How to do Multi-dimensional Analysis: Theoretical background and practical analyses in R

Dr. Katharina Ehret

22 May 2024

universität freiburg

SFU SIMON FRASER UNIVERSITY ENGAGING THE WORLD

Discourse Processing Lab

This presentation is mainly based on

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge University Press: Cambridge.

Biber, Douglas. 1993. "The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Findings". *Computers and Humanities* 26 (5/6): 331-345.

Biber, Douglas, and Jesse Egbert. 2016. "Register Variation on the Searchable Web: A Multi-Dimensional Analysis". *Journal of English Linguistics* 44 (2): 95-137.

Outline

- Introduction
- First steps
- Factor analysis
- Microscopic analysis
- Criticism and new flavours
- Summary

Introduction

Defining register

- register and genre are often used interchangeably, however, they refer to distinct text varieties

- register
- analysis based on text excerpts
- linguistic characteristics:
  - any lexico-grammatical features (e.g. nouns, verbs)
  - (co-)occurrence of frequent features
  - communicative function
  - situational context

- genre
- analysis of entire texts
- linguistic characteristics:
  - specialized expressions (e.g. greeting formulars)
  - rhetorical organization and formatting
  - conventionally associated features

➪ register

(Biber and Conrad, 2009)

Theoretical background

- description of differences between speech and writing based on linguistic features (e.g. Chafe, 1982; DeVito, 1967; Kay, 1977; O'Donnell, 1974)
- drawbacks of previous research:
  - only a small number of texts and features analysed
    ➪ lack of representativity
  - too much significance assigned to particular texts/features
    ➪ biased results
  - often pairwise comparisons of registers
    ➪ no generalisation possible
  - communicative functions of selected registers not considered
    ➪ contradictory results

(Biber, 1988: 47-53)

The multi-dimensional approach

Douglas Biber

- introduced in "Variation across Speech and Writing" (1988)
- also known as multi-dimensional analysis or multi-feature approach
- methodology to analyse register variation:
  - describe (dis)similarities of registers on the basis of linguistic co-occurrence patterns which can be interpreted as variational dimensions

Theoretical assumptions

- different kinds of texts differ in regard to linguistic patterns and their communicative functions
- variation is caused by multiple factors
- no single dimension is adequate to represent linguistic variation in a text
- statistical co-occurrence of patterns relates to their functional properties
- dimensions are continua of linguistic variation rather than dichotomous poles

(Biber, 1993: 332)

Methodological properties

- quantitative ➪ based on feature frequencies
- multidimensional ➪ based on a large set of linguistic features
- macroscopic ➪ global analysis of variation based on different features across different texts
- microscopic ➪ detailed analysis of specific features in specific texts

(Biber, 1993: 332)

First steps

What is involved?

1. creating the dataset:
    - create representative database of texts
    - define linguistic features
    - retrieve feature frequencies
2. exploratory factor analysis:
    - determine co-occurence patterns of features
    - obtain statistical values to group feature patterns into dimensions
3. microscopic analysis:
    - interpret variational dimensions in terms of communicative and functional parameters

Defining features

- define a representative sample of features
  - ➪ include as many features as possible
  - ➪ include features described as functional markers in the literature
- features can have multiple functions independent of their grammatical category

➪ sixteen major grammatical categories

➪ 67 features in total

(Biber, 1988: 70-72)

Feature catalogue

1. tense and aspect markers
2. place and time adverbials
3. pronouns and pro-verb *do* (as in *she did it*)
4. wh-questions
5. nominal forms (e.g. *-tion*, *-ment*, gerunds, nouns)
6. passives
7. stative forms (e.g. existential *there*, full verb *be*)
8. subordination features
9. prepositional phrases, adjectives, adverbs
10. lexical specificity (type token ratio, mean word length)
11. lexical classes (e.g. downtoners, hedges, emphatics)
12. modals
13. specialised verb classes (e.g. private verbs, public verbs, *seem* and *appear*)
14. reduced forms and dispreferred structures (e.g. stranded prepositions, contractions, *that* deletion)
15. coordination
16. negation

(Biber, 1988: 73-75)

