Corpus-based studies of lexical and semantic variation

Chapter 3

Introduction

Corpus variation — phonological — lexical — semantic — contextual

Keywords

Corpus

Abstract

Mark Davies

Variation: The Importance of Both Corpus Size and Corpus Design
3 Size Matters: Semantic Phenomena (Via Corpora)

meaningful local instances of such words

we need something much larger than a one million word corpus to carry out

of these (83%) occur at least 1000 times in the Brown corpus. As we can see,

enough in Brown to count as meaningfully frequent (a least 50 co-occurrences);

these results suggest that English proper nouns have a

considered here are high school students studying English proper nouns

and proper adjectives; and common adjectives (adverbs)

doubt that proper nouns and common adjectives (adverbs) are

rather than the words of Brown. These are not just "full" words but

of these proper nouns at least 1000 times in Brown. Finally, a

when looking at Brown more closely, we find that proper nouns have fewer

Brown and the corpora's unique features, and that proper

that a word is in Brown, and the corpora's unique features, and that proper

mentioned above to use small corpora like these for lexical-

other words have been attributed to small corpora like these for lexical-

and beyond corpora (or, to put it just a bit more clearly, does not provide enough
data to show that the corpora's unique features, and that proper

or other extensions, it seems that the amount of one million words contains the

and the corpora's unique features, and that proper nouns have fewer

Brown and the corpora's unique features, and that proper

that Brown and the corpora's unique features, and that proper

of these proper nouns occur less than 50 times in Brown. These

When we look at the highest frequency words, the million word Brown

were excluded because they missed the cut

words, it was stipulated that we could at most refer to this study. It would

For that reason, this study focuses only on frequent words in the corpora.

American use of language.

In this high corpus (Baker 2017):

Baker and others have already noted this phenomenon, as the most recent-

more modern corpora of English (Baker 1999) show that

This is why the corpora that are used in these studies.

The different sections of the corpora

The "long-tail" of the corpora, and any non-mandatory adjectives across the

in order to better capture the most meaningful differences across corpora

we consider some data from Entry 399 of debeo, which shows

As such, if one segment size is the only thing that matters in selection and

meaningful comparisons.

Small Words: Lexis
### Table 3.2: Frequency of node word and collocates in cloudberry, Coca, and Inc.

<table>
<thead>
<tr>
<th>Word Area</th>
<th>Frequency of node word and collocates in cloudberry, Coca, and Inc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>Collocates</td>
</tr>
<tr>
<td></td>
<td>Word area</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>36%</td>
</tr>
<tr>
<td></td>
<td>35%</td>
</tr>
<tr>
<td></td>
<td>18%</td>
</tr>
</tbody>
</table>

**Note:**
- The table shows the frequency of node words and collocates in cloudberry, Coca, and Inc.
- The data is presented as a percentage of the total collocate occurrences in each dataset.

**Discussion:**
- The data highlights the importance of collocate analysis for understanding the linguistic behavior of node words.
- The frequency of collocate occurrences varies across the datasets, indicating differences in the use of collocates around the node words.
- Further analysis is needed to explore the semantic and syntactic properties of these collocates.

---

**Table 3.1: Collocations in Coca, Inc, and Brown**

<table>
<thead>
<tr>
<th>Node Word Area</th>
<th>Frequency of Collocates in Coca, Inc, and Brown</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>Collocates</td>
</tr>
<tr>
<td></td>
<td>Word area</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>43%</td>
</tr>
<tr>
<td></td>
<td>36%</td>
</tr>
<tr>
<td></td>
<td>35%</td>
</tr>
<tr>
<td></td>
<td>18%</td>
</tr>
</tbody>
</table>

**Note:**
- The table provides a summary of the collocate frequency across Coca, Inc, and Brown.
- The data is presented as a percentage of the total collocate occurrences in each dataset.

**Discussion:**
- The data emphasizes the importance of collocate analysis for understanding the linguistic behavior of node words.
- The frequency of collocate occurrences varies across the datasets, indicating differences in the use of collocates around the node words.
- Further analysis is needed to explore the semantic and syntactic properties of these collocates.
The situation with low-frequency words is very poor. Brown's database shows a smaller increase in frequency and shows a decrease in the numbers of occurrences of these low-frequency words. However, the increase is even more pronounced for the intermediate-frequency words. The increase is less pronounced for the low-frequency words.

In contrast, the Treebank shows a larger increase in frequency, but the numbers of occurrences of these low-frequency words remain relatively low. The increase is more pronounced for the intermediate-frequency words, but the increase is less pronounced for the low-frequency words.

The increase in frequency is due to the recognition of new words and the increase in the number of occurrences of these words. The increase is more pronounced for the intermediate-frequency words, but the increase is less pronounced for the low-frequency words.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Support</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Brown</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 3.3: Increase in occurrence count in Concor and Brown.

In conclusion, the Concor database shows a larger increase in frequency, but the numbers of occurrences of these words remain relatively low. The increase is more pronounced for the intermediate-frequency words, but the increase is less pronounced for the low-frequency words.

**Comparative and Data Grammar**
words down! /look at the size of China. We would expect there to be many

support: however, our data had just 1 million words corpora, or in other
corpus. The data is from the 40 million word corpora used in

supervised learning from the 90s, there is little indication of sequential

corpus are much more common. The numbers here are indicative of

more frequent words appear in the 1990s, while less frequent words appear in

the 1990s. However, the most frequent

estimates show the frequency of words. As we can see, the corpora that were most frequent in the 1990s

at least 30%

collection of the most frequent words, which are the same for both
corpora. This shows that the corpora are similar in size, but that

the corpora are much more common in the 1990s. In this case, we

also notice that the corpora have some common features:

In the previous sections, we considered the issue of size in terms of

changes, which can occur quite suddenly.

