Lectometry and latent variables

Koen Plevoets

FLAMES – KU Leuven

Overview

- Background of lectometry/lectometrics
- Objective of this presentation: Statistical modelling
- Models for lectometry/lectometrics
- Case study: COMURE
  - Research hypothesis
  - Data
  - Linguistic variables
  - Results
- Conclusion
- Appendix: R code
- Bibliography

Background

- Lectometry/Lectometrics = the measurement of distances between different “lects”, i.e. language variaties
- “Lect” = generalization of dialect, sociolect, mesolect,…
- Integration of
  - Stylometry and register analysis

    (Biber 1995; Luyckx 2010;…)
  - Dialectometry

    (Heeringa 2004; Szmrecsanyi 2013;…)
  - …

Background

- Geeraerts et al. (1999).
  - Clothing & football terms
  - Belgian & Netherlandic Dutch
  - Registers: quality newspapers, regional newspapers & labels in shop windows
  - Time periods: ’50s, ’70s & ’90s
- Soares da Silva (2010; 2014).
  - Replication of Geeraerts et al. (1999).
  - European & Brazilian Portuguese

Background

- Plevoets (2008).
  - Vernacular Belgian Dutch (“*tussentaal*”)
  - 14 registers of the Spoken Dutch Corpus
  - 5 social variables
- Ghyselen (2016).
  - Dialects in Ypres, Ghent & Antwerp
  - 5 situations: dialect test, regional conversation, supraregional conversation, interview & standard language test
  - 2 age groups: 25-35 & 50-65

Background

- Delaere (2015).
  - Translated vs. non-translated Dutch
  - 7 genres, based on the text typology in the Dutch Parallel Corpus
  - Source languages: English & French (& original Dutch)
- Prieels et al. (2015).
  - Subtitles vs. translations (& original Dutch)
  - 2 program genres: news & entertainment
  - 2 speaker types: voice-over & actor/interviewee
- …

Objective

- Goal of lectometry/lectometrics: scaling of linguistic distances
- = distances between language varieties (i.e. “lects”)
- Typically based on the frequency counts for variants of several linguistic variables
- Analytically the same as treating the varieties as observations for a multinomial/polytomous response factor
- SO, this paper: can we build (statistical) models for the frequency table of varieties by variants/variables?

Models

Individual observations

Multinomial response factor

	Variant 1.1	Variant 1.2	Variant 2.1	Variant 2.2	Variant 2.3	...	Variant J.1	Variant J.2
Variety A								
Variety B								
Variety C								
⋮								
Variety I								

Models

Individual observations

Multinomial response factor

	Variant 1.1	Variant 1.2	Variant 2.1	Variant 2.2	Variant 2.3	...	Variant J.1	Variant J.2
Variety A								
Variety B								
Variety C								
⋮								
Variety I								

We also have to take heed of the partitioning of the variants into variables (see later)

Models

- More specifically, the varieties are usually characterized in terms of extralinguistic factors:
  - Register
  - Region
  - Age/Period
  - …
- Two possible modelling techniques (as reference points):
  - Log-linear models
  - Correlation models
- This paper: “correspondence regression” = a special case of correlation models

Models

- Log-linear models for a response factor start from the null model of independence between response factor (say, index *j*) and all combinations among external factor (say, indices *k*, *l*,…)
  - I.e. variety *i* = register *k* & region *l* &…
- Every interaction between the response factor and an external factor/combination expresses the explanatory effect of that factor/combination on the response factor
- Formula: $\ell(\hat{P}_{ij}) = u_{I(i)} + u_{J(j)} + u_{JK(jk)} + u_{JL(jl)} + \cdots$

(Christensen 1997: 99-102)

- $\ell(\cdot)$ = "link function"

Interactions with response factor (*J*) express the predictive effect of each external factor (*K* or *L* or…).

Models

- Instead of interactions with the response factor (*J*), correlation models estimate scores for association terms
- Max. no. association terms M = min(I–1, J–1)
  - Because the scores on the association terms are unknown parameters (to be estimated), the association terms can be considered as “latent variables”
- Reduced models arise by choosing R < M
- Formula: $\ell\left(\hat{P}_{ij}\right) = P_i * P_j + P_i * P_j * \sum_{m=1}^{M} \sigma_m * \varphi_{im} * \gamma_{jm}$

(Gilula & Haberman 1988)

- $\ell(\cdot)$ = “link function”

Reduced (non-saturated) models arise by choosing R < M.

Models

- Correspondence regression estimates σ_m, $\varphi_{ikl\ldots m}$ and γ_{jm} by means of (Generalized) Least Squares
  - ~ Correspondence analysis: Singular Value Decomposition
- I.e. maximize σ_m (= correlation between $\varphi_{ikl\ldots m}$ and γ_{jm})
- (Computational) Constraints:
  - Mean(φ_{im}) = 0
  - Var(φ_{im}) = 1
  - Cor(φ_{im} , φ_{in}) = 0
  - Mean(γ_{jm}) = 0
  - Var(γ_{jm}) = 1
  - Cor(γ_{jm} , γ_{jn}) = 0

Models

- The partitioning of the linguistic variants into variables is modelled with conditional independence (of the varieties and variants given the variables)

  (Escofier 1984; Choulakian 1988; Van der Heijden et al. 1989)

- Confidence regions for the estimated parameters (σ_m, $\varphi_{ikl\ldots m}$ and γ_{jm}) are obtained with the Partial Bootstrap procedure

  (Alvarez et al. 2002; 2004; 2006; Lebart 2004)

  (Beh & Lombardo 2014: Chapter 8; Greenacre 2017: Chapter 29)

Case study

- COMURE (= COrpus-based Multivariate research of Register variation in translated and non-translated Belgian Dutch)
- 2010-2014
- PhD: Isabelle Delaere
- More specifically, case study 1 out of 4: Standardization

  (Delaere et al. 2012)

Case study

- Research hypothesis
- Data
- Linguistic variables
- Results

Research hypothesis

- (Universal of) Normalization

  (Baker 1993)

- Law of growing standardization

  (Toury 1995)

- “… translations tend to be less idiosyncratic, more conventionalized, than their originals”.

  (Chesterman 1997)

Research hypothesis

- Is normalization the same in
  - Different registers?
  - Different (source) languages?

Data

- Dutch Parallel Corpus
- 10 million words
- 3 languages
- 6 “text types”/registers

  (Macken et al. 2011)

Data

	Non-translated Dutch	Dutch translated from English	Dutch translated from French
Administrative texts	428 391	237 579	339 826
Journalistic texts	483 714	295 039	272 429
Instructive texts	106 640	0	45 371
External communication	371 154	311 493	261 640
Non-fiction	412 712	0	96 688
Fiction	0	0	116 178

