Language variation and corpora

As you have seen in some assignments, corpora are useful for studying variation in language

- Variation across different social & communicative contexts (regions, registers, etc.)

We’ll look at a few points from studies in the book to help us get a better sense of some of the various issues

- Next class, we’ll put this into practice

Register variation in English

Biber 1995

Goal: “provide comprehensive descriptions of the patterns of register variation”

- Identify underlying linguistic parameters of variation (dimensions)
  - Cover a range of linguistic features, since no feature in and of itself determines a register
  - Goal is not to analyze individual constructions, but to use them to analyze whole texts
- Specify similarities & differences among registers based on these dimensions

Register = groups of texts

- Two registers can be compared in their similar use of co-occurring features
  - co-occurring features = empirically determined set of features that tend to co-occur

Linguistic features

Some example linguistic features that can be used to build up a multi-dimensional analysis

- lexical features: type-token ratio, word length, …
- semantic features: hedges, speech act verbs, …
- grammatical classes: nouns, predicative adjectives, …
- syntactic features: relative clauses, passive postnominal participial clauses

Multi-dimensional analysis

Steps in multifeature/multidimensional (MF/MD) analysis:

1. Collect texts with register information
2. Collect set of potential linguistic features to analyze (based on previous studies)
3. Automatically tag texts with features, post-editing where necessary
4. Compute frequency co-occurrence patterns of linguistic features using factor analysis
   - Functional interpretation of co-occurrence patterns = dimensions of variation
5. Sum the features on each dimension:
   - Mean dimension scores for each register used to analyze similarities and differences

Multi-dimensional analysis (2)

How factor analysis works (roughly):

- Build a correlation matrix of all features
- From this, determine the loading, or weight, of each linguistic feature
  - Loading tells us to what degree we can generalize from this factor to the linguistic feature
  - Positive loading = positive correlation (likewise for negative)
  - High absolute value = more representative the feature is of a factor/dimension/register

Biber removed features with absolute value under 0.35

- Features only kept on factor they had highest loading for (even if occurring on 2+ with scores above 0.35)
Results

Biber found these dimensions for register variation in English:

- involved vs. informational production
- narrative vs. non-narrative concerns
- overt expression of persuasion
- abstract vs. non-abstract style

These were his functional interpretations, based on the linguistic features and the resulting text splits

See table 1, p. 164, in the book for more details

Zero subject relatives in English

Lehmann 2002

Compares zero-subject relative (ZSR) constructions between American and British English

1. It was Joanne Ø said you’d go down there [BNC]

- American: 94 ZSRs/5 million words
- British: 205 ZSRs/4.2 million words

Important: Need to compare zero-subject occurrences against overall subject relatives

- Difference could be in overall relative use
- American: 94/3647 (2.5%); British: 205/1376 (13%)

Likewise, we can find differences in the matrix sentence (higher percentage of existential there in British)

Metadiscourse

Hyland 1999

Goal: Compare metadiscourse features across genre and across discipline

- Metadiscourse features: e.g., hedges, connectives, etc.
- Genre: textbooks vs. research articles
- Discipline: biology vs. applied linguistics vs. marketing

Some findings:

- Textbooks in all three genres use a large amount of logical connectives and code glosses
- Research articles showed a marked increase in interpersonal markers
- “[M]etadiscourse variations were more pronounced between genres than disciplines”

Incorporating social variables

Q1: Do certain ethnic groups use ZSRs more?

- Hard to say, given the low overall frequency of ZSRs

Q2: Do certain ages use ZSRs more?

- American: no clear picture
- British: ZSR usage increases as the ages get older

What does this result show us about language change?

- Speaker age in a synchronic corpus gives indications of language change, but is “certainly not uncontroversial”
- Frequency of use may change over one’s lifetime

Getting started

- Getting a concordancer
  - We know how to use AntConc - but we’ll want to be able to use regular expressions in it
  - And, importantly, we’ll want to extract frequencies

- Getting corpus data: for this to work, you’d want to look at 2 different corpora
  - We’re going to only look at one during class, the Brown corpus, found on miller at:
  - /Volumes/Data/en/brown/
  - You can do comparisons outside of class, assuming (similar) POS tags in the other corpus
Searching for our specific patterns

We can use AntConc or we can use Perl (see help.pl), which is well-maintained for regular expressions.

We want to compare help V and help to V

- and to compare help NP V and help NP to V

For these latter cases, we'll simplify the NP to be a single noun (tag starts with n) or pronoun (p)

The patterns we'll use (read \s as a space):

- help V: \b(help\w*/v\w*/\s+\w+/v\w*?)\b
- help to V: \b(help\w*/v\w*/\s+to/to\s+\w+/v\w*?)\b
- help NP V [approximated]: \b(help\w*/v\w*/\s+s\+to/to\s+s\+\w+/v\w*?)\b
- help NP to V [approximated]: \b(help\w*/v\w*/\s+s\+to/to\s+s\+\w+/v\w*?)\b

Breaking down the regular expression

We'll deal with this more later in the semester, but let's walk through a few things

while (m{\b(help\w*/v\w*/\s+\w+/v\w*?)\b}gi)

Perl options

- The g at the end specifies global matching, in case there is more than one match
- The i at the end makes it case-insensitive

Breaking down the regular expression (2)

This is the formal way of saying help (or its variants), used as a verb (v) and followed by any verb

- \b specifies a word boundary
- \w specifies only word characters, and *? specifies 0 or more of them, i.e., possible word endings
  - FYI (for those “in the know”): *? is the non-greedy version of the usually greedy
  - \s specifies white-space characters, and + specifies one or more of them
    - this is more robust than just putting a space

Breaking down the regular expression (3)

This regular expression:

\b(help\w*/v\w*/\s+s\+\w+/v\w*?)\b

is trying to say this:

help(s|ed|ing|ful|...)/verb <word>/verb

Today's goal 1: Don't worry about all the details, just the gist

- Today's goal 2: feel comfortable making little changes to search for what you want

Frequency counts

Let's alter the Perl code to get frequency counts . . .

<table>
<thead>
<tr>
<th>Inf-type</th>
<th>No NP</th>
<th>With NP</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bare</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Intervening NP

Take a look at the numbers you have for help (to) in Brown

- with an intervening NP
- without an intervening NP

Do you notice any trends?
Infinitive marker

A further claim:
- If *to* precedes *help*, then there is less likely to be a *to* afterwards

Let’s alter the Perl regular expressions to find out whether this is true in the Brown data
- What types of patterns do we need to express?
- Can we use particular POS tags to help?

The passive

Yet a further claim:
- The passive (e.g., *were helped*) only occurs with *to*

So, let’s search for (using the book’s notation):
- *be* verbs (POS starting with *vb*) followed by a *vvn* *help*
  - with vs. without *to* + following verb

Language variety

To fully perform the experiments in the book, we would need 3 more corpora
- Look at table C2.5 on p. 235

We’ll cover statistical tests the next two weeks, to see whether the observed differences are statistically significant
- Without statistical tests for now, what do you observe in the table?