The American National Corpus: More Than the Web Can Provide

Nancy Ide,* Randi Reppen,† Keith Suderman*

* Department of Computer Science
Vassar College
Poughkeepsie, New York 12604-0520 USA
{ide, suderman}@cs.vassar.edu
† Department of English
Northern Arizona University
Flagstaff, Arizona USA
reppen@nau.edu

Abstract

The American National Corpus (ANC) project is developing a corpus comparable to the British National Corpus (BNC), covering American English. Recent interest in the web as a source of corpus materials has caused some in the language processing community to suggest that the development of a corpus of American English is unnecessary. However, we argue that far from being rendered superfluous by the availability of web materials, the ANC is likely to provide a resource for developing web acquisition techniques to support tasks such as genre and language detection and automatic annotation. This paper presents a comparison of the ANC in terms of both content and format with a test corpus compiled from web data, and a discussion of points of intersection and divergence.

1. Introduction

The American National Corpus (ANC) project is developing a corpus comparable to the British National Corpus (BNC), covering American English (Fillmore, et al., 1998; Ide & Macleod, 2001). The project is funded by a consortium of publishers of American English dictionaries and companies with interests in language processing. Consortium members are providing materials for inclusion in the corpus. The Linguistic Data Consortium and the European Language Resource Association (ELRA) are managing the distribution of the ANC.

The ANC will contain a core corpus of at least 100 million words, comparable across genres to the BNC, but containing only texts produced after 1990. Beyond this, the corpus will include an additional component of potentially several hundreds of millions of words, chosen to provide both the broadest and largest selection of texts possible. A first installment of the corpus of 10 million words is scheduled for release on September 1, 2002.

Initially, the corpus will contain only textual data across a variety of genres, including transcriptions of spoken data. Audio speech data, video, etc. will be added in a later phase. All data will be distributed freely for non-commercial research purposes from the outset. Commercial use will be limited to members of the ANC Consortium throughout the development process and for five years after the first installment of the corpus becomes available.

We plan to include as a part of the ANC a "gold standard" portion, comprising 10% of the final 100 million word core, which has been hand-validated and corrected for structural markup (paragraphs, etc.) and part of speech tagging. We feel strongly that a standardized, validated corpus of American English representing a balanced cross-section of genres will provide an invaluable tool for research in language use and the development of language processing tools. However, funding to support the development of a gold standard corpus has been hampered by the claim that the availability of enormous quantities of language samples on the World Wide Web eliminates the need for the ANC in general, and for a gold standard corpus in particular. We have taken this argument seriously; the experiment reported in this paper addresses this issue directly, and at the same time explores the feasibility of collecting samples of American English from the web for inclusion in the ANC.

2. Status of the ANC

In the initial phase of the ANC, we identified texts held by consortium members which can be contributed to create a balanced corpus. Identification of appropriate texts was based, first, on the date of publication so as to include only texts produced after 1990. Text selections were also filtered on the basis of authorship: only those texts written by persons born or educated (and currently living) in the U.S. were included. We have also obtained rights to include various spoken materials in the ANC, including the Switchboard (Godfrey *et al.,* 1992) and CallHome[1] corpora, as well as materials contributed from various projects (e.g., the Linguistic Atlas project at the University of Georgia).

We are currently transducing the contributed texts from their original formats (Quark Express, PDF, SGML, etc.) to an XML format compliant with the EAGLES/ISLE XML Corpus Encoding Standard (XCES) (Ide et al., 2000). In this first phase, all transduction is being done automatically; therefore, in some cases the resulting markup will be imprecise (e.g., for italicized words within paragraphs, lists marked only with paragraph separators, etc.). All formatting information in the original is being retained to enable later refinement. Morpho-syntactic annotation is being automatically added to the entire corpus, using a tagger developed by Douglas Biber at Northern Arizona University (Biber, et al., 1998). Part of speech tags produced by the Biber tagger are compliant with the CLAWS tags (although they include

[1] http://www.ldc.upenn.edu/Catalog/LDC97S42.html

additional information in some cases), thereby enabling cross-linguistic comparisons using the BNC and the ANC. Headers for each text in the corpus are being created semi-automatically.

