What is a corpus and why are corpora important tools?

Kristina Nilsson Björkenstam
Computational Linguistics, Stockholm University

1. Introduction

In 2012, the Republican candidate for US president, Mitt Romney, tried to defend himself against allegations that he was too liberal by saying:

> "But I was a **severely conservative Republican governor**." (Mitt Romney, CPAC 2012-02-10)

People, both within the Republican Party and outside, got upset because of the phrase "severely conservative". It didn't feel right. It was perceived as negative, almost as if Romney didn't want to be a conservative. But if we look up the adjective "severely" in Webster's dictionary we find the following definitions:

1. *harsh or strict, unsparing, stern*
2. *serious, grave, forbidding*
3. *conforming strictly to a rule or standard*
4. *extremely plain or simple*
5. *keen, violent, intense*
6. *difficult, rigorous*

It would seem that Romney used the word "severely" in sense #3, "conforming strictly to a rule or standard". If this is an example of correct usage of the word, why did people (especially within his own party) get so upset with him? What did he do wrong?

The answer is that there is more to language than dictionary definitions. If we look up the word "severely" in the 450 million word Corpus of Contemporary American English,[1] we find that this word typically co-occurs with words like:

- *damaged*
- *injured, wounded, ill, depressed*
- *disabled*
- *limited, restricted, limit*
- *punished, beaten*
- *criticized*
- *affected*

This pattern of co-occurrence with mostly negatively charged words is the reason why "severely" has negative connotations, and this is why people reacted to the phrase "severely conservative".[2] This blending of features of one set of words (e.g., "damaged", "injured", and "depressed") with another word ("severely") through frequent co-occurrence is called *semantic prosody* (Louw,

[1] COCA. URL: http://corpus.buy.edu/coca
[2] See the blog entry "Severely X" by Mark Liberman at Language Log for more on this quote. URL: http://languagelog.ldc.upenn.edu/nll/?p=3762

1993), and this is one of the aspects of language that can be studied by analyzing language production in collections of language samples, so-called *corpora*.

2. What is a corpus?

A corpus is a collection of natural language (text, and/or transcriptions of speech or signs) constructed with a specific purpose. While most available corpora are text only, there are a growing number of multimodal corpora, including sign language corpora.

A multimodal corpus is ”a computer-based collection of language and communication-related material drawing on more than one sensory modality or on more than one production modality” (Allwood, 2007:207), where sensory modalities include sight, hearing, touch, smell or taste, and production modalities e.g., speech, signs, eye gaze, body posture, and gestures. That is, a multimodal corpus is a collection of video and/or audio recordings of people communicating. But any collection of audio and video is not a corpus. Firstly, the audio-visual material should be carefully selected, and the content must be described using meta-data. Secondly, the material should be analyzed and described with transcriptions and annotations in a standardized format.

Ideally, a corpus is a set of language production samples designed to be representative of a language (or sub-language) through careful selection -- not a randomly collected set of data. How representative a corpus is, given a particular research question, is determined by the balance and sampling of the corpus. We can think of representativeness as the answer to the question: how well does this corpus describe relevant aspects of the language? In order to create a *general corpus*, language samples produced by both men and women, of all ages, from different parts of the area where the language is spoken, etc., should be included.

The same principles regarding representativeness, balance and sampling are relevant for both text and multimodal corpora, and there is a large body of work on corpus design to draw from (for an introduction see e.g., (McEnery, Xiao & Tono, 2006) and (Allwood, 2007)). There are different ways to go about selecting data. One way is to focus on language as a "product" and sample different types of language material, e.g., dialogue or monologue, or scripted or spontaneous speech. Another way is to focus on the “producer” of language, and to choose informants based on speaker characteristics such as age, gender, social class, first/second language, level of education, profession, and regional background. In some cases, e.g., when recording communication between co-workers in a specific work place, the informants are selected because they work there, and not based on speaker characteristics. Such corpora are *specialized* rather than general, but speaker characteristics are still important when analyzing the data.

In the case of multimodal corpora, an important aspect is whether the recordings were made in a naturalistic setting in a studio, or in the real world. There is also a difference between unobtrusive observation of an activity (e.g., a parent and a child playing with a set of toys at home), and recordings of people performing a task according to instructions (e.g., two adults discussing a movie in a lab setting).

Corpus selection is important not only for corpus builders but also for also for corpus users because the set of questions that can be investigated depends on the composition of the corpus. Let me give you an example: The Alcohol Language Corpus (ALC; Schiel, Heinrich & Barfüßer, 2011) is a specialized corpus consisting of speech samples from 162 speakers of German (85 male and 77

female police officers). The recordings were made in a car. Each speaker was recorded both sober and drunk (with alcohol level as meta-data), speaking in two different speaking styles: scripted, (that is, as instructed, e.g., reciting addresses and registration numbers), and spontaneous (e.g., descriptions, question-answer, and dialog).

