

# Corpus Linguistics (L615)

## Application #1: Language variation

Markus Dickinson

Department of Linguistics, Indiana University  
Spring 2009

# Language variation and corpora

Corpora are extremely 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 figure out what to do

- ▶ Next class, we'll put this into practice

# Register variation in English

Biber 1995

Corpus Linguistics

Application #1:  
Language variation

**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 and 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

Biber 1995

Hyland 1999

Lehmann 2002

Kachru 2003

help/help to

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

Steps involved 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

Biber 1995

Hyland 1999

Lehmann 2002

Kachru 2003

help/help to

# Multi-dimensional analysis (2)

## How does factor analysis work?

- ▶ 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 are only kept on the factor they had the highest loading for (even if they occur on 2+ with scores above 0.35)

Biber 1995

Hyland 1999

Lehmann 2002

Kachru 2003

help/help to

Biber found the following dimensions for register variation in English:

- ▶ involved vs. informational production
- ▶ narrative vs. non-narrative concerns
- ▶ elaborated vs. situation-dependent reference
- ▶ 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

Biber 1995

Hyland 1999

Lehmann 2002

Kachru 2003

help/help to

# Metadiscourse

Hyland 1999

Corpus Linguistics

Application #1:  
Language variation

Biber 1995

Hyland 1999

Lehmann 2002

Kachru 2003

help/help to

**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”

# Zero subject relatives in American & British English

Lehmann 2002

Corpus Linguistics

Application #1:  
Language variation

Biber 1995

Hyland 1999

Lehmann 2002

Kachru 2003

help/help to

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

(1) It was Joanne  $\emptyset$  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)

# 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?

- ▶ Using speaker age in a synchronic corpus gives indications of language change, but it's “certainly not uncontroversial”
- ▶ Frequency of use may change over one's lifetime

# Definite reference in World Englishes

Kachru 2003

Corpus Linguistics

Application #1:  
Language variation

Biber 1995

Hyland 1999

Lehmann 2002

Kachru 2003

help/help to

Kachru looks at definite reference in world Englishes

- ▶ collected letters comprising 15,000 words (945 definite NPs)
- ▶ hand-marked the definite NPs into 9 major classes

From this, can see some tendencies for definite NPs

- ▶ But would need more data to know for sure

## Subcategorizations of *help*:

- ▶ help to V
- ▶ help NP to V
- ▶ help V
- ▶ help NP V

## Some questions:

- ▶ Is the choice of *to* arbitrary?
- ▶ What factors influence this selection?

# Getting started

- ▶ Getting a concordancer
  - ▶ By the end of the semester (or earlier), most of you will be able to write your own in Perl
  - ▶ We'll take a look at the Multilingual Corpus Toolkit (MLCT)
    - ▶ <http://personalpages.manchester.ac.uk/staff/scott.piao/research/DownLoad/download.htm>
    - ▶ download and unzip `mlct_public.zip`
- ▶ 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
    - ▶ `/Volumes/Data/Corpora/en/brown/`
  - ▶ You can do comparisons outside of class (assuming POS tags)

We might also look at an online concordance for some help

# Using MLCT

- ▶ Make sure you have java installed on your computer
- ▶ Double-click on `mlct_public.jar`
- ▶ For information, click on Help → Readme

To load a corpus, click on File → Open in Left Window

# Searching for patterns

1. Load corpus
2. Tools → RegExp Frequency Matching Table
3. Put RE in the box next to the anchor
  - ▶ Note that word boundaries are sometimes implicit (?)
4. Click on the anchor

This will give us a table with the frequency count of each matched pattern

Concordancing is done a little bit differently in MLCT than in the book

- ▶ It doesn't immediately provide you with frequencies
- ▶ It does give keywords in context, and you can sort this based on left or right context words

How to run the concordancer:

- ▶ Concordance → Select Files for Concordance
- ▶ Press the *Extract Concordance* button (little newspaper-looking graphic)

# Searching for our specific patterns

We're going to switch to Perl (see `help.pl`) ... it's better-documented for using 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*:  
`\b(help\w*?/v\w*?\s+\w+/[np]\w*?\s+\w+/v\w*?)\b`
- ▶ *help NP to V*:  
`\b(help\w*?/v\w*?\s+\w+/[np]\w*?\s+to/to\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

- ▶ Perl is matching the current line against this regular expression, within `m{...}`
- ▶ The `g` at the end specifies *global* matching, in case there is more than one match (also why we have a `while` loop)
- ▶ The `i` at the end makes it case-insensitive

## Breaking down the regular expression (2)

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

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+\w+/v\w*?)\b
```

is trying to say this:

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

Don't worry about the details today; just trust me that it works as it should

- ▶ and focus on making little changes to search for what you want

# Frequency counts

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

Inf-type	No NP	With NP	Total
Full			
Bare			

# 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 or without *to* + following verb

# Language variety

We'll use the numbers from the book for the next part

- ▶ To fully perform the experiments, we need 3 more corpora

We'll get to statistical tests in a couple weeks, to see whether the observed differences are statistically significant

- ▶ We'll first tabulate the totals, ignoring whether there's an intervening NP