We are also developing a web-based interface for access to the corpus, which will also be distributed on CD ROM together with access software for various platforms.

The first release of 10 million words of the ANC will be released in September, 2002. The data in this release will be tagged for part of speech, and software to perform basic extraction tasks, including generation of concordances and collocation information, will be included.

3. Why not a web corpus?

Recent interest in the web as a source of corpus materials has caused some in the language processing community to suggest that the development of a corpus of American English is unnecessary. There are, however, several significant differences between a corpus compiled of web materials and the ANC, the most obvious and important of which is the difficulty of compiling a corpus of exclusively American English drawn automatically from the web. In addition, the core ANC will include extensive texts representing a balance of genres. It is not at all clear that a web corpus can be balanced for genre, and it is likely that certain genres, such as fiction will be under-represented. The ANC will be coherently marked up for structure and annotated for morpho-syntax and shallow syntactic structure; it is our plan to validate at least 10% of the corpus in order to provide a "gold standard" for further work. We argue that far from being rendered superfluous by the availability of web materials, the ANC is likely to provide a resource for developing web acquisition techniques to support tasks such as genre and language detection and automatic annotation.

To validate our claim, we performed an experiment in which we gathered a corpus of texts from the World Wide Web, in an attempt to answer two fundamental questions:

1. Does the availability of materials extracted from the web render the need for a corpus of American English superfluous?
2. Is it possible to devise automatic methods to draw materials from the web for inclusion in the ANC?

4. The Experiment

To validate our argument, we set up an experiment in which we attempted to gather materials from the World Wide Web which would be suitable for inclusion in the ANC. In order to select web texts for the experiment, we identified the following criteria:

1. The texts should be representative of American English;
2. The texts should include spans of prose long enough to serve the purposes of tasks such as concordance-making and extraction of meaningful collocates, and be suitable for various kinds of linguistic analysis (part of speech tagging, syntactic analysis, discourse analysis);
3. The texts should represent a variety of genres and authors;
4. The texts must not be bound by any copyright restrictions;
5. The texts must have been produced after 1990.

It should be obvious that satisfying these criteria using materials drawn from the web presents several problems, not all of which are easily addressed. Therefore, for the purposes of exploring the possibility of using web data, we adopted a strategy of "rough approximation", wherein we attempted to identify a large body of materials of which a high percentage would be likely to satisfy our criteria.

Identifying American English vs. British, Canadian, Australian, or any other brand of English without knowledge of the author's background is perhaps the least straightforward of our criteria to satisfy. The obvious means is to select texts in which American spelling conventions (and possibly, syntactic conventions) are followed; however, even by this measure, a given text cannot be guaranteed to reflect American usage. Our solution for this experiment was to take texts only from web sites with a `.gov` or `.edu` suffix. Although this still does not guarantee American English, we assumed that most of the materials will fall into this category because these sites are almost exclusively located in the U.S.[2] We also assume that all web texts have been produced after 1990.

Using texts from `.edu` and `.gov` sites also contributed to satisfying the copyright criterion, since most educational and government web materials are not likely to be bound by copyright restrictions (government documents may in fact be required to be in the public domain by law). Materials from sites such as CNN, Newsweek, etc. would have been ideal possibilities for gathering texts in American English, but such sites are, in general, restricted by copyright and therefore unsuitable for our purposes. Organizational (.org) sites provided another possible source of unrestricted materials, but these sites, which exist all over the world, cannot be guaranteed to include materials in American English.

