Executive Summary

Corpus linguistics comprises a set of empirical methods for research on language. Central to this enterprise is the construction of the corpus itself: a collection of texts that ideally stand in for a language as a whole. The animating principle behind this is corpus *representativeness*. The concept emerges from several decades of debate regarding the role of empiricism in linguistic research, and it encompass a range of practical and theoretical concerns from sampling to social context. In sum, a representative corpus aims to enable generalizations about a language as a whole.

The value of corpus linguistics to WE1S is twofold. First, the principle of corpus representativeness — with some modification — can guide of the construction of WE1S corpus. The goal of the corpus under this paradigm would be coverage of the *range* of discussions about the humanities. In order to capture this, prior research would be done into the social settings in which the humanities is discussed (e.g. newspapers, television interviews on higher education), and texts would be drawn from these unique contexts.

Second, several publicly available research corpora overlap with anticipated segments of WE1S. In terms of corpus analysis, these other corpora provide an opportunity to validate findings from the project and strengthen scholarly claims. In terms of distribution and public access, they also offer potential models for public-facing components of the WE1S corpus, especially regarding the legal question of fair use.

Citation: Teddy Roland. “Corpus Linguistics Research Report: Representativeness.” WhatEvery1Says Project, 4Humanities.org. July 18, 2017.

Corpus Linguistics

History & Context

Within the larger field of Linguistics, *corpus linguistics* conceives of itself as a methodology.[1] Broadly speaking, the highest level goal of Linguistics research is to develop a model of a human language, while the branches of the field are devoted to different problems or features of language. For example, large bodies of research are devoted to second-language acquisition or phonology. Rather than describing a particular set of problems, however, corpus linguistics offers a method for answering a variety of theoretical questions through the observation of large bodies of texts. McEnery & Wilson describe this method as primarily *empirical*, in distinction to the *rationalist* methods that have dominated Linguistics since the 1950s.[2]

The use of empirical methods in Linguistics has a long history. Whereas the term *corpus linguistics* refers to a particular set of practices that coalesced in the early 1990s, they have clear precedents reaching back at least to the nineteenth century. Many corpus linguists cite F.W. Käding's 1897 study of letter frequencies, for which a workforce of five thousand Prussian analysts tallied these in a collection of documents totaling eleven million words.[3] Through the middle of the twentieth century, the use of written records was a common source of evidence and their empirical study underpinned the linguist's claim to scientific standing.[4]

These methods became relatively marginal during the second half of the twentieth century. Since the 1950s, Chomskyan theories and research practices have enjoyed great popularity in linguistics. As they relate to this discussion, these theories emerged partly as a critique of earlier corpus-based methods. For example, just one of Chomsky's critiques asserted that the space of possible linguistic production is infinitely large, whereas the size of a given corpus will always necessarily be finite. As a consequence, this means that the corpus at hand will always be skewed.[5] As a result of this and other criticisms, corpus-based linguistic research was more or less displaced by rationalist methods, in which an expert native speaker mentally evaluates the wellformedness of given linguistic constructions.

It would overstate the case to suggest that corpus methods disappeared entirely from Linguistics between the 1950s and 1980s. Some areas of research necessarily continued to rely on observational evidence and records, such as childhood language acquisition since the child is thought not yet to possess metalinguistic awareness.[6] Other linguists refined corpus methods in response to the criticisms raised by Chomsky.[7] This is the period that produced, for example, the Brown Corpus and Lancaster-Oslo-Bergen (LOB) Corpus, which sought to represent the English language as it appeared in publication during 1961 in the US and Britain respectively. These corpora continue to be used and emulated today.

Corpus linguistics enjoyed a renaissance with increasing access to personal computers during the 1990s. A review of an edited volume of conference proceedings from 1990

expresses this clearly. The reviewer notes that two of the papers had been dedicated explicitly to the topic of managing corpora on personal computers, and he concurs that "it is now possible for almost anyone to use large corpora. You no longer need an expensive computer center to look at a concordance."[8] Personal computers had removed a significant bar to entry in the field during the 1980s, especially by comparison to the mainframe computers that had been used previously. This is reiterated in McEnery & Wilson a decade later: "Whatever the philosophic advantages we may eventually see in the corpus, it is the computer which allows us to exploit corpora on a large scale with speed and accuracy." Theorizing corpus linguistics on firmer, post-Chomsky grounds was not enough to bring it into the mainstream: this would depend on "[t]he marriage of machine and corpus."[9]

Corpus Methods

Since the 1990s, corpus linguistics has comprised three major research activities: corpus, concordance, and statistics. This triad frames the title of John Sinclair's 1991 book *Corpus, Concordance, Collocation* — where the last term indicates his disciplinary commitment to discovering relationships between linguistic objects through their frequency distributions. Similarly, it is not uncommon to find textbook chapters devoted to each of the three topics, such as in McEnery & Hardie 2012 or Biber et al 1998.