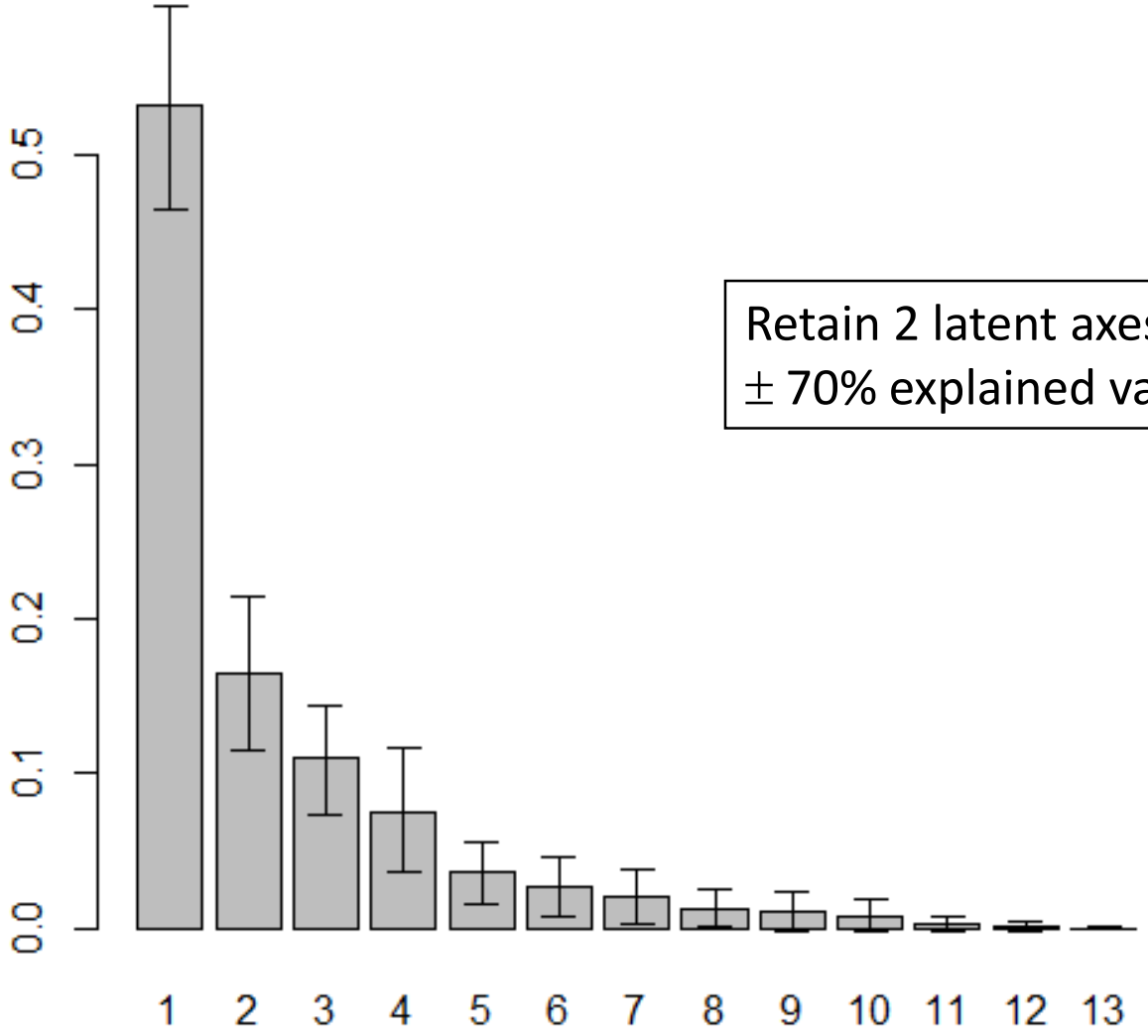
Linguistic variables

	Variant 1	Variant 2	Variant 3	Translation
1	akkoord gaan met	akkoord zijn met		to agree with
2	een van de	één van de		one of the
3	vd pv inf	pv inf vd	pv vd inf	order of the verbal end group
4	te veel	teveel		too much
5	ten(minste) – goed	ten(minste) – fout		at least
6	zulke + mv	zo'n + mv		such + plural
7	een beroep doen op	beroep doen op		to make an appeal to
8	zodra	van zodra		as soon as
9	verkrijgen	bekomen		to obtain
10	raken	geraken		to get
11	proberen te + inf	proberen + inf		to try to
12	op het eerste gezicht	op het eerste zicht		at first sight
13	beginnen te + inf	beginnen + inf		to start to

Results

- Correspondence regression of 27 variants in function of 18 varieties (= 6 registers * 3 languages) given 13 variables
- Done with the R package `corregp`

  (Plevoets 2015)



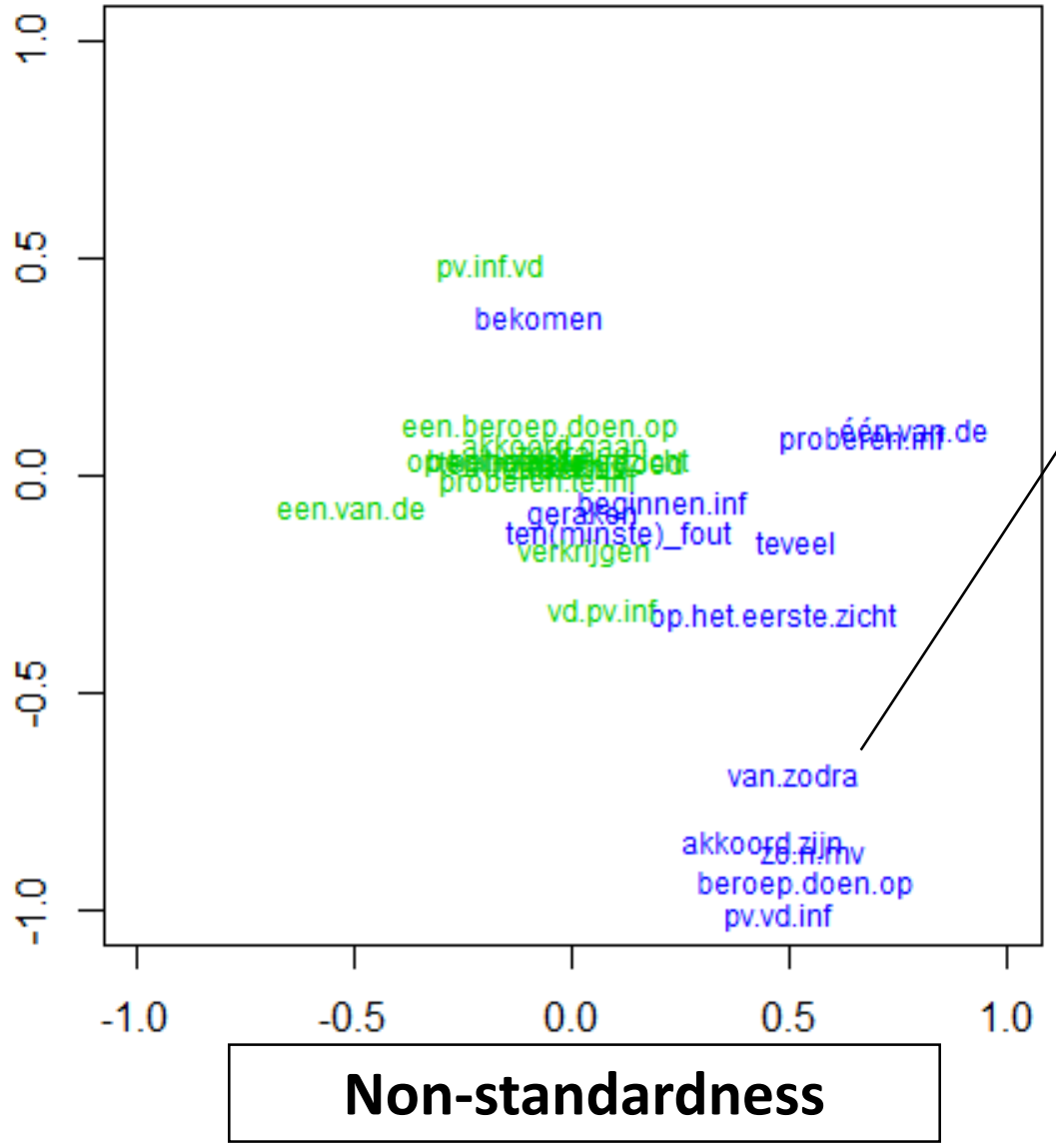
Retain 2 latent axes:
± 70% explained variation

ANOVA Table
(Type III effects)

	X^2	Lower (95%)	Upper (95%)
Register	370,3959	304,7964	448,1488
Language	83,5268	51,9235	120,4936
Register:Language	226,8943	176,9724	295,3711