At the same time, materials extracted from `.edu` and `.gov` sites, while representing prose produced by a variety of different authors, skews the range of domains that are represented in the data. Government documents, in particular, are likely to reflect a consistent and somewhat idiosyncratic prose style. This may be offset to some degree by materials from `.edu` sites: here, we are likely to see a broader range of styles (for example, consider course materials across the full range of academic disciplines, students' home pages, etc.). We do, however, believe that even if we broadened our selection criteria to include .com and other sites (even if it were possible to identify those including American English), the range of genres would likely not be meaningfully diversified.

Fiction is likely the genre which is most under-represented in web documents. We attempted to address this by looking at sites identified by a search for

[2] We recognize the larger question of whether the ANC should attempt to determine what is "American English" *a priori,* or whether the data should drive the definition. This nonetheless leaves us with the task of determining authorship, whatever the definition of an "American author" may be taken to be.

"American novel", "American short story" etc. However, many of the sites we examined include materials whose copyright restrictions are, at best, unclear: often, such sites reproduce published materials or are themselves copyrighted by the author of the web documents in which they are included. Identifying potentially copyright-free documents would therefore involve considerable effort, which for this preliminary experiment we did not pursue.

To ensure texts of adequate length, we considered only pages with 2000 or more words (excluding tags, and determined solely on the basis of white space separation). However, length alone was not sufficient to ensure that the documents contained appropriate text samples. Web documents frequently consist primarily of tables, bulleted lists of cryptic prose, links, etc. and in many cases contain few or no stretches of running prose. To address this problem, we computed the ratio of words/paragraph for each document retrieved from the web, by dividing the number of words in the document (excluding tags) by the number of "paragraphs". The number of paragraphs was computed by counting the number of *out-of-line* HTML tags—i.e., tags which generate a break to a new line, as opposed to *in-line* tags such as **<i>**, **<a>**, **<b>**, etc.[3] It was not possible to use the HTML ** ** (paragraph) tag to determine the number of paragraphs because they are very often used in web documents for presentational rather than descriptive purposes (insertion of blank lines, delimiting list items, etc.).[4] More to the point, tag usage in web documents is highly unreliable in terms of content description: header tags and list item tags often delimit what would normally be regarded as simple prose paragraphs. Despite the roughness of our measure, manual inspection of the data showed that we could reliably *extract* web pages that include stretches of running prose by applying the rough measure based on out-of-line tags (although we do, on occasion, eliminate pages that would otherwise be suitable for inclusion in our test corpus).

To determine the appropriate cut-off point for discarding web documents, we used a simple but powerful scripting language for web data, *WebL*[5], to harvest 9757 web documents. From this data, we extracted only those documents containing 2000 or more words, which yielded a corpus of 542 web pages and a total of 2.9 million words (about 5% of the harvested data). We then plotted the distribution of words/paragraph for these texts. The resulting graph is shown in Figure 1. Manual examination of a sample of the corpus revealed that in fact, most of the pages whose words/paragraph ratio was 30 or below (66% of the pages) did not contain significant stretches of prose and were therefore unsuitable for inclusion in our corpus.[6] Thus, only 1-2% of the sites visited satisfied our criteria for total length and paragraph length.

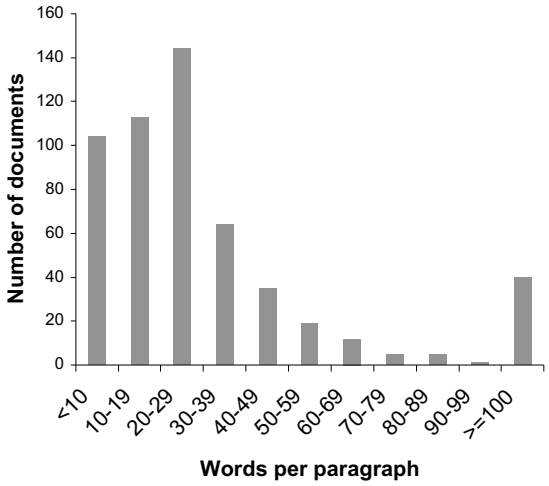

Figure 1 : Words/paragraph in the unfiltered corpus

Using *WebL*, we then extracted a corpus of 2.2 million words from the web, using the following criteria:

- Page name must end in `.htm` or `.html` to avoid automatically generated pages (`.php`, `.asp`, etc.) that are unlikely to contain text;
- Domain must be `.edu` or `.gov`;
- Page must contain 2,000 or more words and 30 or more paragraphs.