The latter terms — concordance and statistics — describe different modes of analysis within the corpus. The concordance presents the researcher with a comprehensive listing of a phenomenon of interest. For example, one may wish to observe every instance of a given part of speech in the corpus, and each instance is typically shown within a contextual window for interpretability. Statistical analysis, especially frequency distribution, enables researchers to determine how common features are compared to one another and the strength of relationships between different phenomena. Corpus linguists do not see these techniques as competing with one another but as complementary, providing different forms of evidence.

The corpus itself has been intensely theorized in recent decades. The guiding principles of modern corpus construction go under the heading of *representativeness*. One important definition, from Douglas Biber, articulates this as "the extent to which a sample includes the full range of variability in a population," with the goal of offering a "basis for generalizations concerning a language as a whole."[10] The research goal of producing a generalized model of language is clear in this definition, where variation in usage is precisely what one hopes to observe. An appropriately representative corpus will contain phenomena of interest to the researcher, as well as their variations, in the same proportions as they appear in the broader language type under study. Ideally, this enables the researcher to reach beyond the finite bounds of the corpus that had been the subject of critique.

In order to ensure representation of linguistic variability, the researcher begins with the external (i.e. non-linguistic) criteria of the social situations in which language is

produced. For example, a private conversation versus a televised interview between the same participants is likely to cover different subject matter and syntax, yet a model of a given language as a whole would have to account for speech patterns in both of these situations. When compiling the corpus, sociological research precedes textual acquisition in order to identify the conditions of production for texts that contain the language type of interest. This has the consequence of articulating bounded sets of text that are eligible for inclusion in the corpus.[11]

In corpus linguistics terminology, situationally-defined categories of texts are referred to as *registers* and each of these may be *stratified* into subgroups. In the example above, both personal and televised conversations occur in the register of spoken language. However they belong to different sub-strata: *private* and *public* speech. Particular registers of texts and their strata are not universally useful for all corpora. Rather these are chosen to reflect the research priorities of a given project. Atkins et al 1992 offer a taxonomy of registers that illustrates how each is typically stratified when employed. Biber 1993 attempts to schematize these hierarchically as general categories. See the appendices to this article for those schema.

Brown Corpus Stratification, with lettered categories & number of texts			
Format	Publication: English, US, 1961		
Factuality	Informative (374)		Imaginative (126)
Topic	A. Reportage (44) B. Editorial (27) C. Reviews (17)	D. Religion (17) E. Skills & Hobbies (36) F. Popular Lore (48) G. Belles Lettres (75) H. Miscellaneous (30) L. Learned (80)	K. General (29) L. Mystery (24) M. Science Fiction (6) N. Adventure (29) P. Romance (29) R. Humor (9)
Channel	Periodical	Periodical, Book, Other	Novel & Short Story

Table: Brown Corpus stratification scheme. Note that each row further stratifies the categories of the row above. Lettered categories designate the major categories of the corpus, however these were compiled through a process of stratification by format. In the case of Reportage (A), this was also stratified by news paper sections and Miscellaneous (B) by institution responsible (government, business, etc). Texts were sampled from each of these strata individually and collected under the lettered topics.

To demonstrate the process of stratification, we may look for example to the Brown Corpus. This corpus aimed to represent language in the register of US publication in 1961. That register was initially stratified in terms of factuality: texts were categorized as informative or imaginative. In turn, these were further stratified by subject matter and finally by their avenue of publication. Based on this stratification, the researchers produced 15 topic-based categories covering the range of published texts, including

Press: Reportage, *Press: Editorial*, *Fiction: General*, *Fiction: Mystery*, and so on. Each category's importance in the universe of American English publications was determined by a subjective measure. More important categories received a larger number of texts in the corpus than others.[12]

In practice, the strata of a corpus are operationalized as circumscribed populations of texts from which random samples are drawn. Returning to the Brown Corpus, each of the fifteen categories of texts roughly corresponds to a subject heading in the card catalog of the Brown University Library and the Providence Athenaeum. From these catalog listings, a random selection was made by the researchers. In this sense the "universe of publication" is operationalized as the holdings of these libraries. Perhaps a more evenhanded example is the LOB corpus. It used the same categories but drew from bibliographies of periodical and book publications, rather than potentially biased library acquisitions.[13]