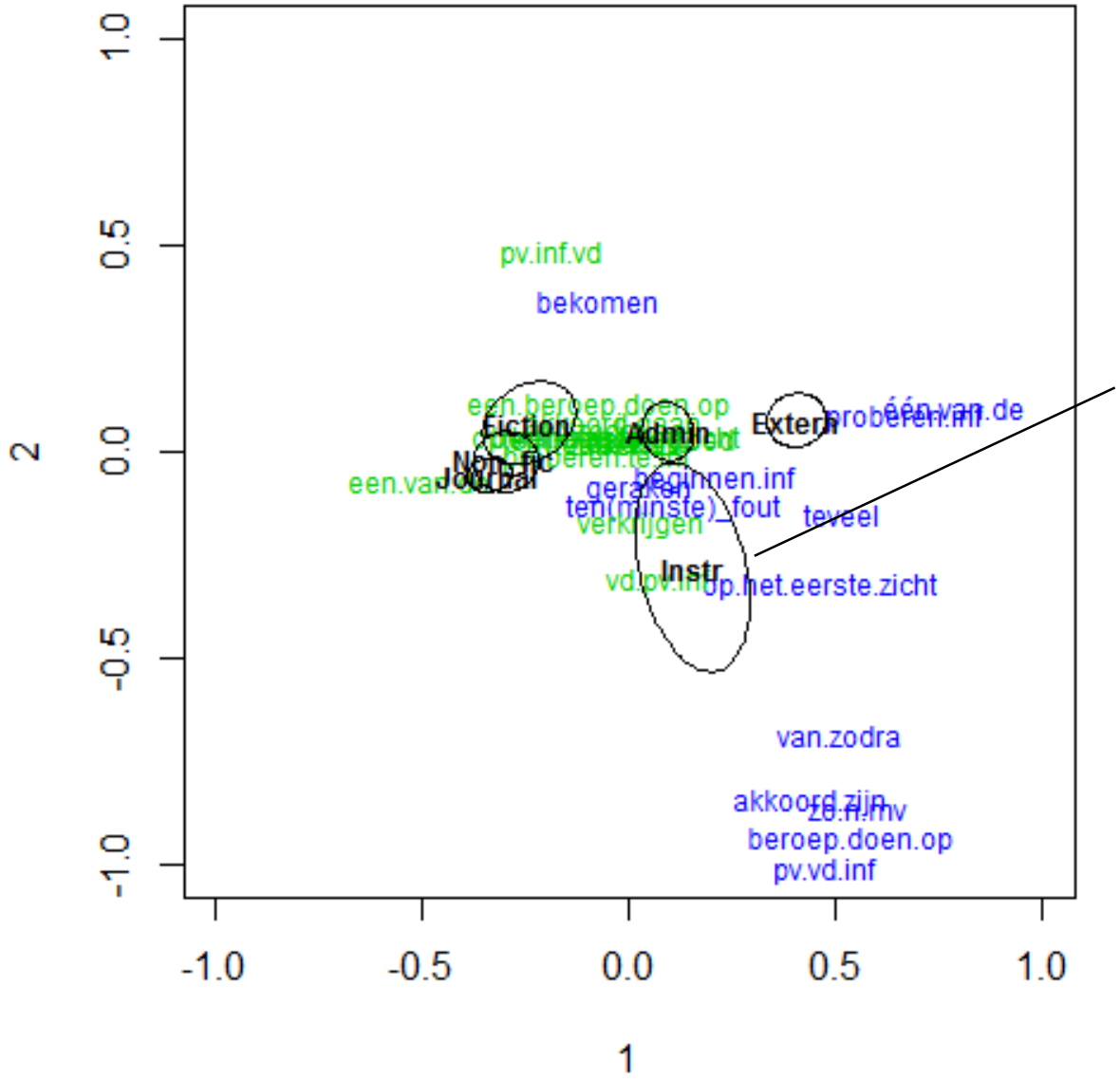
Both main effects and interaction significant predictors of the variation in the variants