Table 1 summarizes the extraction results. Of the texts extracted, 20 were eliminated from the final test corpus because they consisted entirely of tables, figures, addresses, etc.; were duplicates; or (in two instances) were in Spanish, yielding a final corpus of 2,125,240 words.

The corpus was automatically tagged for part of speech together with over seventy additional linguistic features, using the Biber tagger (Biber, et al., 1998). The tagged texts were then fed to a program developed at Northern Arizona University (Biber, 1988; Conrad & Biber, 2001) that classifies texts over five textual "dimensions":

1. *Involved vs. Informational Production*
2. *Narrative vs. Non-Narrative Discourse*
3. *Situation-Dependent vs. Elaborated Reference*
4. *Overt Expression of Argumentation*
5. *Non-Abstract vs. Abstract Style*

The score for each dimension is determined by co-occurrence patterns for over seventy linguistic features; for example, the score on Dimension 1 is computed using weighted scores for each of the following features:

> *Positive features:* private verbs, THAT deletion, contractions, present tense verbs, 2nd person pronouns, DO as pro-verb, analytic negation, demonstrative. pronouns, general emphatics, first pers. pronouns, pronoun IT, BE as main verb, causative subordination, discourse particles, indefinite pronouns, general hedges, amplifiers, sentence relatives, WH questions, possibility

[3] The full list of in-line tags is `tt, i, b, u, s, strike, big, small, em, string, dfn, code, samp, kbd, var, cite, acronym, a, img, applet, object, font, basefont, script, map, q, sub, sup, span, bdo, iframe, input, select, textarea, label, button.`

[4] In fact, programs such as MS Word that automatically generate HTML typically include sequences such as **   ** to generate white space.

[5] http://www.compaq.com/WebL

[6] Note that the large number of documents containing 100 or more words/paragraph results from the inclusion of pages in which the entire text was contained in **<pre>** tags, with little or no other markup.

modals, non-phrasal coordination, WH clauses, final prepositions

Negative features: nouns, word length, prepositions, type/token ratio, attributive adjectives.

(See Biber, 1988 for a complete list of features computed for each dimension.) Dimension 1 can be interpreted as an "oral – literate" continuum: texts at the upper end are highly involved (e.g., face-to-face conversations) containing a significant number of the positive linguistic features for this dimension; texts at the lower end reflect careful production and packaging of information (high percentage of nouns, long words, a high type/token ratio, etc.). The five dimensions and associated features and weights were determined based on extensive textual analysis (Biber, 1995), and have since served as the basis for a range of text and register analyses and comparisons, including stylistic analysis (Connor-Linton, 2001); gender comparisons (Biber and Burges, 2001; Rey, 2001); and diachronic changes across registers (Biber & Finegan, 1992).

Pages visited	16,270
Pages collected	456
Percent retained	2.8
Words collected	2,190,196
Average words/page	4,803
Average paragraphs/page	81

Table 1: Automatic collection statistics

To gain better insight into the range of text types and features included in our web corpus, we manually categorized all of the web documents into four main groups:

- **Institutional documents :** manuals, handbooks, statements of policies and procedures, etc.
- **Expository :** letters, essays, etc.
- **Science :** scientific reports, etc.
- **Prepared speeches :** transcripts of delivered speeches

Table 2 summarizes the text categorization.