Feature retrieval

- annotation of features
- automatic retrieval of features
- normalisation of feature frequencies per 1,000 words

Feature retrieval

- annotation of features
- automatic retrieval of features
- normalisation of feature frequencies per 1,000 words

➪ input for multivariate analysis: exploratory factor analysis

Exploratory factor analysis

The basics

- primary tool for multi-dimensional approach
- statistical method for variable reduction
- identifies clusters of linguistic features based on their co-occurrence frequencies
- returns factors that represent the maximum amount of shared variation

➪ these factors are interpreted as dimensions of variation based on the communicative functions of the majority of shared features

(Biber, 1988: 79)

Factor analysis vs. Principal component analysis

- factor analysis (FA) accounts for shared variance
- factors change depending on the number of factors selected
- factor indeterminacy
  - ➪ there is no single unique factor solution
- the sample must have more observations than variables

➪ detect underlying theoretical constructs

(Hair et al., 2009)

- principal component analysis (PCA) accounts for the total variance
- PCA components are stable
- orthogonal projection preserves distances between data points

➪ obtain purely statistical summary of data

Technical procedure

1. calculate a correlation matrix of all variables
2. calculate factors
   - ➪ factors indicating shared variance of features
   - ➪ eigenvalues indicating how much variance is explained by a factor
3. include only the most important factors ➪ scree plot
4. use factor rotation to obtain simplified solution and facilitate interpretation
5. compute dimension scores (factor scores) for individual texts and mean dimension scores for individual registers

(Biber, 1988: 79-95)

Correlation matrix

	1st pers. pro.	*questions*	*passives*	*nominal-izations*
1st pers. pro.	*1.00*			
questions	*.85*	*1.00*		
passives	*−.15*	*−.21*	*1.00*	
nominal-izations	*.08*	*−.17*	*.90*	*1.00*

(adapted from Biber (1988: 79))

- negative coefficients: complementary co-variation
- positive coefficients: systematic co-occurrence
- R^2: percentage of variance shared by two variables

➪ correlation of feature frequencies defines factors

Correlation matrix

	1st pers. pro.	*questions*	*passives*	*nominal-izations*
1st pers. pro.	*1.00*			
questions	*.85*	*1.00*		
passives	*−.15*	*−.21*	*1.00*	
nominal-izations	*.08*	*−.17*	*.90*	*1.00*

(adapted from Biber (1988: 79))

- R^2 72% of shared variance
  - ➪ extremely high likelihood for 1st person pronouns and questions to co-occur in a text

From correlations to factors

	1st pers. pro.	questions	passives	nominal-izations
1st pers. pro.	1.00			
questions	.85	1.00		
passives	−.15	−.21	1.00	
nominal-izations	.08	−.17	.90	1.00

(adapted from Biber (1988: 79))

- factors are defined when several features are highly correlated

➪ Factor 1:

1st person pronouns + questions - passives - nominalisations

➪ Factor 2:

-1st person pronouns - questions + passives + nominalisations

Factor loadings

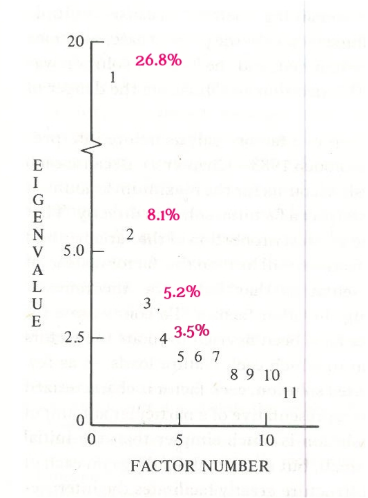
- factor loadings indicate the strength of correlation between a feature and a factor
- no one-to-one correspondence between correlation coefficients and loadings
  - ➪ the higher the absolute value, the more representative is a feature for a given factor

(adapted from Biber (1988: 79))

➪ Factor 1:
.82(1st person pronouns) + .82(questions) - .23(passives) - .11(nominalisations)

➪ Factor 2:
-.16(1st person pronouns) - .19(questions) + .91(passives) + .76(nominalisations)

Choosing the optimal number of factors



(adapted from Biber (1988: 83): Tab. 5.2)