Using this corpus, Schiel and colleagues (2011) show that there is a gender difference in spontaneous speech: men talk less when drunk, whereas women are not affected. They also show that both men and women make more speech errors when drunk. That is, using this corpus we can study how speech performance is affected by alcohol, and this knowledge can be used e.g., to device new methods to stop drunk drivers. But there are inherent limitations to this corpus that affect how this corpus can be used: the recordings were made in cars, all participants are police officers, and the speakers are drunk in some of the recordings. This, of course, means that we cannot use this corpus to study e.g., German in general. For that purpose, we need a balanced corpus consisting of language material produced by men and women of all ages, from all regions, with different speaker characteristics.

3. What corpora can tell us...

The first thing I do when I get access to a new corpus is to explore the content using some basic methods, typically by counting the words. We can find out which words are the most frequent in the corpus, and by ranking the words by corpus frequency we can study the distribution of the vocabulary of the corpus. By using normalized frequencies, we can make comparisons between different corpora. We can e.g., compare the vocabulary frequency distribution of English (British National Corpus[3]), Swedish (Stockholm-Umeå Corpus[4]), and Swedish Sign Language (Swedish Sign Language Corpus[5]).

By using basic corpus linguistic tools, either built-in web interface tools for corpora such as COCA or BNC, or software such as AntConc,[6] we can also look at recurring sequences of words or signs, either as sequences of tokens (called n-grams) or as collocations.

Starting with basic methods such as these, we can move on to study many aspects of language production using both quantitative and qualitative methods.

4. ... and what corpora cannot tell us

However, there are limitations to what corpora can tell us.

- No negative evidence: just because a word or a sign does not occur in a corpus (however large and well balanced) does not mean that the word or sign never can occur in the language. However, a representative corpus can show us what is central and typical in a language.
- The findings of a study can tell us something about the subset of language that is included in that corpus, but not necessarily about language as a whole. However, if the corpus is

[3] BNC. URL: http://www.natcorp.ox.ac.uk/
[4] SUC. http://spraakbanken.gu.se/eng/resource/suc3
[5] SSLC. http://www.ling.su.se/forskning/forskningsprojekt/teckenspr%C3%A5k/teckenspr%C3%A5kskorpus
[6] AntConc. URL: *www.antlab.sci.waseda.ac.jp/*

representative of the language we are interested in we can make careful generalizations about the language.
- A corpus can rarely provide explanations, and thus most corpus studies combine quantitative and qualitative work. Sometimes other methods, such as questionnaires, eye gaze or EEG experiments are better suited to answer a particular question. Sometimes a descriptive corpus study can give new ideas on what to look for using other methods.

To summarize: make sure that you select the right corpus for your study, find out as much as you can about the corpus, take the characteristics and limitations of the corpus into account, and make careful generalizations!

5. Why are corpora important tools?

Corpus analysis provides quantitative, reusable data, and an opportunity to test and challenge our ideas and intuitions about language. Further, analysis applied to corpora as transcriptions or other types of linguistic annotation can be checked for consistency and inter-annotator agreement, and the annotated corpus can be reviewed and reused by others.

Corpora are essential in particular for the study of spoken and signed language: while written language can be studied by examining the text, speech, signs and gestures disappear when they have been produced and thus, we need multimodal corpora in order to study interactive face-to-face communication.

References

Allwood, J. 2007. Multimodal Corpora. In: Lüdeling, A. & M. Kytö (eds) Corpus Linguistics. An International Handbook. Mouton de Gruyter. Berlin: 207-225

Davies, M. 2008. *The Corpus of Contemporary American English: 450 million words, 1990-present*. URL: http://corpus.byu.edu/coca/

Louw, B. 1993. *Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies*. In Baker, M., Francis, G. & Tognini-Bonelli, E. (eds) "Text and Technology". Philadelphia/Amsterdam: John Benjamins.

McEnery, T., R. Xiao & Y. Tono. 2006. Corpus-Based Language Studies: An Advanced Resource Book. Taylor & Francis US.

Schiel F, C. Heinrich & S. Barfüßer. 2011. Alcohol Language Corpus. In: Language Resources and Evaluation, Springer, Berlin, New York, Vol 45.

The British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

The Stockholm-Umeå Corpus, version 3.0. 2012. Distributed by the Swedish Language Bank at Gothenburg University. URL: http://spraakbanken.gu.se/eng/resource/suc3

The Swedish Sign Language Corpus. 2013. Distributed by the Section for Sign Language at Stockholm University. URL: http://www.ling.su.se/english/research/research-projects/sign-language