Register	# of texts	# of words
Science	153	739,474
Expository	185	858,250
Planned speech	18	118,661
Institutional	80	408,855

Table 2 : Web texts by register

Dimension scores were computed for each group and plotted against known scores for a variety of other text types, including telephone conversations, face-to-face conversations, personal letters, spontaneous speeches, prepared speeches, general fiction, professional letters, broadcasts, editorials, academic prose, press reportage, and official documents.

5. Results

Table 3 gives the scores across the five dimensions identified by Biber (1988) for web texts and for several other text types. Transcriptions of spoken texts are given in capital letters; texts drawn from the web are given in bold. Table 4 shows the text types sorted by their scores for each of the five dimensions (web texts are highlighted in gray). As the table shows, the written web-based texts appear at the lower end of all five dimensions. In general, this implies that the web texts are more informationally dense and elaborated in their content than texts from other sources. However, the web text scores are comparable to those for texts which are similar in genre (official documents, academic prose).

The one clear anomaly in the data is web-based "planned spoken" texts (pl), which consist of transcriptions of formal speeches. Web-based speech transcriptions are, apparently, significantly denser informationally and more "literary" than their paper-based equivalents (Figure 2). Just why this is so is not clear; further investigation of the linguistic features contributing to the dimension scores for web and non-web based speeches may provide some insight.

The conclusion we can draw from the dimension scores is that in general, texts taken from the web represent a particular type of prose—in particular, a formalized, dense type of prose characteristic of formal documents. It makes intuitive sense that materials produced for the web would not exhibit characteristics of informal or even argumentative prose (e.g., editorials); it is likely that even if pages from additional types of sites were included, the result would be roughly similar. Therefore, in spite of the similarities found among the web written texts and their paper-based counterparts, there remain major gaps in the range of texts that would be needed to construct a representative and balanced corpus.

The web is obviously not a source of face-to-face conversation or other spoken interactions[7], but it also not a source of the range of written texts that readers frequently encounter. As such, web texts lack the variety and distribution of linguistic features that can be found in many texts. In addition, our data suggest that web texts differ from much standard prose in their rhetorical structure: the average length of a web "paragraph" is about 50 words[8], whereas the average in other text types is often much higher (for example, the average paragraph length in this paper is well over 100 words). As a result, studies of American English based on web materials alone are likely to be inappropriate for tasks such as dictionary and lexicon creation, language teaching, etc. Because web texts do not seem to include significant amounts of informal prose, they may be even more inappropriate for development of language processing applications dealing with user input.

[7] Materials from chat rooms and newgroups may provide a source of informal prose, but because of the restrictions of the medium, they do not exhibit many of the characteristics of transcriptions of spoken data (overlaps, pauses, hedges, etc.).

[8] Paragraph length was re-computed to eliminate headers, etc., and their content in order to obtain this statistic.

TEXT TYPE	1	2	3	4	5
General fiction	-2	6	3.8	1	0
SPONTANEOUS SP	18	1.5	-1.2	0.2	2.5
FACE-TO-FACE SP	35	-0.5	4	-0.2	3.2
TELEPHONE CONV	38	-2	5.2	0.5	3.8
Personal letters	20	0.2	3.7	1.5	2.8
BROADCASTS	2	-3.2	9	-4.5	1.8
PLANNED SPEECH	2.5	0.8	-0.2	0.5	2
Planned speech	**-17**	**-2**	**-7**	**-0.8**	**-1**
Press reportage	-15	0.5	0.2	-0.8	-0.5
Professional letters	2.5	-2.2	-6.8	3.6	-0.3
Academic prose	-15	-2.5	-4.4	-0.7	-5.6
Official documents	-18	-2.8	-7.5	0	-4.8
Editorials	-10	-0.8	-1.8	3	-0.3
Science	**-25**	**-3.5**	**-4**	**-4.5**	**-1.2**
Expository	**-20**	**-3.5**	**-5**	**-0.8**	**-1.6**
Institutional	**-18**	**-3.5**	**-8**	**0**	**-2.5**