- use a scree plot of eigenvalues
- guidelines:
  1. which factors explain most variance?
  2. search for breaks in plot
  3. include larger number of factors to avoid loss of information
  4. discard unnecessary factors

Factor solution

- factors represent maximum amount of shared variation
  - most features weigh on the first factor
  - underlying linguistic constructs of other factors are hidden

➪ use rotation technique to obtain simple factor solution

Table 5.3 *Rotated factor pattern for the 7 factor solution (Promax rotation)*

LX FEATURE	FACTOR1	FACTOR2	FACTOR3	FACTOR4	FACTOR5	FACTOR6	FACTOR7
PRO1	0.744	0.088	0.025	0.026	-0.089	0.008	-0.098
PRO2	0.860	-0.043	-0.018	0.016	0.007	-0.168	-0.064
PRO3	-0.053	0.727	-0.074	-0.018	-0.167	-0.076	0.138
PANY	0.618	0.046	0.011	0.085	-0.094	-0.085	-0.032
PDEM	0.756	-0.166	-0.001	-0.108	0.004	0.306	-0.077
PERFECTS	0.051	0.480	0.049	-0.016	-0.101	0.146	0.143
PASTTNSE	-0.083	0.895	0.002	-0.249	-0.049	-0.052	0.021
N	-0.799	-0.280	-0.091	-0.045	-0.294	-0.076	-0.213
N_NOM	-0.272	-0.237	0.357	0.179	0.277	0.129	-0.019
N_VBG	-0.252	-0.127	0.216	0.177	0.087	-0.052	0.052
PREP	-0.540	-0.251	0.185	-0.185	0.234	0.145	-0.008
ADVS	0.416	-0.001	-0.458	-0.020	-0.156	0.053	0.314
CONJNCTS	-0.141	-0.160	0.064	0.108	0.481	0.180	0.217
SUB_COS	0.661	-0.080	0.110	0.023	-0.061	0.078	-0.076
SUB_CON	0.006	0.092	0.100	-0.071	0.010	-0.056	0.300
SUB_CND	0.319	-0.076	-0.206	0.466	0.120	0.103	-0.007
SUB_OTHR	-0.109	0.051	-0.018	0.008	0.388	0.102	0.109
INF	-0.071	0.059	0.085	0.760	-0.274	-0.005	-0.074
PRO_DO	0.821	0.004	0.071	0.049	-0.057	-0.077	-0.056
SEEM	0.054	0.128	0.160	-0.010	0.015	0.045	0.348
DOWNTONE	-0.084	-0.008	0.021	-0.080	0.066	0.113	0.325
AMPLIFR	0.563	-0.156	-0.028	-0.124	-0.124	0.225	-0.018
PL_ADV	-0.417	-0.060	-0.492	-0.094	-0.067	-0.018	-0.023
TM_ADV	-0.199	-0.062	-0.604	-0.020	-0.290	0.116	-0.046
TH_CL	0.045	0.228	0.125	0.265	0.053	0.558	-0.122
ADJ_CL	-0.124	0.066	-0.080	0.123	0.171	0.360	0.183
CONTRAC	0.902	-0.100	-0.141	-0.138	-0.002	-0.057	-0.032
TYPETOKN	-0.537	0.058	0.002	-0.005	-0.311	-0.228	0.219
SYNTHNEG	-0.232	0.402	0.046	0.133	-0.057	0.176	0.110
NOT_NEG	0.778	0.149	0.017	0.125	0.019	0.001	0.037
BE_STATE	0.713	0.056	0.075	0.008	0.014	0.292	0.180
POS_MOD	0.501	-0.123	0.044	0.367	0.122	-0.022	0.115
NEC_MOD	-0.007	-0.107	-0.015	0.458	0.102	0.135	0.042
PRD_MOD	0.047	-0.056	-0.054	0.535	-0.072	0.063	-0.184
PUB_VB	0.098	0.431	0.163	0.135	-0.030	0.046	-0.279
PRV_VB	0.962	0.160	0.179	-0.054	0.084	-0.049	0.106
SUA_VB	-0.240	-0.035	-0.017	0.486	0.051	0.016	-0.237
PRTCLE	0.663	-0.218	-0.128	-0.029	-0.096	0.165	-0.140
GEN_HDG	0.582	-0.156	-0.051	-0.087	-0.022	-0.145	0.096
GEN_EMPH	0.739	-0.216	0.015	-0.027	-0.188	-0.087	0.210
SENT_REL	0.550	-0.086	0.152	-0.118	-0.025	0.048	-0.041
WH_QUES	0.523	-0.024	0.117	-0.111	-0.032	0.036	-0.094
P_AND	-0.253	-0.091	0.355	-0.066	-0.046	-0.324	0.126
O_AND	0.476	0.041	-0.052	-0.161	-0.139	0.218	-0.125
WHIZ_VBN	-0.382	-0.336	-0.071	-0.137	0.395	-0.128	-0.103
WHIZ_VBG	-0.325	-0.114	0.080	-0.169	0.212	-0.070	-0.093
CL_VBN	-0.025	-0.154	0.029	-0.050	0.415	-0.142	-0.059
CL_VBG	-0.211	0.392	-0.142	-0.076	0.268	-0.217	0.121
EX_THERE	0.262	0.108	0.113	-0.124	-0.004	0.318	0.017
DEM	0.040	-0.062	0.113	0.010	0.132	0.478	0.153
WRDLNGTH	-0.575	-0.314	0.270	-0.009	0.023	0.028	0.081