Corpus Construction

This, finally, is the rough template used by Biber and others for producing a representative corpus. Researchers begin by determining the social conditions in which a type of language is produced. Conditions of textual production are organized initially by their register and then increasingly stratified by their major sub-groups. This produces several discrete categories of texts within the corpus, each of whose prominence is determined by prior theoretical research on its relative importance. Each category is operationalized as a circumscribed list of texts. The list is sampled randomly, and the results are collected as the preliminary corpus.

Biber and Atkins et al both emphasize that, in reality, this process is cyclical.[14] Once a preliminary corpus of texts has been collected, the researcher assesses the documents based on their content (*internal* features of the text) and potentially adjusts the categories or the balance among them. For example, a rare linguistic feature may compel the researcher to increase the size of a small category in order to produce more observations for statistical robustness. At a more intuitive level, one may simply find that the documents are skewed in a way that was originally unexpected. These adjustments are integral to corpus development if it is to have broad usefulness in research.

What the WhatEvery1Says Project Can Learn

- Evaluation: adequate coverage over *range* of real-word discussions on the humanities
- Identify registers: research social situations where humanities discourse occurs
- Stratify the corpus: registers sub-divided into small number of discrete categories; categories become basis for text collection
- Balance the corpus: each category is represented in proportion to its importance
- Parallel corpus: create a mirror corpus of non-humanities articles (same balance)
- Legalize the corpus: emulate fair-use policies of existing, public corpora

Before discussing the application of corpus linguistic methods to WE1S, there is an important distinction to make. Whereas corpus linguistics aims to model a language type as a whole, WE1S aims to model public discourse on the humanities. The fact that WE1S relies on an internal feature of the text — its narrowly defined topic — as a condition for inclusion in the corpus indicates that the idea of representativeness needs critical examination before it can be imported from corpus linguistics.

The practical goal of representativeness had been the expression of a "full range of variability" in a language type. Translating this as a consideration of subject matter, perhaps we revise the concept so that it accounts for the range of ways that an article can talk about the humanities. The test of representativeness for the WE1S corpus would be an evaluation of whether it adequately represents different arguments and rhetorical frames surrounding the humanities.

Research in corpus linguistics indicates that the greatest variability in texts' internal features occurs across lines of social setting. In fact, WE1S has already laid the groundwork for such exploration. The "Statement of Corpus Expansion" from the 2017 grant proposal includes discussion of sub-corpora that include government documents, professional association reports, academic mission statements, and academic research articles. Each of these constitutes an important and unique register of discourse on the humanities. WE1S may wish to consider including spoken registers, as well, such as political speeches and television interviews concerning higher education. Beyond these, directed research in the field of linguistics may identify additional registers of public discourse (as well as resources from which these can be collected).

The stratification of texts, from newspapers as well as other registers, would enable WE1S to encode its research priorities in the structure of the corpus. Again, from the "Statement," WE1S intends to stratify texts by demographics of newspaper readership in order to learn for example about discussions of the career goals of first-generation college students. Other proposed strata include region/nation of publication. It may be desirable to use strata such as factuality or headline subject matter as well. Breaking out the corpus into sub-groups enables WE1S to explicitly indicate the contribution of each textual category to humanities discourse.

Stratification constitutes the basis for sampling methods in corpus linguistics, since each stratum is operationalized as a finite list of texts. Yet sampling raises serious questions

in its application for WE1S. To wit, is representativeness (sampling) preferable to completeness (totality) when creating the corpus? The advantage of sampling is that it has a strong claim to reach beyond the finite bounds of the corpus. Rather than drawing all articles from a few prominent newspapers, the corpus would consist of a few articles from a wide range of newspapers. In any case, the ability to sample may be constrained by the availability of a complete bibliography of humanities articles, as well as limitations on access to a range of periodicals. Whether WE1S will sample from newspapers (or any other register) depends on theoretical as well as practical considerations.

Stratification also grounds questions about balance in the corpus, which may prove a more useful application. One research goal of WE1S is to identify diachronic trends in discussion of the humanities. Insofar as it is possible, then, it is desirable to have the same distribution of texts across sub-strata for each year or decade (or another chosen unit of time). That is, strata of texts must have the same balance. The same may be said of the geographic coverage of the corpus. In each case, stratification is a practical method for ensuring similarity of distribution.