Formality

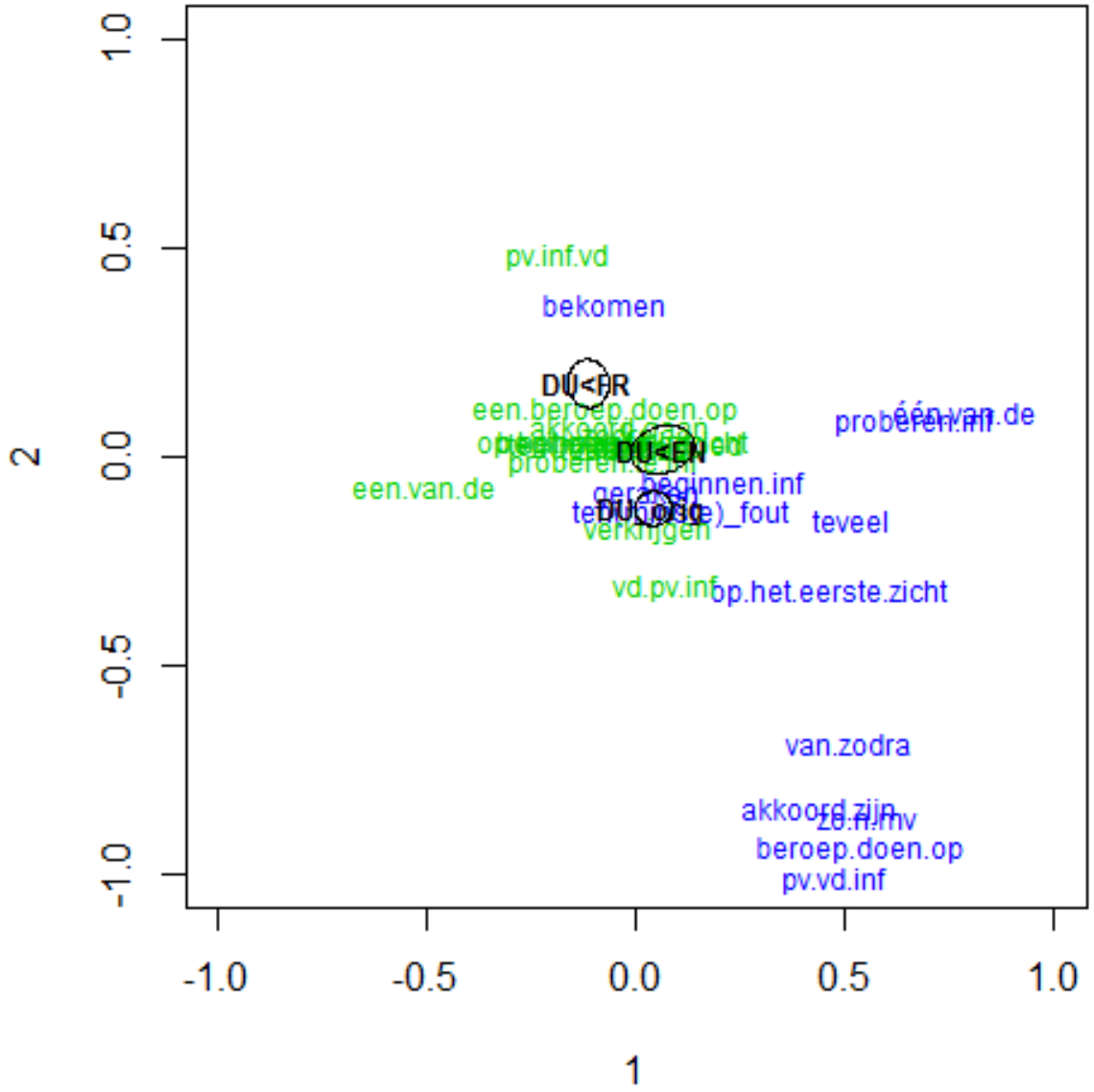


Non-standardness

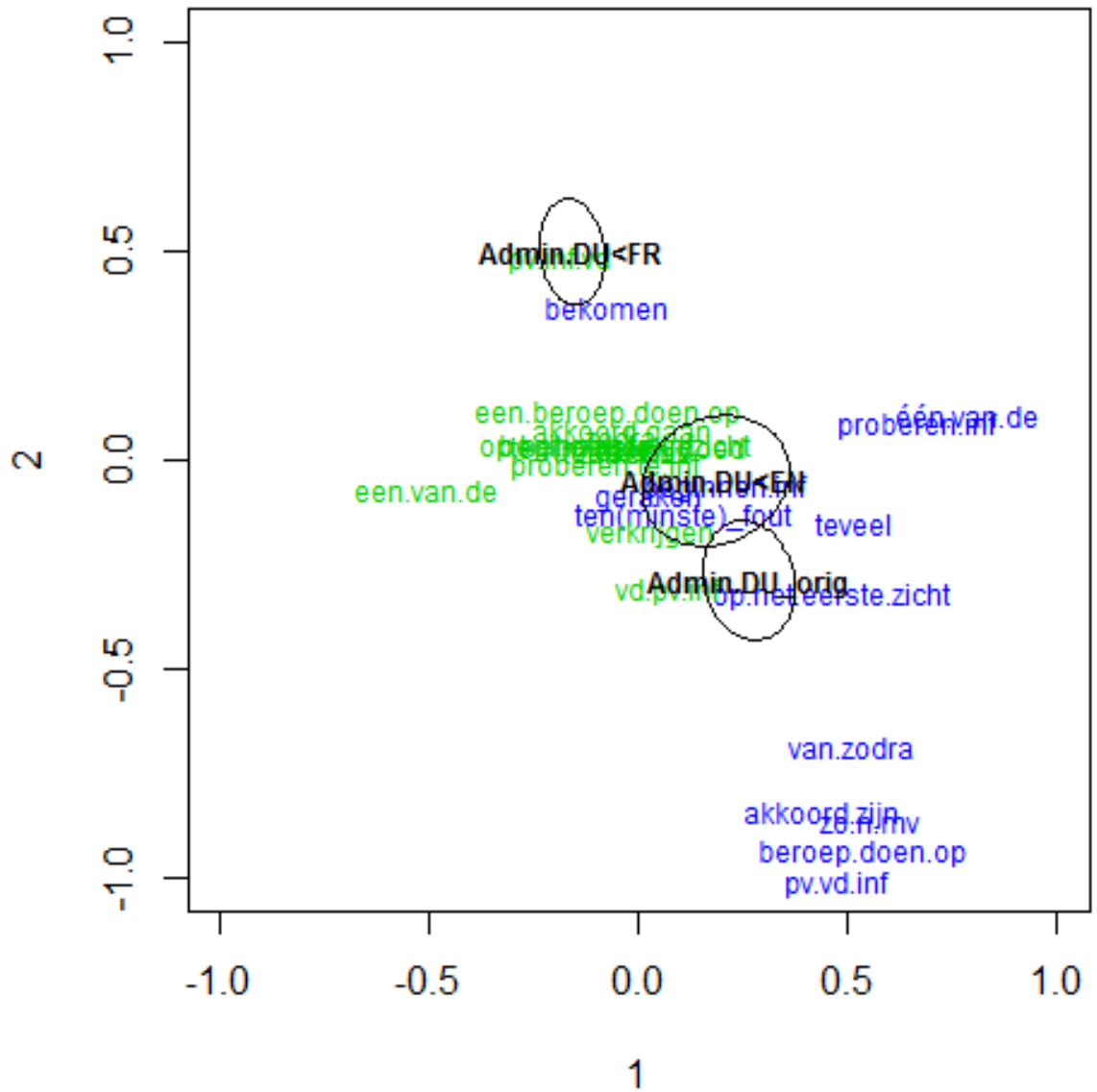
e.g.
van.zodra
=
0.51 * Non-standardness
+
–0.69 * Formality