Table 3 : Dimension scores for web and other texts

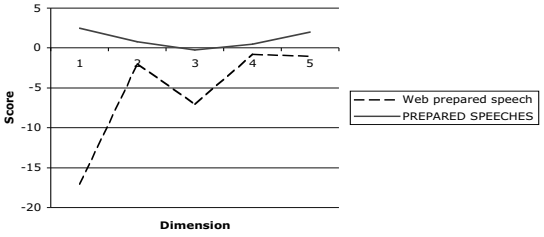

Figure 2 : Web vs. other planned speech

DIMENSION 1		DIMENSION 2		DIMENSION 3		DIMENSION 4		DIMENSION 5	
TELEPHONE	38	TELEPHONE.	38	BROADCASTS	9	Prof. letters	3.6	TELEPHONE	3.8
FACE-TO-FACE	35	FACE-TO-FACE	35	TELEPHONE	5.2	Editorials	3	FACE-TO-FACE	3.2
Pers. letters	20	Pers. letters	20	FACE-TO-FACE	4	Pers. letters	1.5	Pers. letters	2.8
SP. SPEECH	18	SP. SPEECH	18	Fiction	3.8	Fiction	1	SP. SPEECH	2.5
PLANNED SP.	2.5	PLANNED SP.	2.5	Pers. letters	3.7	TELEPHONE	0.5	PLANNED SP.	2
Prof. letters	2.5	Prof. letters	2.5	Reportage	0.2	PLANNED SP.	0.5	BROADCASTS	1.8
BROADCASTS	2	BROADCASTS	2	PLANNED SP.	-0.2	SP. SPEECH	0.2	Fiction	-
Fiction	-2	Fiction	-2	SP. SPEECH	-1.2	Official Docs	0	Prof. letters	-0.3
Editorials	-10	Editorials	-10	Editorials	-1.8	Institutional	0	Editorials	-0.3
Reportage	-15	Reportage	-15	Science	-4	FACE-TO-FACE	-0.2	Reportage	-0.5
Acad. Prose	-15	Acad. Prose	-15	Acad. Prose	-4.4	Acad. Prose	-0.7	PLANNED SP.	-1
PLANNED SP.	-17	**PLANNED SP.**	-17	Expository	-5	Reportage	-0.8	Science	-1.2
Official Docs	-18	Official Docs	-18	Prof. letters	-6.8	PLANNED SP.	-0.8	Expository	-1.6
Institutional	-18	Institutional	-18	PLANNED SP.	-7	Expository	-0.8	Institutional	-2.5
Expository	-20	**Expository**	-20	Official Docs	-7.5	BROADCASTS	-4.5	Official Docs	-4.8
Science	-25	**Science**	-25	Institutional	-8	Science	-4.5	Acad. Prose	-5.6

Table 4 : Text types sorted by dimension scores

6. Future Work

We are continuing to investigate the characteristics of texts harvested from the web, in order to determine more precisely the characteristics of the "web text" genre. Our next steps include more detailed consideration of the linguistic features that contribute to the various dimension scores. We will also look at syntactic properties of the texts, to determine the degree to which usage in web texts reflects that of texts from other sources and representing other genres. In addition, we hope to broaden the range of web materials we gather to include chat rooms, newsgroups, etc., but the problem remains of identifying those that include strictly American English.

The web materials that we have gathered so far are similar to other texts of similar genre, but they appear to be more cryptic and terse. Paragraph length, and possibly sentence length as well, seem to be considerably shorter in the web materials. We are currently gathering statistics on paragraph length in non-web texts and segmenting the web materials into sentences, in order to more precisely assess the differences.

