Creating a Corpus: Planning, Collection, and Analysis

Andrew Petrun, Joleen Hanson, University of Wisconsin-Stout

What is a corpus?

A corpus is a collection of tagged text that is designed to be easily indexed and searched for the purpose of conducting research on the context and frequency of given words or phrases.

For example:

A parent would like to communicate better with their high school-age children and wants to know what various words or phrases are popular. A search in a corpus of dialogue between high school students would produce exactly what this parent needs, allowing them to know what the current lingo is and how to use it appropriately. Pretty groovy, eh home slice?

The purpose of our corpus

The corpus I helped to create was a collection of self-introductions from several different sources:

- Senior citizen profiles written upon joining a retirement community
- Freshman college student introductory essays from an English composition course
- Online social networking profiles from Google Orkut and Myspace

The corpus was designed for defining the self-introduction as a genre and researching various ways in which language use varies with respect to variables in demographics and mode of introduction.

Planning

There were three main aspects of the planning process:

- Determining what needed to be done and by whom. We needed to delegate tasks evenly between me and my research partner and ensure that the information we gathered separately could be easily integrated when it was finally entered into the corpus.

- Deciding on a consistent criteria for collecting text for the corpus. This included decisions on which social networking sites would be used, how to determine who was eligible from the sites chosen, and what software would be used for collection and collaboration.

- Deciding on a consistent protocol for preparing the collected text for the corpus. This included decisions on how the text would be redacted, coded, and tagged, as well as what software would be used for building the corpus.

Collecting and Preparing Text

As stated, the text used in the corpus was collected from three different sources.

Two of the sources were collected as physical documents and had to be manually “digitized,” while the other source was already in digital format and only needed to be copied and pasted.

For all three sources, personal information was removed or “redacted” from the text and then the text was tagged for topic or rhetorical move, lexical category, and complex sentence structure.

Analyzing the Corpus

Once the information from all three sources had been collected, redacted, and tagged it was ready to be entered into the corpus program that we had selected called Wordsmith.

The information was transferred from a plain text document into Wordsmith and searched using the “concordance” function.

This project is still a work in progress, and there are many advanced features in Wordsmith that we will be learning and applying in the future.

N	Concordance	Word #	Sent. #	Sent. Pos.	Para. #	Para. Pos.	Sect. #	Sect. Pos.	File	%
1	Couple PA0003 [female first name] I was born in [city] and raised in [city,	5	0	28%	0	2%	0	2%	Couple PA0003.txt	2%
2	name] with a major in speech-drama, I went to [university name] for a	28	1	52%	0	9%	0	9%	Couple PA0003.txt	11%
3	church in [country]. In later jobs I was a group worker at the [company	60	3	20%	0	19%	0	19%	Couple PA0003.txt	22%
4	name] in [city, state]. In [1945-1955] I had my first trip out of the US to the	83	4	20%	0	26%	0	26%	Couple PA0003.txt	29%
5	of Women in Guatemala. The next year I married and we eventually traveled to	103	5	28%	0	32%	0	32%	Couple PA0003.txt	34%
6	the U.S. and Canada. In [1965-1975] I was widowed and continued to live the	145	9	20%	0	45%	0	45%	Couple PA0003.txt	46%
7	14 years in [state]. Ten of those years I spent 2-3 months each year in New	162	10	30%	0	51%	0	51%	Couple PA0003.txt	52%
8	"side trips" to Australia. In [1975-1985] I took my first trip with Elderhostel to	179	11	33%	0	56%	0	56%	Couple PA0003.txt	57%
9	Elderhostel to Australia. On that trip I met [male full name]. A year later we	191	12	56%	0	60%	0	60%	Couple PA0003.txt	60%
10	[male first name] has done over 100 and I've done 90-some. We've done a lot of	236	14	93%	0	74%	0	74%	Couple PA0003.txt	72%
11	and have taken bus trips to enjoy them. I volunteered for 10 years at [company	277	18	10%	0	87%	0	87%	Couple PA0003.txt	86%
12	Female PA0007 I grew up in [city], [state] and attended	2	0	10%	0	2%	0	2%	Female PA0007.txt	5%
13	[university name] in [state] where I met my husband, [male first name].	15	0	71%	0	12%	0	12%	Female PA0007.txt	13%
14	away on February 19, [2005-2015]. I have been a member of the League of	83	3	13%	0	66%	0	66%	Female PA0007.txt	69%
15	. Genealogy is one of my hobbies and I found that my great grandparents lived	104	4	53%	0	83%	0	83%	Female PA0007.txt	83%
16	Male PA0005 I was born in [city], [state], but grew up	2	0	10%	0	1%	0	1%	Male PA0005.txt	2%
17	High School in [1945-1955]. That year I enrolled in the pre-veterinary program	24	1	31%	0	8%	0	8%	Male PA0005.txt	8%
18	program at [abb university name]. I earned a V.M.D. from the Veterinary	34	2	40%	0	11%	0	11%	Male PA0005.txt	11%
19	. During my graduate studies, I served as a pathologist at the [city]	65	5	32%	0	21%	0	21%	Male PA0005.txt	21%
20	at the Veterinary School. In [1955-1965] I received an appointment as a	81	6	17%	0	27%	0	27%	Male PA0005.txt	25%
21	reproductive endocrinology and biology. I also was the scientific director for	112	8	10%	0	37%	0	37%	Male PA0005.txt	35%

concordance | collocates | plot | patterns | clusters | filenames | follow up | source text | notes

136 Set Couple PA0003 [female first name] I was born in [city] and raised in [city, state], another great university town! After college at [university name] with a major in speech-drama, I went to [university name] for a master's degree in religion. My fir

Figure 1. An example of concordance search results using the word “I”