

# Corpus Linguistics (L615)

## Multidimensional Analysis

Markus Dickinson

Department of Linguistics, Indiana University  
Spring 2009

# Multidimensional Analysis of Registers

Doug Biber

We talked a while back about multidimensional analysis, and I want to revisit it.

- ▶ Recall that it's a way for us to determine which linguistic features tend to predict different register properties

We'll walk through a few more details today, but by no means will you become experts

- ▶ You can find some software out there to use it (e.g., VARBRUL, R)

# 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 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

# 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

General 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

# 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 (range from -1.0 to 1.0)
  - ▶ Loading tells us to what degree we can generalize from this factor to the linguistic feature
  - ▶ Positive loading = positive correlation (similarly for negative)
    - ▶ Positive and negative features occur in complementary patterns
    - ▶ 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)

# Results (overview)

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

# Steps in Multivariate Analysis

Biber et al 1992

A more specific listing of the process:

1. data collection
2. determine linguistic features
3. automatic POS and syntactic analysis
4. post-edit annotation manually
5. count frequencies of features
6. analyze co-occurrence patterns of features (factor analysis)
7. functional interpretation of co-occurrence patterns
8. compute dimensional scores for each text
9. revise functional interpretation of dimensions based results from previous step

# After Step 2:

Linguistic features for written/spoken continuum:

- ▶ first person pronoun
- ▶ second person pronoun
- ▶ third person pronoun
- ▶ nouns
- ▶ temporal adverbs
- ▶ nominalizations
- ▶ prediction modals
- ▶ agentless passives
- ▶ *by* passives
- ▶ ...

# Biber tagger output (Step 3)

```
The ^ati++++
dissolved ^jj+atrb++xvbn+
components ^nns++++
that ^tht+rel_subj++
precipitate ^vb++++
to ^to++++
form ^vbi++++
these ^dt+dem+++
rocks ^nns++++
are ^vb+ber+aux++
decomposed ^vpsv++agls+xvbnx
from ^in++++
...
```

# After Step 5:

|                       | mean  | min. | max. |
|-----------------------|-------|------|------|
| first person pronoun  | 27.2  | 0    | 122  |
| second person pronoun | 9.9   | 0    | 72   |
| third person pronoun  | 29.9  | 0    | 124  |
| nouns                 | 180.5 | 84   | 298  |
| temporal adverbs      | 5.2   | 0    | 122  |
| nominalizations       | 19.9  | 0    | 71   |
| prediction modals     | 5.6   | 0    | 30   |
| agentless passives    | 9.6   | 0    | 38   |
| <i>by</i> passives    | 0.8   | 0    | 8    |
| ...                   |       |      |      |

Actually uses *rate of occurrence* of linguistic features (e.g., number of nouns per 1000 words)

# Factor Analysis

- ▶ hypernym: multivariate analysis
- ▶ hyponyms:
  - ▶ principal component analysis
  - ▶ **common factor analysis (principal factor analysis)**

Goal: reduce large number of features to a small set of *factors*

- ▶ each factor groups features with a high correlation
- ▶ factor loading ranges from -1 to 1
- ▶ features with a positive loading occur together frequently in a text, while features of the same factor with a negative loading are mostly absent
- ▶ and the other way round

### After Factor Analysis (Step 6):

|                             | factor | loading |
|-----------------------------|--------|---------|
| first person pronoun        | F1     | 0.74    |
| second person pronoun nouns | F1     | 0.86    |
| ...                         |        |         |
| third person pronoun        | F2     | 0.73    |
| ...                         |        |         |
| temporal adverbs            | F3     | 0.60    |
| nominalizations             | F3     | -0.36   |
| ...                         |        |         |
| prediction modals           | F4     | 0.54    |
| ...                         |        |         |
| agentless passives          | F5     | -0.43   |
| by passives                 | F5     | -0.41   |
| ...                         |        |         |

### After Interpretation (Step 7)

|                      | load. | dimension                      |
|----------------------|-------|--------------------------------|
| 1st p. pronoun       | 0.74  | involved vs.                   |
| 2nd p. pronoun nouns | 0.86  | informational production       |
| ...                  |       |                                |
| 3rd p. pronoun       | 0.73  | narrative vs.                  |
| ...                  |       | nonnarrative discourse         |
| temp. adv. nomin.    | 0.60  | situation-dep. vs.             |
| ...                  | -0.36 | elaborated reference           |
| pred. modals         | 0.54  | overt expression of persuasion |
| ...                  |       |                                |
| agt. less passives   | -0.43 | nonimpersonal vs.              |
| by passives          | -0.41 | impersonal style               |
| ...                  |       |                                |

### Calculating dimension scores for texts

Now, we want to take each dimension and calculate a dimension score for a given text

- ▶ for one text: sum over all frequency counts of the salient dimensional features of the text
  - ▶ salient means greater than 0.35 absolute value
  - ▶ to sum = add up positive loadings, subtract out negative loadings
- ▶ But: individual linguistic variables are first normalized to mean 0.0 and standard deviation 1.0
  - ▶ This gives each feature equal weight in calculating a dimension score

### Example for dimension score calculation

Let's walk through an example for dimension 2:

- ▶ dimension 2 features: past tense, 3rd p. pronouns, perfect aspect, public verbs (*say, explain*), present part. clauses, synthetic negation

Assume that text X has:

- ▶ 113 past tense forms, 124 3rd p. pronouns, 30 perfect aspect forms, 14 public verbs, 5 present part. clauses, and 3 synthetic negations
  - ▶ non-standardized sum:  $113 + 124 + 30 + 14 + 5 + 3 = 289$

### Example for dimension score calculation (2)

- ▶ past tense forms in general have a mean of 40.1 and a standard deviation of 30.4:
  - ▶  $113 = (2.4 \times 30.4) + 40.1$
  - ▶ So, 2.4 is the dimension score contributed by past tense forms
- ▶ positive score of text X:  $2.4 + 4.2 + 1.5 + 2.3 + 1.5 = 15.9$ 
  - ▶ After subtracting out the negative loadings, we'll have this text's score

For all texts of a genre: calculate the mean over all text scores

### Graphical Display of Dimension 2

Biber et al 1992