(adapted from Biber (1988: 86): Tab. 5.3)

Factor rotation

- each feature loads on as few factors as possible
- factors are based on the most significant/representative features only
  - ➪ only features with high factor loadings
  - ➪ Biber uses a cut-off of |.3|

Table 5.3 *Rotated factor pattern for the 7 factor solution (Promax rotation)*

LX FEATURE	FACTOR1	FACTOR2	FACTOR3	FACTOR4	FACTOR5	FACTOR6	FACTOR7
PRO1	0.744	0.088	0.025	0.026	-0.089	0.008	-0.098
PRO2	0.860	-0.043	-0.018	0.016	0.007	-0.168	-0.064
PRO3	-0.053	0.727	-0.074	-0.018	-0.167	-0.076	0.138
PANY	0.618	0.046	0.011	0.085	-0.094	-0.085	-0.032
PDEM	0.756	-0.166	-0.001	-0.108	0.004	0.306	-0.077
PERFECTS	0.051	0.480	0.049	-0.016	-0.101	0.146	0.143
PASTTNSE	-0.083	0.895	0.002	-0.249	-0.049	-0.052	0.021
N	-0.799	-0.280	-0.091	-0.045	-0.294	-0.076	-0.213
N_NOM	-0.272	-0.237	0.357	0.179	0.277	0.129	-0.019
N_VBG	-0.252	-0.127	0.216	0.177	0.087	-0.052	0.052
PREP	-0.540	-0.251	0.185	-0.185	0.234	0.145	-0.008
ADVS	0.416	-0.001	-0.458	-0.020	-0.156	0.053	0.314
CONJNCTS	-0.141	-0.160	0.064	0.108	0.481	0.180	0.217
SUB_COS	0.661	-0.080	0.110	0.023	-0.061	0.078	-0.076
SUB_CON	0.006	0.092	0.100	-0.071	0.010	-0.056	0.300
SUB_CND	0.319	-0.076	-0.206	0.466	0.120	0.103	-0.007
SUB_OTHR	-0.109	0.051	-0.018	0.008	0.388	0.102	0.109
INF	-0.071	0.059	0.085	0.760	-0.274	-0.005	-0.074
PRO_DO	0.821	0.004	0.071	0.049	-0.057	-0.077	-0.056
SEEM	0.054	0.128	0.160	-0.010	0.015	0.045	0.348
DOWNTONE	-0.084	-0.008	0.021	-0.080	0.066	0.113	0.325
AMPLIFR	0.563	-0.156	-0.028	-0.124	-0.124	0.225	-0.018
PL_ADV	-0.417	-0.060	-0.492	-0.094	-0.067	-0.018	-0.023
TM_ADV	-0.199	-0.062	-0.604	-0.020	-0.290	0.116	-0.046
TH_CL	0.045	0.228	0.125	0.265	0.053	0.558	-0.122
ADJ_CL	-0.124	0.066	-0.080	0.123	0.171	0.360	0.183
CONTRAC	0.902	-0.100	-0.141	-0.138	-0.002	-0.057	-0.032
TYPETOKN	-0.537	0.058	0.002	-0.005	-0.311	-0.228	0.219
SYNTHNEG	-0.232	0.402	0.046	0.133	-0.057	0.176	0.110
NOT_NEG	0.778	0.149	0.017	0.125	0.019	0.001	0.037
BE_STATE	0.713	0.056	0.075	0.008	0.014	0.292	0.180
POS_MOD	0.501	-0.123	0.044	0.367	0.122	-0.022	0.115
NEC_MOD	-0.007	-0.107	-0.015	0.458	0.102	0.135	0.042
PRD_MOD	0.047	-0.056	-0.054	0.535	-0.072	0.063	-0.184