One further area where balance would be necessary is the collection of texts representing non-humanities public discourse. In order to identify unique features of humanities discourse, such a collection would be instrumental. Practically, the WE1S corpus could be modeled on the structure of bilingual corpora that are said to be *parallel*. The distribution of texts across strata in one language is mirrored in the other. In this case, categories would be balanced according to criteria regarding humanities discourse and non-humanities texts simply sampled in matching proportions.

After these decisions have been made about corpus construction and texts have been collected, WE1S will then proceed with analysis and public distribution. At that point, extant corpora offer significant guidance, particularly the Corpus of Contemporary American English (COCA). COCA aims to represent American English as it was used 1990-2015 and includes newspapers and other publications.[15] Perhaps most importantly for WE1S, the corpus is hosted online by Brigham Young University (along with several other major corpora, including the British National Corpus) with a public user interface.[16]

COCA public interface gives WE1S the opportunity to extend its own scholarly research and to fulfill its public humanities mission. Minimally, WE1S will be able to validate its findings by comparing trends found in the WE1S corpus to those in COCA. This may be as simple as comparing frequencies of keywords, but such a procedure would enable WE1S to determine how far its claims reach in other discursive contexts. In the realm of legal concerns, COCA has been made publicly accessible and includes a web interface, despite its containing copyrighted material. Indeed, the website has a page devoted to its fair-use standing and anticipated legal defense should a suit be brought against it.[17] The legal issues that WE1S faces regarding end-user access to the corpus have a clear precedent in COCA and other similar corpora.

Conclusion

As a mature field of research, corpus linguistics sharpens the questions that we are able to ask about the WE1S text collection process. What are the various social situations in which humanities discourse is produced? How do such texts break out into sub-groups and what do we believe is the relative importance of each? These questions constitute a pathway that connects prior research on the humanities and public discourse to the structure of the corpus itself.

The goals of corpus linguistics and WE1S remain distinct, however, in ways that constrain its application to this project. Rather than representativeness *per se*, balance may become the priority for the WE1S corpus. Longitudinal axes of inquiry have been proposed for the corpus, crossing nations and time periods. Appropriately balancing the corpus across those lines will ensure the robustness of our findings. In theory and example, corpora produced for corpus linguistics will offer guidance during the construction of the WE1S corpus.

As a final note, the resemblance between corpus linguistics and distant reading is striking. Methodologically, this is entirely the case thanks to shared points of intellectual formation, such as the founding of the journal *Literary and Linguistic Computing*. Institutionally, as well, both methods are similarly situated within their home disciplines, somewhat marginalized but enthusiastically advocated by their practitioners. The contentious yet long-standing role of empiricism in Linguistics and Literature departments has strongly contoured each of these methods as they are deployed today, and the complementary roles of statistical patterns and expert interpretation are frequently emphasized.

One current debate regarding the validity of distant reading for literary study concerns appropriate corpus construction. Without rehearsing this debate, suffice it to say that the positions taken echo those among linguists as late as the 1990s, which led to some of the resolutions described in this article. Seated as it is in the English department, WE1S has a unique opportunity to intervene on this literary argument with the benefit of several decades of hindsight from corpus linguistics theory and practice.

Appendix 1

Below is an abridged version of the "Taxonomy of Text Types" that Atkins et al offer in their article "Corpus Design Criteria." The full list given in the article is not meant to be exhaustive and the authors admit there there is potential overlap among categories. However, it effectively demonstrates the animating logics of register and stratification.

- Spoken
  - Dialogue
    - Private
      - Face-to-face conversation
      - Distanced conversation
    - Public
      - Broadcast discussion/debate
      - Legal proceedings
  - Monologue
    - Commentary
    - Unscripted speeches
- Written
  - To be spoken
    - Lectures
    - Broadcasts (news, documentary)
  - Published
    - Periodicals
      - Magazines
      - Newspapers
      - Journals
    - Books
      - Fiction
      - Non-fiction
        - (Auto-)biography
        - Educational textbooks
    - Miscellaneous
      - Leaflets
      - Adverts
  - Unpublished
    - Letters
      - Personal
      - Business
    - Memos
    - Minutes of meetings

Appendix 2

Biber attempts to roughly formalize the priorities that are expressed at each level of such a taxonomy in Table 1 of his article “Representativeness in Corpus Design.” For example, public and private speech can be thought of analogous to published and unpublished essays. Several of these analogies run throughout the kind of taxonomy given above, and Biber arranges them hierarchically.