We are also continuing to assess the feasibility of automatically gathering web texts for inclusion in the ANC. Our experiment so far suggests that development of a fully automated procedure for gathering web texts appropriate for inclusion in the ANC may not be possible. We noted several problems in the previous section, including the difficulty of identifying texts written by speakers of American English, copyright, and the general difficulty of selecting texts which satisfy our criteria for length and contiguous stretches of running prose. Without manual examination of the harvested texts, it is likely impossible to ensure that texts harvested from the web are appropriate for the ANC. However, we are seeking means to minimize the amount and extent of manual intervention by refining our selection criteria. For example, by fine-tuning the words/paragraph requirement, both in terms of minimum and maximum values. we may be able to ensure that a high percentage of the harvested texts contain appropriately long stretches of prose. We may also be able to use various linguistic filters to identify appropriate texts; this is, however, dangerous, since we do not want to pre-determine the characteristics of a representative corpus of American English; rather, these characteristics should be derived from a representative sample.

7. Summary and Conclusion

Texts drawn from the web exhibit characteristics that are similar, but not identical, to other text types, suggesting that they can be regarded as falling into a genre of their own. In particular, written web materials contain dense, information-packed language that is also found in official documents and academic prose. However, they also appear to be more cryptic and terse, containing shorter paragraphs than those found in paper-based materials. Our study suggests that web-based texts are, in any case, representative of only a small slice of the range of genres encountered by human readers everyday, and therefore cannot be used to provide a comprehensive view of American English in the 1990's.

A drawback of our approach is the limited range of web sites we included in the study, and the possibility that our data reflect only a small range of the types of materials that can be found on the web. In order to broaden the range of text types we consider, however, we would have to find some way to reliably identify American English in web texts, which is far from a straightforward task.

Our experiment demonstrates that gathering a corpus of materials from the web requires considerable work—some of it manual—if the materials are to be useful for many language-analytic tasks. To avoid large samples of dubious materials, such as pages consisting entirely of links or tables (which comprised a significant proportion of the pages we examined), it is necessary to identify and program precise selection criteria. Even then, the harvested texts must be rendered in a form that is amenable to analysis. Given the high variability in the ways that HTML tags are used in web pages, it is nearly impossible to identify structural components such as headers, paragraphs, etc. We are therefore unconvinced that creation of a web corpus is a simple matter, precluding the work involved to create the ANC.

8. References

Biber, D. & Finegan, E. (1992). The linguistic evolution of five written and speech-based English genres from the 17th to the 20th centuries. In M. Rissanen, O. Ihalainen, T. Nevalainen & I. Taavitsainen (eds.) *History of Englishes: New methods and interpretations in historical linguistics*. Amsterdam: Mouton de Gruyter, 688 – 704.

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Biber, D., & Burges, J. (2001). Historical shifts in the language of men and women: Gender differences in dramatic dialogue. In S. Conrad & D. Biber (eds.) *Variation in English: Multi-Dimensional studies*. Harlow, Essex: Pearson Education, 157 – 170.

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press

Connor-Linton, J. (2001). Author's style and world view: A comparison about American nuclear arms policy. In S. Conrad & D. Biber. (eds.). *Variation in English: Multi-Dimensional studies*. Harlow, Essex: Pearson Education, 84 – 93.

Conrad, S., & Biber, D., eds. (2001). *Variation in English: Multi-Dimensional studies*. Harlow, Essex: Pearson Education.

Fillmore, C., Ide, N., Jurafsky, D., & Macleod, C. (1998). An American National Corpus: A Proposal. *Proceedings of the First International Language Resources and Evaluation Conference,* Granada, Spain, 965-70.

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

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, 825-30.

Ide, N. & Macleod, C. (2001). The American National Corpus: A Standardized Resource of American English. *Proceedings of Corpus Linguistics 2001*, Lancaster UK.

Rey, J. (2001). Changing gender roles in popular culture: Dialogue in Star Trek episodes from 1966 to 1993. In S. Conrad & D. Biber (eds.). *Variation in English: Multi-Dimensional studies*. Harlow, Essex: Pearson Education, 138-156.