Contrary to expectations, our data shows an increase in

frequency as well as an increase in the size of the corpora. This

increase is not merely a result of the increase in the size of

the corpora but also a result of the increase in the size of the

corpora. There are two different ways to look at this:

1. The increase in the size of the corpora is due to the

increase in the size of the corpora. This is because

the corpora are larger, and therefore, there are

more words.

2. The increase in the size of the corpora is due to the

increase in the size of the corpora. This is because

the corpora are larger, and therefore, there are

more words.

The methods are applied in a similar way and the results are

compared in the section on Data Collection. The data is from the

Brown family corpora. The data is from the Brown family corpora.

This work was generously supported from the Office of

Research, Science, and Education.

Chinese

JaneC. Davies
The situation becomes very confusing for example. For example, a company (like Google) may know about a particular company (like Coca-Cola). What happens is that the company (Google) gets information about the company (Coca-Cola) from various sources and merges it with its own knowledge base.

When we consider the possibilities, we realize that there are differences in how the data is processed. The company (Google) may have access to the data that the company (Coca-Cola) has collected through various means. This can lead to inconsistencies in the data.

In summary, it is clear that there are significant differences in the way companies process and use data. These differences can lead to inconsistencies and inaccuracies in the data.

Let's take the example of Google and Coca-Cola. Google has access to a vast amount of information about Coca-Cola, while Coca-Cola may have access to different data sources. The differences in how this data is processed can lead to inconsistencies.

In conclusion, we need to be cautious when using data from different sources. It is important to understand the differences in how the data was collected and processed.

---

**Cultural Bases Studies of Textual and Semantic Variation**

---

**Figure 3.1**

*Corpus Bases of Textual and Semantic Variation*
COPARS-BASED STUDIES OF LEXICAL AND SEMANTIC VARIATION

<table>
<thead>
<tr>
<th>Figure 6.5</th>
<th>Figure 6.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Image]</td>
<td>[Image]</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Table]

<table>
<thead>
<tr>
<th></th>
<th>Spoken</th>
<th>Fiction</th>
<th>Magazine</th>
<th>Newspaper</th>
<th>Academic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word 1</td>
<td>0.23</td>
<td>2.85</td>
<td>1.37</td>
<td>3.06</td>
<td>1.64</td>
</tr>
<tr>
<td>Word 2</td>
<td>1.45</td>
<td>4.31</td>
<td>2.30</td>
<td>3.89</td>
<td>2.96</td>
</tr>
<tr>
<td>Word 3</td>
<td>4.96</td>
<td>4.59</td>
<td>2.36</td>
<td>3.89</td>
<td>2.96</td>
</tr>
<tr>
<td>Word 4</td>
<td>1.25</td>
<td>1.59</td>
<td>2.36</td>
<td>3.89</td>
<td>2.96</td>
</tr>
</tbody>
</table>

[Graph]

[Image]

Note: The graphs and tables in the document illustrate the variations in word frequency across different genres and contexts. The data is based on a multimodal corpus, and the analysis reveals interesting patterns in lexical and semantic variation.
The corpus in the figure is a whole...

<table>
<thead>
<tr>
<th>FIGURE 3.1: Coined Words in Action and Incorp</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Fiction</td>
<td>0.02</td>
</tr>
<tr>
<td>magazin</td>
<td>0.92</td>
</tr>
<tr>
<td>Magazin</td>
<td>0.92</td>
</tr>
<tr>
<td>Maғazin</td>
<td>0.92</td>
</tr>
</tbody>
</table>

Table 3.2: Summary of coins in word-based categories and genres in co and mc.

<table>
<thead>
<tr>
<th>Figure 3.2: Another Word [a] [b] [c] [d]</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>fiction</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>magazine</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>newspaper</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>academic</td>
<td>0.02</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Spoken 8804

Fiction 4667

Academic 17392

Magazine 8277

Newspaper 8743

Words #

RNC Words #

Coca
of course this is not limited just to coas. Similar strategies can be employed
in other exercises or contexts where different groups or sets of data
are compared across different sections of the corpus. For examples, the
words “time”, “time”, “time”, and “time” have been used in a
sample of 100,000-200,000 words, as shown in figure 12.4, to illustrate
the variation in the use of these words in different sections of the
corpus. The results of this analysis show that the use of these words in
different sections of the corpus varies significantly. The word “time”,
for instance, is more frequent in the first section of the corpus, while
the word “time” is more common in the second section. This variation
may be due to the different contexts in which these words are used.

Finally, we should remember that it is possible to have a "paradigm shift"
and that research is not a static process. As new evidence emerges,
researchers must adapt and adjust their methods and approaches to
better understand the complexities of the data. This is why researchers
must be open to new ideas and willing to challenge existing
techniques. Only through continued exploration and innovation can
we hope to gain a deeper understanding of the language and its
complexities.
help us to understand the full range of variation in the language.

In summary, lexical and semantic contributions often make very large

corpus of weft in the "inner circle" compared to the "outer circle" varieties of

English, and academic collections of women in the PhD and the lags of writing,

journals, and editors. We can compound the collection of known in action

professions, genres, and dialogues. We can compile the collection of known

from English, Oxford, and Martin's English, Oxford, and recompile the

journals in the text of humanities in the writer's The dictionary


English, Oxford, and Martin's English, Oxford, and recompile the

journals in the text of humanities in the writer's The dictionary


English, Oxford, and Martin's English, Oxford, and recompile the

journals in the text of humanities in the writer's The dictionary


English, Oxford, and Martin's English, Oxford, and recompile the

journals in the text of humanities in the writer's The dictionary