language as a whole, it is necessary that a corpus include samples across a variety of texts that reflect the range of syntactic and semantic phenomena across that language. Balance among resources is especially crucial for tasks such as lexicon building. For example, COMLEX Syntax, a large syntactic dictionary developed at New York University under the auspices of the Linguistic Data Consortium used a large "unbalanced" corpus to create entries.1 Because the corpus consisted predominantly of newspaper data, statistical information was skewed, resulting in an unrepresentative preponderance of particular phenomena at the same time as others were under-represented. For example, there is a disproportionate number of complex NUNITP complements for some verbs, which appear in sentences typical of newspaper style, such as "The price rose two percent to 102 dollars per share from 100 dollars per share". This type of complement was shown to exhibit a significantly different distribution in the Brown Corpus (Macleod, *et al.*, 1998), which represents a range of texts and is therefore more representative of linguistic usage; however, in general the Brown Corpus is too small to provide adequately large samples for the purposes of lexicon construction.

Similar problems have arisen in work on word sense disambiguation, which has relied heavily on newspaper data for its samples: it has been noted that for some typical test words such as "line", certain senses (for example, the common sense of "line" as in the sentence, "He really handed her a line") are absent entirely from resources such as the *Wall Street Journal.*

The problem of balance is acute in speech recognition. Speech recognition systems are notoriously dependent on the characteristics of their training corpora. Corpora large enough to train the trigram language models of modern speech recognizers (many tens of millions of words) are invariably composed of written rather than spoken texts. But the differences between written and spoken language are even more severe than the differences balanced corpora like the Brown and newspaper corpora like the *Wall Street Journal*. Therefore, whenever a state-of-the-art speech recognition research effort moves to a new domain, a new large training corpus of speech must be collected, transcribed at the word level, and the transcription must be aligned to the speech.

3. The Need for 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. For example, phrases such as "omit to", "endure to", etc. are common in British usage but occur only in highly constrained collocations in American English. Other similar 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, American usage typically involves the gerund (e.g., "omit paying", etc. vs. the British "omit to pay"). British usage also differs in phrases such as "take a decision" vs. the American "make a decision". 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. 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 U.S. English.

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 very 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, which 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.

4. The Need for an Annotated Corpus

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 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. Part of this is because the Brown corpus has been lemmatized and tagged. Lemmatization (Francis and

[1] The corpus was composed of the Brown Corpus (7 MB), the *Wall Street Journal* (part 1, 8.5 MB), the *San Jose Mercury News* (30 MB), the Associated Press (29.5 MB), the *Wall Street Journal* (part 2, 18.5 MB) and miscellaneous texts, including literature (1.5 MB).

Kucera, 1982) has made it possible for the researcher to explore the role of the stem independently of the inflected form. 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). While it is possible to train part-of-speech taggers (such as hidden Markov model taggers) without hand-labeled training data, Merialdo (1994) has argued that using hand-labeled data results in better performance. Part-of-speech tagged data has been used to automatically acquire subcategorization dictionaries (Manning, 1993), in numerous applications in speech recognition, 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.

Annotated speech corpora have played an important role as well. The TIMIT corpus (Lamel *et al.,* 1986, Fischer *et al.,* 1987) of hand-annotated sentences provided original training data used in most if not all of the speech recognition laboratories in the nation, as well as significantly contributing to studies of American pronunciation (Withgott and Chen, 1993, etc.). Recently, the small *Switchboard* database (Godfrey *et al.,* 1992) and others such as *CallHome* have begun to play part of the role in spoken language processing that the Brown Corpus has played for written language processing. Four hours of Switchboard have been hand-annotated with phonetic transcriptions (Greenberg *et al.,* 1996). Approximately 1.4 million words of Switchboard has been tagged and parsed, has also been hand-segmented into turns and utterances (Meteer *et al.,* 1995) and annotated with dialog-act tags (Jurafsky *et al.,* 1997) for the next release of the Penn Treebank. While by modern standards this is far too small an amount of data, it shows that the same annotations that were applied to written data can successfully be applied to spoken data, as well as others.

One of the strongest arguments for an American National Corpus is the possibility that different sites, each funded in different ways, could contribute annotations at different levels, as was the case for *Switchboard*. The availability of a single corpus that could be augmented with annotations at different linguistic levels from different sites would be invaluable.