PUB_VB	0.098	0.431	0.163	0.135	-0.030	0.046	-0.279
PRV_VB	0.962	0.160	0.179	-0.054	0.084	-0.049	0.106
SUA_VB	-0.240	-0.035	-0.017	0.486	0.051	0.016	-0.237
PRTCLE	0.663	-0.218	-0.128	-0.029	-0.096	0.165	-0.140
GEN_HDG	0.582	-0.156	-0.051	-0.087	-0.022	-0.145	0.096
GEN_EMPH	0.739	-0.216	0.015	-0.027	-0.188	-0.087	0.210
SENT_REL	0.550	-0.086	0.152	-0.118	-0.025	0.048	-0.041
WH_QUES	0.523	-0.024	0.117	-0.111	-0.032	0.036	-0.094
P_AND	-0.253	-0.091	0.355	-0.066	-0.046	-0.324	0.126
O_AND	0.476	0.041	-0.052	-0.161	-0.139	0.218	-0.125
WHIZ_VBN	-0.382	-0.336	-0.071	-0.137	0.395	-0.128	-0.103
WHIZ_VBG	-0.325	-0.114	0.080	-0.169	0.212	-0.070	-0.093
CL_VBN	-0.025	-0.154	0.029	-0.050	0.415	-0.142	-0.059
CL_VBG	-0.211	0.392	-0.142	-0.076	0.268	-0.217	0.121
EX_THERE	0.262	0.108	0.113	-0.124	-0.004	0.318	0.017
DEM	0.040	-0.062	0.113	0.010	0.132	0.478	0.153
WRDLNGTH	-0.575	-0.314	0.270	-0.009	0.023	0.028	0.081

(adapted from Biber (1988: 86): Tab. 5.3)

How does a factor look like?

FACTOR 2

past tense verbs	.90
third person pronouns	.73
perfect aspect verbs	.48
public verbs	.43
synthetic negation	.40
present participial clauses	.39

(present tense verbs	-.47)
(attributive adjs.	-.41)
(past participial WHIZ deletions	-.34)
(word length	-.31)

(adapted from Biber (1988: 89): Tab. 5.4)

Microscopic analysis

Interpreting variational dimensions

- identify widest shared communicative functions of features on a factor
  - interpret as textual dimensions
  - e.g. past tense, 3rd person pronouns, perfect aspect indicate narrative purposes ➪ narrative dimension
- interpretations are tentative and require confirmation
- compute factor scores to “confirm” hypothesised textual dimensions based on the distribution of texts/registers

(Biber, 1988: 91-93)

Factor scores

- calculate factor score for each text: sum of salient feature frequencies (based on normalised frequencies; with statistical cut-off point)
- take means to obtain average factor scores for registers
- plot average factor scores on textual dimensions ➪ register distribution

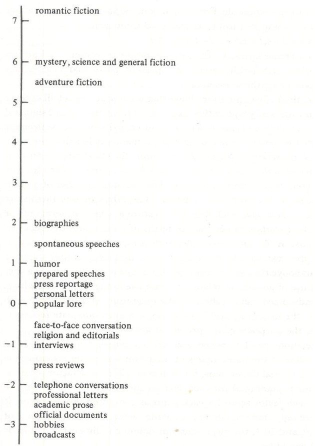


Figure 5.2 *Mean scores of Dimension 2 for each of the genres*
Dimension 2 ($F=32.30$, $p<.0001$, $R^*R=60.8\%$)