1. *Primary channel*. Written/spoken/scripted speech

2. *Format*. Published/not published (+ various formats within ‘published')

3. *Setting*. Institutional/other public/private-personal

4. *Addressee*.
    (a) Plurality. Unenumerated/plural/individual/self
    (b) Presence (place and time). Present/absent
    (c) Interactiveness. None/little/extensive
    (d) Shared knowledge. General/specialized/personal

5. *Addressor*
    (a) Demographic variation. Sex, age, occupation, etc.
    (b) Acknowledgement. Acknowledged individual/institution

6. *Factuality*. Factual-informational/intermediate or indeterminate/imaginative

7. *Purposes*. Persuade, entertain, edify, inform, instruct, explain, narrate, describe, keep records, reveal self, express attitudes, opinions, or emotions, enhance interpersonal relationship, . . .

8. *Topics . . .*

Note that the degree to which topic or subject matter, as an internal feature of the text, can be used for stratification remains a matter of debate. Corpus linguists prefer external criteria, since they typically constitute the greatest axes of variability in a language type and because they enable linguistic findings to be connected to social questions. Organizing texts by their internal features requires interpretation on the part of the researcher that potentially negates these goals. In light of this, Atkins et al recommend stratifying texts like newspaper articles under the broadest headings possible when necessary. For example, the researcher might stratify articles by section (e.g. politics, sports). More recently Sinclair 2005 has argued against the use of topic for stratification, although his reasoning resembles that of Atkins et al.

In Biber’s terms the Brown Corpus, which discussed in this article, is organized hierarchically by Format, Factuality, Topic, and Primary Channel.

Works Cited

Atkins, Sue, Jeremy Clear, Nicholas Ostler. “Corpus Design Criteria.” Literary and Linguistic Computing 7:1 (1992): 1-16. doi: 10.1093/llc/7.1.1

Biber, Douglas. "Representativeness in Corpus Design." Literary and Linguistic Computing 8:4 (1993): 243-257. doi: 10.1093/llc/8.4.243

Biber, Douglas, Susan Conrad, & Randi Reppen. *Corpus Linguistics: Investigating Language Structure and Use*. New York: Cambridge University Press. 1998.

Church, Kenneth Ward. “Theory and Practice in Corpus Linguistics.” (Review). *Computational Linguistics* 17:1 (1991): 99-103.

Davies, Mark. “The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, Architecture, Linguistic Insights.” *International Journal of Corpus Linguistics* 14:2 (2009), 159–190. doi: 10.1075/ijcl.14.2.02dav

Francis, W. N. & H. Kucera. *Brown Corpus Manual: Manual of Information to accompany* A Standard Corpus of Present-Day Edited American English*, for use with Digital Computers*. Revised edition. Providence, RI: Brown University, Department of Linguistics. 1979. Accessed: July 18, 2017. http://clu.uni.no/icame/manuals/BROWN/INDEX.HTM

Gries, Stefan Th. "What is Corpus Linguistics?" *Language and Linguistics Compass* 3:5 (2009): 1225–1241. doi: 10.1111/j.1749-818X.2009.00149.x

McEnery, Tony & Andrew Wilson. *Corpus Linguistics: An Introduction*. 2nd ed. Edinburgh: Edinburgh University Press. 2001.

McEnery, Tony & Andrew Hardie. *Corpus Linguistics: Method, Theory and Practice*. New York: Cambridge University Press. 2012.

Sinclair, John. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press. 1991.

Sinclair, John. "Corpus and Text: Basic Principles." in Developing Linguistic Corpora: a Guide to Good Practice. ed. Martin Wynne. Oxford: Oxbow Books. 2005. Accessed: July 18, 2017. http://ota.ox.ac.uk/documents/creating/dlc/

Endnotes

[1] Gries, 1225.

[2] McEnery & Wilson, 5.

[3] McEnery & Wilson, 12.

[4] McEnery & Wilson, 8.

[5] McEnery & Wilson, 7-10.

[6] McEnery & Wilson, 13.

[7] McEnery & Wilson, 14.

[8] Church, 100.

[9] McEnery & Wilson, 17-19.

[10] Biber, 243.

[11] Biber, 243.

[12] Francis & Kucera: Contents

[13] Biber, 244.

[14] Biber, 256; Atkins et al 5-6.

[15] Davies, 160.

[16] http://corpus.byu.edu/corpora.asp

[17] http://corpus.byu.edu/copyright.asp