5. Composition of the ANC

An American National Corpus should be developed with an eye toward serving the needs and interests of research across a wide range of areas, including not only computational linguistics, but also lexicography, speech recognition and synthesis, literary studies, and all varieties of linguistics.

The American National Corpus should comprise at least 100 million words of data; ideally, it should be considerably larger than this. Like the British National Corpus, the proposed American National Corpus should be designed to cover as wide a range of written modern American English as possible, including national and regional newspapers, specialist periodicals and journals for all ages and interests, academic books and popular fiction, published and unpublished letters and memoranda, essays, etc. The corpus should also contain a substantial spoken component, comprising 10-15% of the corpus and including transcriptions of formal and informal speech by speakers of differing ages, regions, and social classes.

The spoken component should include speech signal for the transcribed data, aligned to the orthographic transcription, and ideally including phonetic and prosodic annotation as well as annotation for dialog acts (similar to the Switchboard data--Godfrey et al., 1992; see also Greenberg et al., 1996; Meteer et al., 1995; Jurafsky et al., 1997). The natural speech corpora that are currently widely available *(Switchboard, CallHome, CallFriend),* besides being small and often domain-specific, are not closely aligned with their text transcriptions, making it very difficult to use these corpora for any research purpose which requires the careful correlation of speech and text. The availability of a single corpus, which could be augmented with annotations at different linguistic levels from different sites, would be invaluable.

We also propose that a portion of both the written and spoken portions of the corpus be composed of texts in other major languages of North America, in particular, Spanish and Canadian French. At least a portion of this data should comprise parallel translations aligned to the English and/or the other languages in the sample.

All texts in the corpus, including written texts and transcriptions of speech, should be marked for major structural divisions, paragraphs, and sentence boundaries, as well as part of speech annotation and alignment (where applicable). Speech data should be segmented and marked for turn and utterance.

In order to be maximally usable, the corpus should be encoded and annotated using agreed-upon international standards. We therefore propose to encode the corpus using the Corpus Encoding Standard (CES) (Ide, 1998), which was developed expressly to serve the needs of corpus-based work in language engineering applications. It provides encoding conventions suitable for encoding various linguistic phenomena in speech and text as well as for various kinds of linguistic annotation. The CES also defines a data architecture that allows for the separation of linguistic annotation (such as part of speech tagging and alignment information) in distinct documents, thus facilitating the layering of and retrieval from different annotations (including variants of the same kind of annotation--e.g., part of speech analysis by several taggers). This architecture would enable a distribution of development and enhancement, as mentioned above in section 4, by enabling different sites to develop separate documents containing particular encoding or annotations, ultimately linked together and retrievable as a hyper-document.

Annotation for linguistic phenomena, such as part of speech, syntactic annotation, etc. should follow *de facto* standards such as those established by the Penn Treebank and EAGLES. The choice of tagset is essential; it would be, for example, necessary to develop a larger tagset than the Penn set used for the Brown 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 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). Similar modern standards for encoding various speech phenomena are under development (e.g., the SABLE project for speech synthesis).

Much of the markup and annotation of a corpus of the size we propose will be done automatically, and will therefore contain errors. We recommend that at least a portion of the encoding and annotation of the data be hand-validated, providing a benchmark corpus that can be used for training, etc. Ideally, the markup for common sub-paragraph level elements (for example, names, dates, abbreviations, etc.) should also be hand-validated and, where possible, enhanced for greater precision (for example, by replacing a simple `<name>` tag with `<name type=person>`, etc.).

5. Implementation

The development of an American National Corpus is a vast undertaking and will require substantial resources and support. It will in particular be essential that the major North American agencies provide funding for the effort. The effort will also demand the cooperation of publishers who will necessarily have to provide at least a portion of the texts to be included and agree to distribution rights. Here we may follow the example of the British National Corpus, which involved a consortium of publishers who contributed directly to the effort.

Most importantly, the development of an American National Corpus will require the commitment of the research community. No one site will be able to perform all of the work involved; it will be necessary to develop a consortium of sites who work together to reach the common goal. The four main areas of activity will involve data collection, encoding, annotation, and distribution. A particular site or group of sites working in collaboration may address each of these tasks.