(adapted from Biber (1988: 96): Fig. 5.2)

Criticism, issues, and new flavours

Issue: text length

- classic MDA relies on relative feature frequencies
- requirement: minimum text length of at least 500 to 1000 words for reliable frequency estimates

➪ not suitable for short texts like Tweets or online comments

➪ multiple correspondence analysis (MCA)

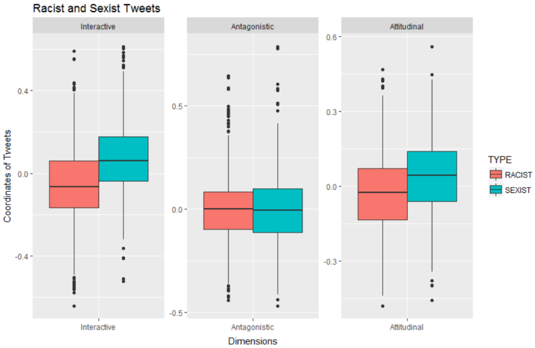
Multiple correspondence analysis

- dimension reduction method
- based on simple occurrence of lexico-grammatical features (i.e. absence vs. presence)
- returns positive or negative coordinates for each linguistic feature on each dimension ➪ similar coordinates indicate co-occurrence
- returns values indicating the variable contribution of features to this dimension
- returns positive or negative coordinates for individual texts ➪ similar coordinates indicate that texts share linguistic features
- interpretation of features as in classic MDA

(Clarke and Grieve, 2017)

MCA: Sexist vs. racist tweets

Figure 1: Boxplots of Racist and Sexist Tweets for Dimension 2, 3, and 4



(adapted from Clarke and Grieve (2017): Fig. 1)

- sexist tweets are more interactive and attitudinal
- ➩ raise new topics, ask questions to regain control
  - e.g. hashtags, question marks, wh-words
- ➩ attitudinal judgements to silence/dismiss previous tweets
  - e.g. comparatives, BE + predicative adjective, 1st ps pronouns

Criticism: feature selection and text-level patterns

- classic MDA depends on choice of selected features
  - strong theoretical assumptions or researcher's expectations
  - ➪ results may be influenced or "tweaked"
- classic MDA does not allow text-level investigations
  - focus on major dimensions only
  - broad feature patterns based on correlations do not reveal text-level distributions and variation
- poor visualisation

➪ Geometric multivariate analysis

(Diwersy et al., 2014)

Geometric multivariate analysis

- strongly visualisation-based approach to multivariate analysis
- ➪ visualise linguistic differences between texts in multidimensional feature space
- data-driven selection and weighting of linguistic features
- combination of PCA and supervised linear discriminant analysis (LDA)
- use of theory-neutral information as target for supervised learning
  - ➪ highlight subtle variational patterns

(Diwersy et al., 2014)

GMA step-by-step

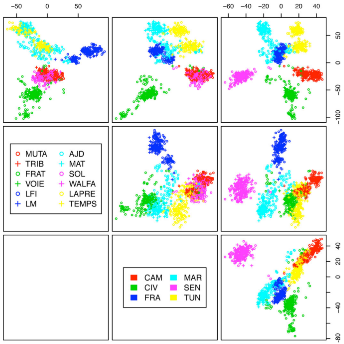
1. conduct PCA to obtain dimensions
   - inspect dimensions visually: do they capture all relevant linguistic patterns?
2. apply LDA to selected PCA dimensions
   - use pre-determined theory-neutral information as target for supervised learning
   - ⇨ detect more subtle patterns
3. validate LDA output via classification accuracy
4. visualise and inspect dimensions
5. (repeat)
6. visualise and interpret dimensions based on feature weights

(Diwersy et al., 2014)

GMA: PCA vs. LDA dimensions

- explore noun colligations in 12 different Francophone newspapers in 6 countries

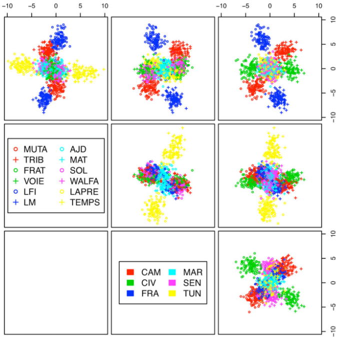
Fig 5: PCA



(adapted from Diwersy et al. (2014))

Fig: 6 LDA

with newspapers as target for LDA



Summary

What next

- MDA is a quantitative, corpus-based method to describe variation in texts
- what can we do with it?

