Unit 7: A multivariate approach to linguistic variation

Statistics for Linguists with R – A SIGIL Course

Stephanie Evert
Computational Corpus Linguistics Group
FAU Erlangen-Nürnberg

Linguistic variation

Variation of a quantitative linguistic feature

– frequency of passive, past perfect, split infinitive, ...
– frequency of expression, semantic field, topic, ...
– association strength, lexical density, productivity, ...

across

– languages and language varieties
– regions & social strata
– time (diachronic change)
– individual speakers & discourses

Studying linguistic variation

▪ Univariate approach
  – compare single feature across two or more conditions
  – e.g. AmE vs. BrE vs. IndE vs. ... / male vs. female / etc.
  – corpus frequency comparison

▪ Regression approach
  – predict single quantity from multiple explanatory factors

▪ Multivariate approach
  – identify common patterns of variation across multiple different features ➞ correlation analysis
  – inductive techniques don’t require pre-defined conditions

Variation as a nuisance parameter

▪ Many aspects of linguistic variation are nuisance parameters in corpus linguistics
  – e.g. difference in frequency of passives between AmE and BrE, as well as development from 1960s to 1990s (Unit #2)
  – ignore other dimensions such as genre/register variation by pooling frequency data from all texts of each corpus
  – corpus is analyzed as a random sample of VP tokens

▪ Consequences
  – variation ➞ non-randomness ➞ overestimate significance
  – discussed in much more detail in Unit #8
The multivariate approach

- Different linguistic features often show similar patterns of variation
- E.g. passives and nominalizations

The multivariate approach

- Different linguistic features often show similar patterns of variation
- E.g. passives and nominalizations
- Such correlations can be exploited to determine major dimensions of variation

The multivariate approach

- Multivariate analysis exploits correlations between features in order to determine latent dimensions
  - interpreted as underlying “causes” of variation
- An inductive, data-driven approach
  - no theoretical assumptions about linguistic variation and categories / sub-corpora to be compared
  - “multidimensional analysis” of register variation
- Related approaches: correspondence analysis, distributional semantics, topic modelling, …
Biber's multidimensional analysis (MDA)

### Table 3.3: Correlations used in the analysis of English

<table>
<thead>
<tr>
<th>Feature</th>
<th>Dimension 1 (Inferential)</th>
<th>Dimension 2 (Narrator)</th>
</tr>
</thead>
<tbody>
<tr>
<td>�</td>
<td>Newspeak verbs</td>
<td>-0.73</td>
</tr>
<tr>
<td>word length</td>
<td>-0.76</td>
<td></td>
</tr>
<tr>
<td>propositional phrases</td>
<td>0.48</td>
<td></td>
</tr>
<tr>
<td>type/token ratio</td>
<td>0.43</td>
<td></td>
</tr>
<tr>
<td>attributive adj.</td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>intra-correlation</td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td>inter-correlation</td>
<td>-0.47</td>
<td></td>
</tr>
</tbody>
</table>

### Table 3.4: Basic sentence patterns

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Feature</th>
<th>Dimension 1 (Inferential)</th>
<th>Dimension 2 (Narrator)</th>
</tr>
</thead>
<tbody>
<tr>
<td>®</td>
<td>Newspeak verbs</td>
<td>-0.73</td>
<td></td>
</tr>
<tr>
<td>word length</td>
<td>-0.76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>propositional phrases</td>
<td>0.48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>type/token ratio</td>
<td>0.43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>attributive adj.</td>
<td>0.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>intra-correlation</td>
<td>0.36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>inter-correlation</td>
<td>-0.47</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 3.5: Summary of the co-occurrence patterns underlying five major dimensions of English.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Feature</th>
<th>Dimension 1 (Inferential)</th>
<th>Dimension 2 (Narrator)</th>
</tr>
</thead>
<tbody>
<tr>
<td>®</td>
<td>Newspeak verbs</td>
<td>-0.73</td>
<td></td>
</tr>
<tr>
<td>word length</td>
<td>-0.76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>propositional phrases</td>
<td>0.48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>type/token ratio</td>
<td>0.43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>attributive adj.</td>
<td>0.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>intra-correlation</td>
<td>0.36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>inter-correlation</td>
<td>-0.47</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Pitfalls

- **Design bias:** choice of quantitative features
- **Design bias:** selection of text samples
- **Involves a miracle** — not clear what quantitative patterns are captured by FA
- **Magic number:** how many factor dimensions?
- **Interpretation bias** — arbitrary cutoff for feature weights (“loadings”)
- **Risk of reading one’s own expectations into features**
- **More subtle patterns of variation invisible**
- **Significance & reproducibility of results?**
**Reproducing Biber's dimensions**

- Sample of 923 medium-length published texts from written part of British National Corpus (BNC)
- Covers 4 different text types + male/female authors — academic writing, non-academic prose, fiction, misc.
- Biber features extracted automatically with Python script (Gasthaus 2007)
  - all frequencies normalized per 1000 words
  - data available in R package *corpora* ([BNCbiber](https://www.linguistik.fau.de))
- Factor analysis with 4 latent dimensions + varimax — seems to yield the most clearly structured analysis

**Design bias: choice of features**

- Correlated with noun frequency
- Correlated with verb frequency (all features measured per 1000 words)

**Design bias: choice of text samples**

- 4-Factor Analysis
And there's the magic number ... 