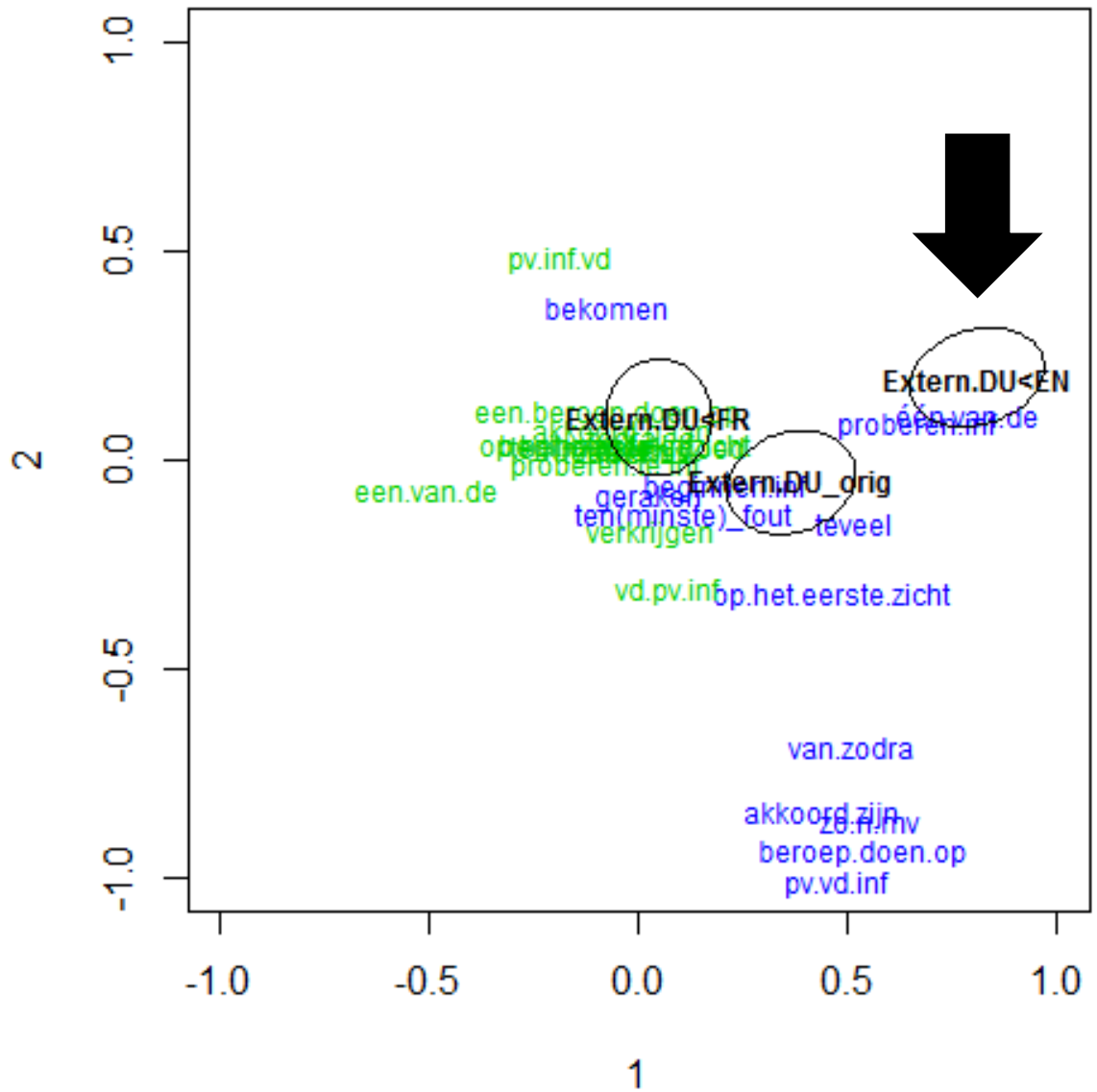


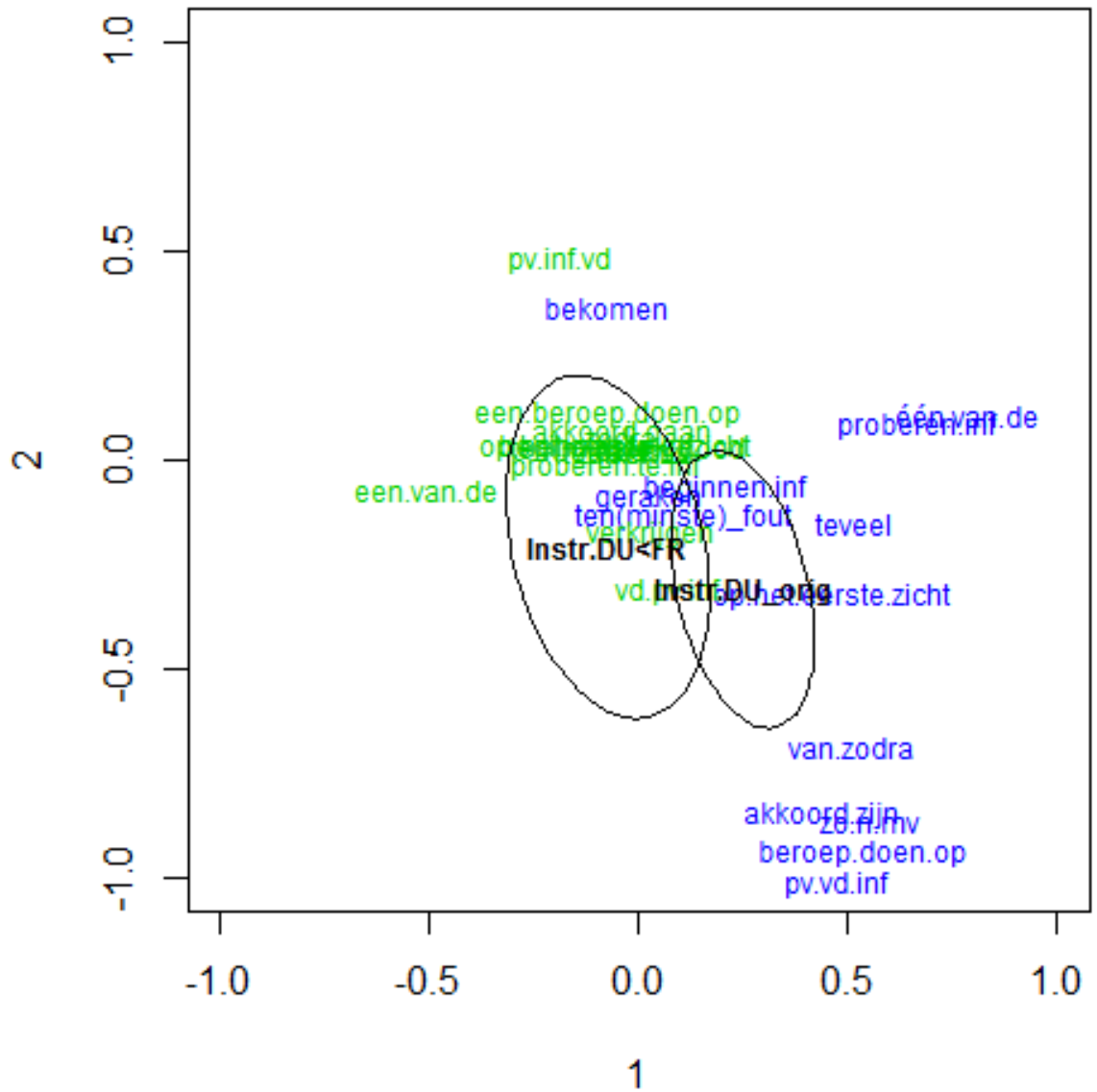
e.g.
Instr(uctive texts)
=
0.15 * Non-standardness
+
–**0.28** * Formality