What next

- MDA is a quantitative, corpus-based method to describe variation in texts
- what can we do with it?
  - ➪ categorise registers on the basis of Biber's dimensions
  - ➪ explore new registers
  - ➪ find new dimensions

What next

- MDA is a quantitative, corpus-based method to describe variation in texts
- what can we do with it?
  - ➪ categorise registers on the basis of Biber's dimensions
  - ➪ explore new registers
  - ➪ find new dimensions

➪ podcasts, online news comments and more

THANK YOU!

References I

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. (1993). The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Findings. *Computers and Humanities 26*(5/6), 331–345.

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Biber, D. (2012). Register as predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory 8*(1), 9–37.

Biber, D. (2014). Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contrast 14*(1), 7–34.

Biber, D. and S. Conrad (2009). *Register, genre, and style*. Cambridge University Press.

Biber, D. and J. Egbert (2016). Register Variation on the Searchable Web: A Multi-Dimensional Analysis. *Journal of English Linguistics 44*(2), 95–137.

Biber, D. and E. Finegan (1989). Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. *Text 9*(1), 93–124.

References II

Chafe, W. L. (1982). Integration and involvement in speaking, writing, and oral literature. In D. Tannen (Ed.), *Spoken and Written Language: Exploring Orality and Literacy*, pp. 35–54. Norwood, N.J.: Ablex.

Clarke, I. and J. Grieve (2017). Dimensions of abusive language on twitter. In *Proceedings of the first workshop on abusive language online*, pp. 1–10.

DeVito, J. A. (1967). Levels of abstraction in spoken and written language. *Journal of Communication 17*, 354–361.

Diwersy, S., S. Evert, and S. Neumann (2014). A weakly supervised multivariate approach to the study of language variation. *Aggregating dialectology, typology, and register analysis. linguistic variation in text and speech*, 174–204.

Godes, D. and D. Mayzlin (2004). Using Online Conversations to Study Word-of-Mouth Communication. *Marketing Science*, 545–560.

Hair, J. F., W. C. Black, B. J. Babin, and R. E. Anderson (2009). *Multivariate Data Analysis-7th*. Pearson Education Limited.

Kay, P. (1977). Language evolution and speech style. In B. G. Blount and M. Sanches (Eds.), *Sociocultural dimensions of language change*, pp. 21–33. New York: Academic Press.

Lampe, C. and P. Resnick (2004). Slash(dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Vienna, pp. 543–550.

References III

Napoles, C., J. Tetreault, E. Rosato, B. Provenzale, and A. Pappu (2017). Finding Good Conversations Online: The Yahoo News Annotated Comments Corpus. In *Proceedings of the 11th Linguistic Annotation Workshop*, Valencia, pp. 13–23.

Neumann, S. and S. Evert (2021). A register variation perspective on varieties of english. *Corpus Based Approaches to Register Variation*, 143–178.

Nini, A. Multidimensional Analysis Tagger.

North, S. (2007). 'The Voices, the Voices': Creativity in Online Conversation. *Applied Linguistics 28*(4), 538–555.

O'Donnell, R. C. (1974). Syntactic differences between speech and writing. *American Speech 49*, 102–110.

of the GM and Editor in Chief, O. (2015, November). Uncivil dialogue: Commenting and stories about indigenous people.

Pavalanathan, U., J. Fitzpatrick, S. F. Kiesling, and J. Eisenstein (2017, August). A Multidimensional Lexicon for Interpersonal Stancetaking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, pp. 884–895. Association for Computational Linguistics.

Wolff, A. G., K. A. Libbey, P. Kang, and P. D. Andrea. Unified online conversation application and platform.