3-Factor Analysis (bootstrap sample #3)

Blindness to subtle patterns

- But research shows that author gender can be identified with high accuracy
  - Koppel et al. (2003): 77.3% with function words + POS n-grams
  - Gasthaus (2007): 82.9% with SVM on Biber features
- This dataset: 82.3% accuracy
  - baseline: 73.1%

Blindness to subtle patterns

Geometric Multivariate Analysis
(Diversy, Evert & Neumann 2014; Evert & Neumann 2017; Neumann & Evert 2021)

Online supplements:
https://www.stephanie-evert.de/PUB/EvertNeumann2017/
https://www.stephanie-evert.de/PUB/NeumannEvert2021/
Geometric Multivariate Analysis
(Diwery, Evert & Neumann 2014; Evert & Neumann 2017; Neumann & Evert 2021)

- Axiom: (Euclidean) distance = similarity of texts
  - depends crucially on theoretically motivated features
- Visualization → interpret geometric configuration
  - latent dimensions as geometric projections
  - orthogonal projection = perspective on data
  - method: principal component analysis (PCA)
- Minimally supervised intervention
  - based on externally observable, theory-neutral information
  - method: linear discriminant analysis (LDA)
- Bootstrapping / cross-validation to assess significance
- Cautious interpretation of feature weights

Feature design:
avoid “obvious” correlations

Feature scaling:
same contribution to Euclidean distances

Case study: CroCo
(Neumann 2013; Evert & Neumann 2017)

- CroCo: parallel corpus English/German
  - English-German and German-English translation pairs
  - we use 298 texts from 5 different genres
    (excluded: instruction manuals, tourism, fiction)
- 28 lexico-grammatical features (Neumann 2013)
  - comparable between languages
  - inspired by SFL and translation studies
- Text = point in 28-dimensional feature space
- Research hypotheses: shining through (Teich 2003), prestige effect (Toury 2012)
Feature scaling: optional signed log transformation

Latent dimensions as perspective on data configuration

- Instead of "magical" latent dimensions we focus on orthogonal projections as perspectives on the data
  - cf. photograph as 2D perspective on 3D object
- Different perspectives highlight different aspects
- Multivariate analysis ➞ choice of perspective
  - principal component analysis (PCA) = perspective that reflects distances between texts as accurately as possible
  - should reveal major dimensions of variation
  - advantage over factor analysis (FA): dimensionality does not have to be fixed a priori

CroCo: correlation matrix

CroCo: 3-dimensional projection
**CroCo: 4-dimensional projection**

- Focus on latent dim's 1 and 3 (register variation)
- Describe genre by centroid + ellipse
- Comparison with Hotelling’s $t^2$ test
  - essays vs. Web
  - $t^2=4.21$, df=2/141, $p=.0167$ *

**How about translationese?**

- PCA dim's can’t separate translations from original texts
  - 62.1% accuracy on first 3 PCA dim's
- But SVM machine learner can do this with >80% accuracy
  - RBF kernel
  - 10-fold c.v.
- Hints at shining through, but no clear-cut evidence

**Minimally supervised LDA**

- Add minimal amount of supervised knowledge to find a more informative perspective
  - evidence for shining through hypothesis from dimension that corresponds to contrast German vs. English
  - supervised knowledge: language of original texts only
- Linear discriminant analysis (LDA)
  - maximize separation between German / English originals
  - minimize variability within each group
  - classical technique related to PCA and ANOVA
- Project all texts onto LDA discriminant
  - complemented by additional PCA dim's for visualization
LDA significance: bootstrapping / cross-validation

- LDA is a supervised ML technique → overtrained?
  - Would we find the same discriminant if we trained on a different set of texts?
- Verification with bootstrap resampling or 10-fold cross-validation
  - LDA trained on 90% of data
  - Discriminant axis shows “wobble” of approx. 10°
- Discriminant scores from c.v. (10% test data per fold)
Case study 2: French regional varieties

(Diwersy, Evert & Neumann 2014)

- Lexical differences in regional varieties of French
- Two nation-wide newspapers each from 6 countries
  - Cameroon, France, Ivory Coast, Morocco, Senegal, Tunisia
  - two consecutive volumes from each newspaper
  - total size approx. 14.5 million tokens
- Text samples = one week each
- Features: frequencies of shared colligations
  - colligation = lemma-function pairs
  - must occur in all subcorpora with \( f \geq 100 \)

FRV: poor choice of features

PCA not excluding country-specific words as features: perfect separation

Design bias results in a completely uninteresting model

FA not applicable: features >> texts

FRV: PCA dimensions

Using only shared words as features, PCA no longer reveals any patterns (just a few outliers)

Use LDA to find a meaningful perspective, based on newspaper source

Country would presume regional varieties exist!

FRV: LDA dimensions (newspapers)
FRV: LDA dimensions (newspapers)

FRV: discriminant axes

References


SIGIL Unit #7