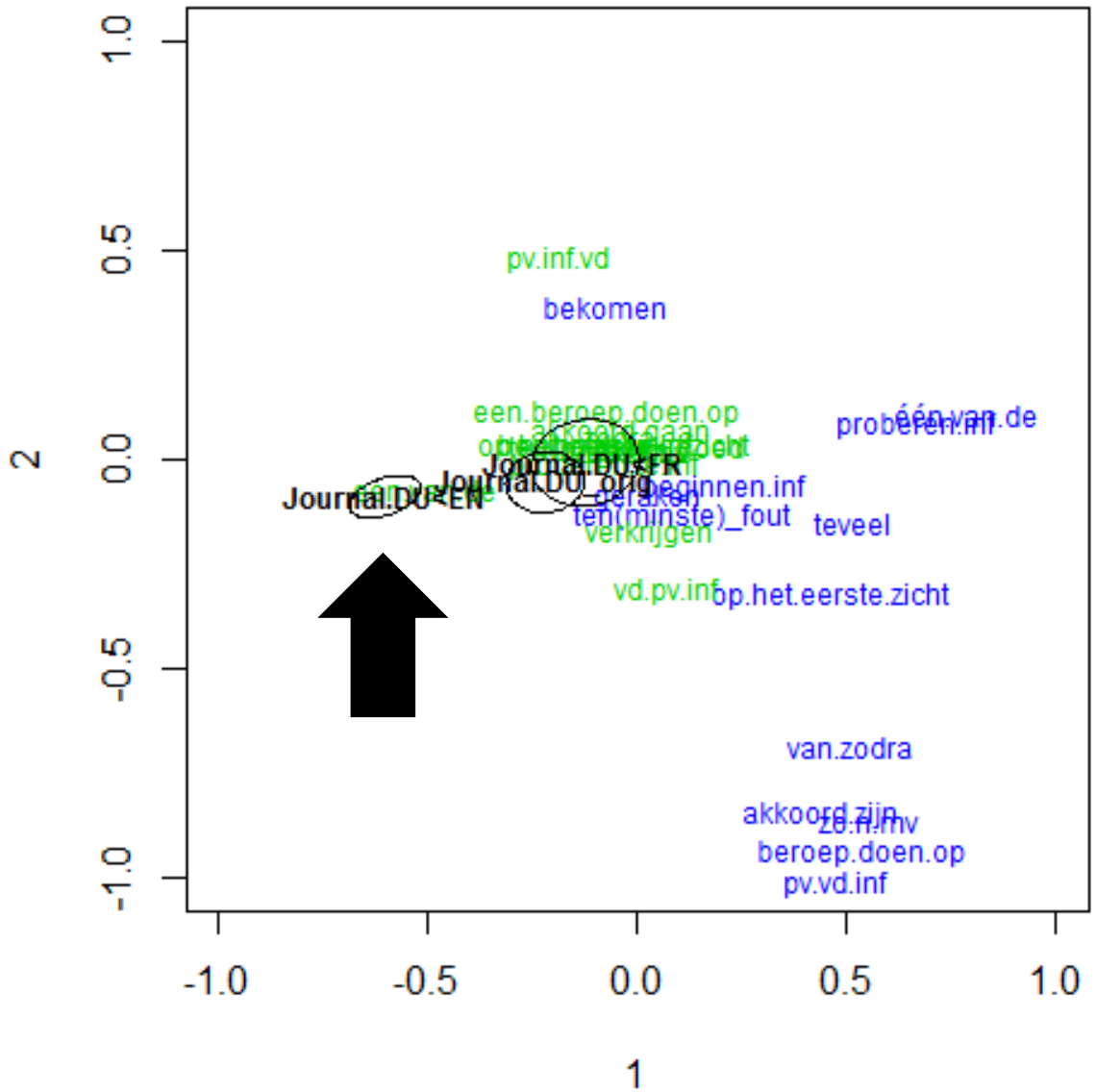


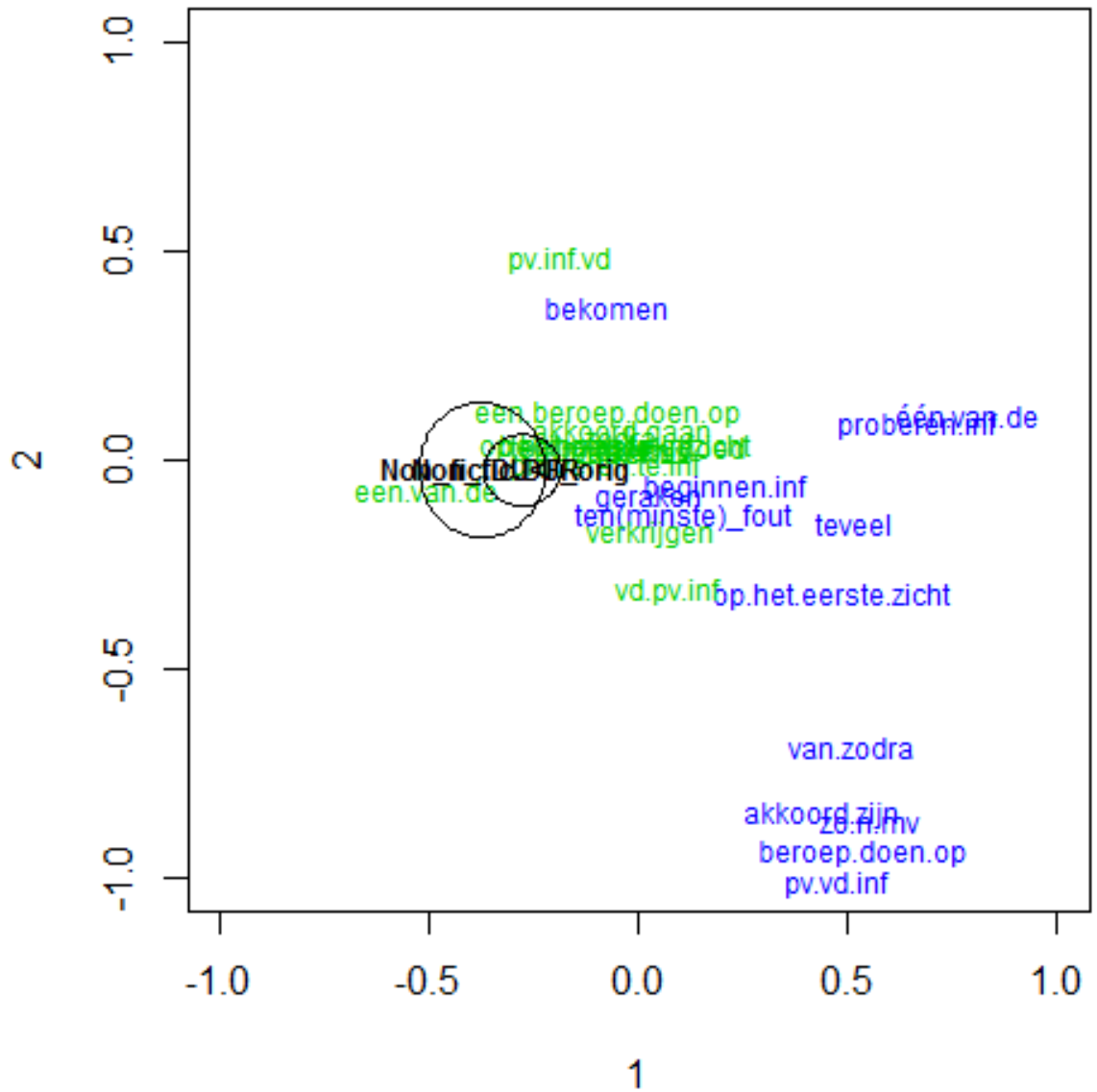
The languages differ to each other in terms of formality, **not** non-standardness: orginal Dutch contains the least formal language use.

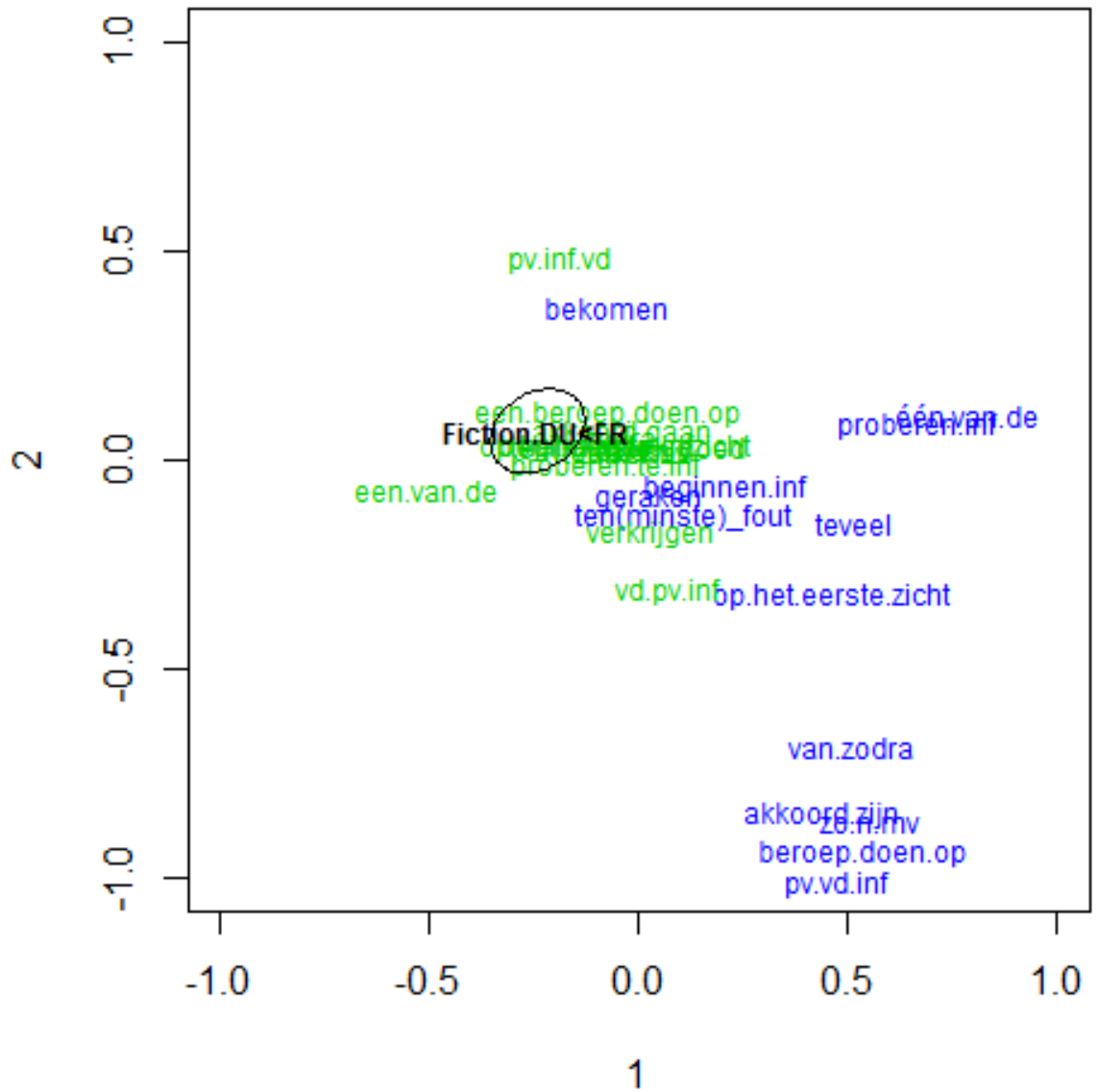

Admin.DU<FR
bekomen
een.beroep.doen.op
één.van.de
proberen.inf
een.van.de
Admin.DU<EU
teveel
Admin.DU_orig
van.zodra
akkoord.zijn
beroep.doen.op
pv.vd.inf
1.0
0.5
0.0
-0.5
-1.0
2
-1.0
-0.5
0.0
0.5
1.0
1