In addition to the development of the annotated corpus itself, it will be necessary to develop software to enable the retrieval and use of the corpus by a wide community of users. We see this part of the effort as fully compatible with the goals of established initiatives and projects, such as the U.S. Digital Libraries Initiative and various NSF and other government-sponsored projects. We feel it is essential for some mechanism to be established to identify software development methodologies that can be applied across the board so that data and software are maximally compatible and therefore, reusable. With such collaboration, it is likely that software for retrieval and manipulation of the ANC data, as well as for automated encoding and annotation, can be potentially developed within the scope of existing projects.

We believe it is essential for funding agencies to initiate actions to promote and enable inter-project collaboration and the establishment of methods and standards for such software development. A framework for software development can be established which will enable distributed development among different sites. Only in this way can we minimize the incompatibility and redundancy that now characterizes the development of software for creation and manipulation of textual data, and avoid paying several times over for the same work.

6. Conclusion

The American National Corpus will play a vital and wide-ranging role in speech and natural language processing for this next century, contributing directly to the U.S. National Digital Libraries Program, the National Science Foundation's focus on Human-Centered Computing, and many other government and corporate projects and fundamental scientific advances in augmentative communication, spelling and grammar checking, speech recognition, topic detection, and message understanding.

We propose that this corpus be made freely available to North American researchers for use in language-related work, including not only computational linguistics but also literary studies, social science research, etc. If the corpus is well designed, it can ultimately be a part of a National Digital Library that is accessible to educational institutions. Overall, its development should provide an invaluable contribution to research and education.

References

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.

Francis, W. Nelson Francis & Kucera, Henry (1982). *Frequency Analysis of English Usage*. Boston, MA: Houghton Mifflin.

Godfrey, J., E. Holliman, & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. *Proceedings of ICASSP-92*, San Francisco, 517-520.

Greenberg, S., Ellis, D. & Hollenback, J. (1996). Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus. *Proceedings of ICSLP-96*. Philadelphia. ftp://ftp.icsi.berkeley.edu/pub/real/dpwe/

Fisher, W. M., Zue, V., Bernstein, J., and Pallet, D. (1987). An Acoustic-Phonetic Data Base. *Proceedings of the 113th Meeting of the Acoustical Society of America.*

Ide, Nancy (1998). The Corpus Encoding Specification: SGML Guidelines for Encoding Linguistic Corpora. *First International Language Resources and Evaluation Conference*, Granada, Spain (this volume). See also http://www.cs.vassar.edu/CES/

Jurafsky, Daniel, Elizabeth Shriberg, and Debra Biasca (1997). The Switchboard DAMSL (SWBD-DAMSL) Labeling System. *1997 LVCSR Summer Research Workshop Technical Reports*, Johns Hopkins University (to appear).

Kucera, Henri & Winthrop Francis (1967). *Computational Analysis of Present-Day American English*. Brown University Press, Providence, RI.

Lamel, L. F., Kassel., R. H., & Seneff, S. 1986. Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus. In Baumann, L. S. (Ed.), *Proceedings of the February 1986 DARPA Speech Recognition Workshop,* 100-109.

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

Merialdo, Bernard (1994). Tagging English Text with a Probabilistic Model. *Computational Linguistics,* 20.2 155-172.

Macleod, Catherine, Grishman, Ralph, & Meyers, Adam (1998). Dictionaries and Balanced Corpora: The Interdependence of Resources. *First International Language Resources and Evaluation Conference*, Granada, Spain (this volume).

Manning, Christopher D. (1993). Automatic Acquisition of a Large Subcategorization Dictionary from Corpora. *Proceedings of ACL,* Columbus, OH, 235-242.

Marcus, M., Santorini, B., & Marcinkiewicz (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics,* 19:2, 313-330.

Meteer, Marie (1995). Dysfluency Annotation Stylebook for the Switchboard Corpus. Linguistic Data Consortium. Revised June 1995 by Ann Taylor. ftp://ftp.cis.upenn.edu/pub/treebank/swbd/doc/DFL-book.ps.

Sinclair, John (1991). *Corpus, Concordance, and Collocation.* Oxford University Press, Oxford.

Sperberg-McQueen, C.M., and Lou Burnard, Eds. (1994). *Guidelines For Electronic Text Encoding and Interchange*. ACH-ACL-ALLC Text Encoding Initiative, Chicago and Oxford.

Withgott, M.M. & Chen, F.R. (1993). *Computational Models of American Speech.* Stanford: CLSI.