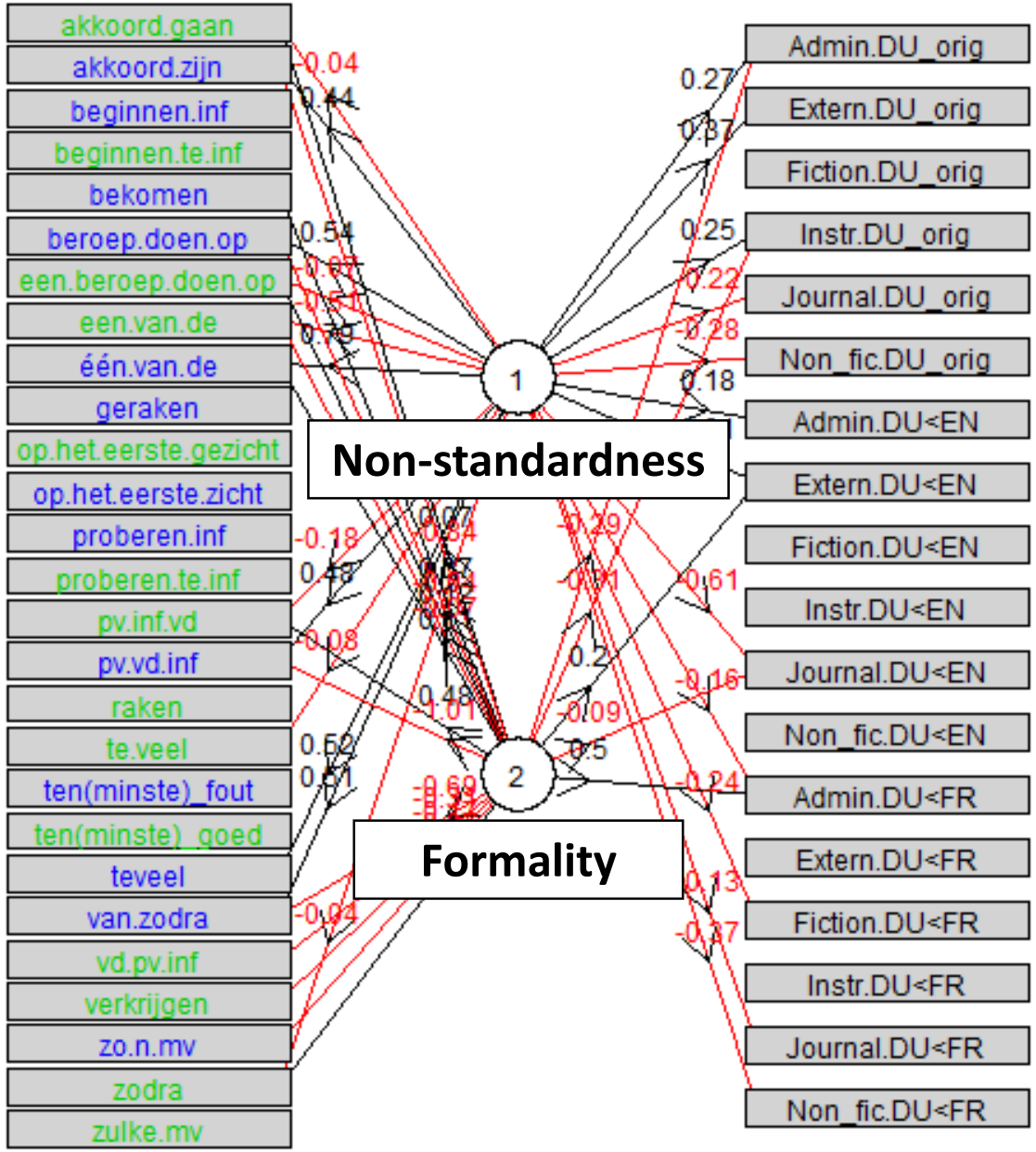

pv.inf.vd
bekomen
Extern.DU<EN
één.van.de
proberen.inf
Extern.DU<FR
een.van.de
Extern.DU_orig
geraken
ten(minste)_fout
teveel
verkrijgen
vd.pv.inf
op.het.eerste.zicht
van.zodra
akkoord.zijn
beroep.doen.op
pv.vd.inf
2
1
1.0
0.5
0.0
-0.5
-1.0
-1.0
-0.5
0.0
0.5
1.0


pv.inf.vd
bekomen
een.beroep.doen.op
een.van.de
beginnen.inf
teveel
Instr.DU<FR
Instr.DU_eng
van.zodra
akkoord.zijn
beroep.doen.op
pv.vd.inf
1.0
0.5
0.0
-0.5
-1.0
2
-1.0
-0.5
0.0
0.5
1.0
1


pv.inf.vd
bekomen
een.beroep.doen.op
proberen.inf
één.van.de
Journal.DU<FR
Journal.DU<EN
Journal.DU_orig
beginnen.inf
ten(minste)_fout
teveel
verkrijgen
vd.pv.inf
op.het.eerste.zicht
van.zodra
akkoord.zijn
beroep.doen.op
pv.vd.inf
2
1
1.0
0.5
0.0
-0.5
-1.0
-1.0
-0.5
0.0
0.5
1.0


pv.inf.vd
bekomen
een.beroep.doen.op
proberen.inf
een.van.de
beginnen.inf
geraken
ten(minste)_fout
teveel
verkrijgen
vd.pv.inf
op.het.eerste.zicht
van.zodra
akkoord.zijn
ze.n.mv
beroep.doen.op
pv.vd.inf
1.0
0.5
0.0
-0.5
-1.0
2
-1.0
-0.5
0.0
0.5
1.0
1


pv.inf.vd
bekomen
een.beroep.doen.op
Fiction.DU+FR
een.van.de
beginnen.inf
geraken
ten(minste)_fout
verkrijgen
teveel
vd.pv.inf
op.het.eerste.zicht
van.zodra
akkoord.zijn
beroep.doen.op
pv.vd.inf
1.0
0.5
0.0
-0.5
-1.0
2
-1.0
-0.5
0.0
0.5
1.0
1


akkoord.gaan
akkoord.zijn
beginnen.inf
beginnen.te.inf
bekomen
beroep.doen.op
een.beroep.doen.op
een.van.de
één.van.de
geraken
op.het.eerste.gezicht
op.het.eerste.zicht
proberen.inf
proberen.te.inf
pv.inf.vd
pv.vd.inf
raken
te.veel
ten(minste)_fout
ten(minste)_goed
teveel
van.zodra
vd.pv.inf
verkrijgen
zo.n.mv
zodra
zulke.mv
1
2
Non-standardness
Formality
Admin.DU_orig
Extern.DU_orig
Fiction.DU_orig
Instr.DU_orig
Journal.DU_orig
Non_fic.DU_orig
Admin.DU<EN
Extern.DU<EN
Fiction.DU<EN
Instr.DU<EN
Journal.DU<EN
Non_fic.DU<EN
Admin.DU<FR
Extern.DU<FR
Fiction.DU<FR
Instr.DU<FR
Journal.DU<FR
Non_fic.DU<FR

Conclusion

- The variation in the COMURE data can be accounted for by one underlying axis of Non-standardness and one underlying axis of Formality
- The 3 registers with high editorial control (*Journal*, *Non_fic* & *Fiction*) use standard language the most, and are not significantly different from each other
- Instructive texts is the register with more informal elements
- The 3 languages differ with each other in terms of formality, not standardness: translated Dutch tends to avoid informal elements more than original Dutch

Conclusion

- The interactions/combinations between register and language show some specific effects: the variation between the languages depends on register (and vice versa)
- Possible explanation (for future research): differences in the editing process?

Appendix: R code

```
> library(corregp)
> data(COMURE)
> comure.crg <- corregp(Variant~Register*Language, data=COMURE,
+ part="Variable", b=3000)

# Scree plot:
> screeplot(comure.crg, add_ci=TRUE, type="%")

# ANOVA Table:
> anova(comure.crg, nf=2)

# Colors for plotting:
> comure.col <- ifelse(xtabs(~Variant+Variety, data=COMURE)[,"Standard"] > 0,
+ "green3", "blue")

# (Mono)Plot of the variants:
> plot(comure.crg, xsub=NA, col_btm=comure.col, cex_btm=0.75,
+ hlim=c(-1,1), vlim=c(-1,1), add_ori=FALSE)
```

Appendix: R code

```
# (Bi)Plot of the registers:
> plot(comure.crg, x_ell=TRUE, xsub="Register", col_btm=comure.col, col_top="black",
+ cex_btm=0.75, cex_top=0.75, font_top=2, hlim=c(-1,1), vlim=c(-1,1), add_ori=FALSE)

# (Bi)Plot of the languages:
> plot(comure.crg, x_ell=TRUE, xsub="Language", col_btm=comure.col, col_top="black",
+ cex_btm=0.75, cex_top=0.75, font_top=2, hlim=c(-1,1), vlim=c(-1,1), add_ori=FALSE)

# (Bi)Plot of the interactions for Admin:
> plot(comure.crg, x_ell=TRUE, xsub=c("Admin.DU_orig","Admin.DU<EN","Admin.DU<FR"),
+ col_btm=comure.col, col_top="black", cex_btm=0.75, cex_top=0.75, font_top=2,
+ hlim=c(-1,1), vlim=c(-1,1), add_ori=FALSE)

# Idem for Extern, Instr, Journal, Non_fic and Fiction.

# Association graph:
> agplot(comure.crg, axes=1:2, xsub="Register.Language", cex=0.75, ycol=comure.col,
+ lcol=c("black","red"))
```

Bibliography

Alvarez, R., M. Becue, J.J. Lanero & O. Valencia. (2002). “Results stability in textual analysis: Its application to the study of Spanish investiture speeches (1979-2000)”. *Proceedings of the Journées internationales d'Analyse statistique des Données Textuelles 2002*, 1–12.

Alvarez, R., M. Becue & O. Valencia. (2004). “Etude de la stabilité des valeurs propres de l'AFC d'un tableaux lexical au moyen de procédures de rééchantillonage”. *Proceedings of the Journées internationales d'Analyse statistique des Données Textuelles 2004*, 42–51.

Alvarez, R., M. Becue & O. Valencia. (2006). “Partial bootstrap in CA: correction of the coordinates. Application to textual data”. *Proceedings of the Journées internationales d'Analyse statistique des Données Textuelles 2006*, 43–53.

Baker, M. (1993). “Corpus Linguistics and Translation Studies. Implications and Applications”. In: M. Baker, G. Francis & E. Tognini-Bonelli (eds), *Text and technology. In honour of John Sinclair*, Amsterdam: John Benjamins, 233–250.

Beh, E.J. & R. Lombardo. (2014). *Correspondence analysis. Theory, practice and new strategies*. Chichester: Wiley.

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

Chesterman, A. (1997). *Memes of translation: The spread of ideas in translation theory.* Amsterdam: John Benjamins.

Choulakian, V. (1988). “Exploratory analysis of contingency tables by loglinear formulations and generalizations of correspondence analysis”. *Psychometrika* 53 (2), 235–250.

Christensen, R. (1997). *Log-linear models and logistic regression*. New York: Springer.

Delaere, I. (2015). *Do translations walk the line? Visually exploring translated and non-translated texts in search of norm conformity*. Ghent: Doctoral dissertation.

Delaere, I., G. De Sutter & K. Plevoets. (2012). “Is translated language more standardized than non-translated language? Using profile-based correspondence analysis for measuring linguistic distance between language varieties.” *Target* 24 (2), 203–224.

Escofier, B. (1984). “Analyse factorielle en référence à un modèle: Application à l'analyse de tableaux d‘échanges”. *Revue de Statistique Appliquée* 32 (4), 25–36.

Gilula, Z. & S.J. Haberman (1988). “The analysis of multivariate contingency tables by restricted canonical and restricted association models”. *Journal of the American Statistical Association* 83 (403), 760–771.

Ghyselen, A.-S. (2016). *Verticale structuur en dynamiek van het gesproken Nederlands in Vlaanderen. Een empirische studie in Ieper, Gent en Antwerpen*. Ghent: Doctoral dissertation.

Bibliography

Geeraerts, D., S. Grondelaers & D. Speelman. (1999). *Convergentie en divergentie in de Nederlandse woordenschat. Een onderzoek naar kleding- en voetbaltermen*. Amsterdam: Meertens Instituut.

Greenacre, M. (2017). *Correspondence analysis in practice. Third edition*. Boca Raton: Chapman & Hall/CRC Press.

Heeringa, W. (2004). *Measuring dialect pronunciation differences using Levenshtein distance*. Groningen: Doctoral dissertation.

Lebart, L. (2004). “Validité des visualisations de données textuelles”. *Proceedings of the Journées internationales d'Analyse statistique des Données Textuelles 2004*, 708–715.

Luyckx, K. (2010). *Scalability issues in authorship attribution*. Antwerp: Doctoral dissertation.

Macken, L., O. De Clercq & H. Paulussen. (2011). “Dutch Parallel Corpus: A balanced copyright-cleared parallel corpus”. *Meta* 56 (2): 374–390.

Plevoets, K. (2008). *Tussen spreek- en standaardtaal. Een corpusgebaseerd onderzoek naar de situationele, regionale en sociale verspreiding van enkele morfosyntactische verschijnselen in het gesproken Belgisch-Nederlands*. Leuven: Doctoral dissertation.

Plevoets, K. (2015). *corregp: Functions and Methods for Correspondence Regression*. R package.

Prieels, L., I. Delaere, K. Plevoets & G. De Sutter. (2015). “A corpus-based multivariate analysis of linguistic norm-adherence in audiovisual and written translation”. *Across Languages and Cultures* 16 (2), 209–231.

Soares da Silva, A. (2010). “Measuring and parameterizing lexical convergence and divergence between European and Brazilian Portuguese”. In: D. Geeraerts, G. Kristiansen & Y. Peirsman (eds.), *Advances in Cognitive Sociolinguistics*. Berlin: de Gruyter, 41–83.

Soares da Silva, A. (2014). “The pluricentricity of Portuguese: A sociolectometrical approach to divergence between European and Brazilian Portuguese”. In: A. Soares da Silva (ed.), *Pluricentricity: Language variation and sociocognitive dimensions*. Berlin: de Gruyter, 143–188.

Szmrecsanyi, B. (2013). *Grammatical variation in British English dialects: A study in corpus-based dialectometry*. Cambridge: Cambridge University Press.

Toury, G. (1995). *Descriptive translation studies and beyond.* Amsterdam: John Benjamins.

Van der Heijden, P.G.M., A. De Falguerolles & J. De Leeuw. (1989). “A combined approach to contingency table analysis using correspondence analysis and log-linear analysis”. *Applied Statistics* 38 (2), 249–292.