On the inevitability of multivariate statistics in corpus-based translation studies: some whys and hows

In collaboration with Isabelle Delaere

Gert De Sutter

gert.desutter@hogent.be
Faculty of Applied Language Studies
University College Ghent

HoGent

Purpose

1. Argue that multivariate statistics is an inevitable part of corpus-based translation studies
   → Because CBTS community is ready for it
2. Explore the field of what could constitute *multivariate corpus-based translation studies*
   → Multidimensionality of translational behaviour

Purpose

1. Argue that multivariate statistics is an inevitable part of corpus-based translation studies
   → why?
2. Explore the field of what could constitute *multivariate corpus-based translation studies*
   → what and how?

A time to look back

1993

A major year for the public life of corpus-based translation studies (CBTS)

- Baker, M. (1993). “Corpus Linguistics and Translation Studies: Implications and Applications.”

A time to look back

The importance of Baker's seminal paper:

- Introduction of corpus methodology in TS
- Shift of focus: from the primacy of the source text to the position of the target text in the target system
- Highly attractive research agenda (translation universals, translational norms)

→ Institutionalisation of CBTS as a major paradigm within TS while setting it apart from neighbouring disciplines (esp. linguistics)

A time to look back

However, 20 years later, and many publications later, Baker sometimes appears to be public enemy n° 1 in CBTS

- “It is therefore misleading to call explicitation a (possible) universal of translation, as e.g. Baker (1993, 1996) does.” (Becher 2010)
- Also: House 2008, Bernardini & Ferraresi 2011, Kruger & van Rooy 2012, Hansen-Schirra, Neumann, Steiner 2012, our own work

A time to look back

Much of the debate centres around the universality assumption

- “features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems” (Baker 1993: 243)

A time to look back

Much of the debate centres around the universality assumption

- "features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems" (Baker 1993: 243)

→ Translation features are pervasive, hence irrespective of other factors, such as genre

A time to look back

Much of the debate centres around the universality assumption

- "features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems" (Baker 1993: 243)

→ Translation features are pervasive, hence independent of source language

A time to look back

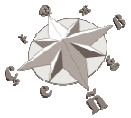
Two of the most obvious potentially influencing factors of linguistic behaviour in translations were relegated from the very beginning

- Bernardini & Ferraresi (2011): interference and normalisation
- Kruger & van Rooy (2012): register variation within translations

A time to look back

More or less ackowledged by Baker herself (1999)

- “we take the view that language in general, and the language of translation in particular, reflects constraints which operate in the context of production and reception: these constraints are social, cultural, ideological, and of course also cognitive in nature”
- “Are certain linguistic features or strategies more likely to occur in certain types of translation genres, like translated fiction, news, inflight magazines?”



A time to grow up

20 years after Baker's seminal paper, CBTS has come of age

→ How can we adequately capture the multidimensionality of linguistic behaviour in translations?

A time to grow up

Answer can be found in Baker 1993:

"It [the paper] argues that the techniques and methodology developed in the field of corpus linguistics will have a direct impact on the emerging discipline of translation studies […]."

→ The institutionalisation of CBTS also led to an auto-isolation wrt neighbouring disciplines

A time to grow up

Let's tie up with common practices in corpus linguistics:

"multifactorial data must be analyzed multifactorially (Gries 2010: 143)

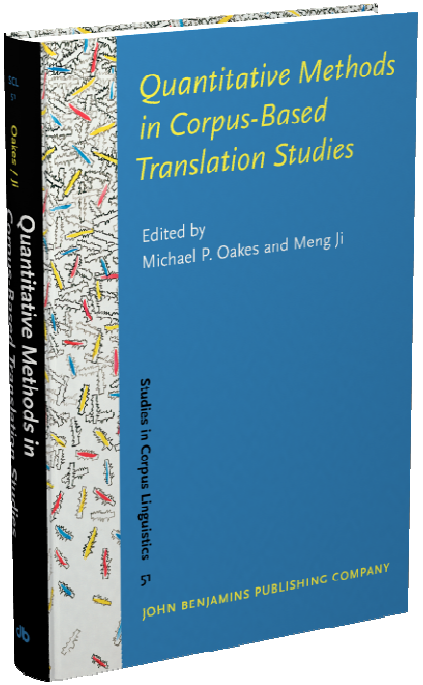
CBTS community is ready for it

Purpose

1. Argue that multivariate statistics is an inevitable part of corpus-based translation studies
   → why?
2. Explore the field of what could constitute *multivariate corpus-based translation studies*
   → what and how?

Explore the field of investigation

Depending on the research question, different multivariate statistics are needed in order to adequately grasp the multifactorial nature of translation behaviour



Correspondence analysis

Logistic regression analysis

Mixed-effect modelling

HoGent

Central hypothesis: conservatism

Conservatism hypothesis: translators tend to use more often conservative language than non-translators (Baker 1996, 2004)

→ Operationalisation: formal (archaic) lexemes

Data, method, analysis

Dutch Parallel Corpus (DPC)

- 10 M tokens
- Parallel and comparable corpus: Dutch is SL (> FR, EN) and TL (< FR, EN)
- Stratified in text types: ADMIN, JOURNAL, INSTR, NON-FIC, FIC

Data, method, analysis

- This case study: selection from DPC
  1. Only Dutch data (target-oriented)
  2. Exit Dutch Dutch data (n = ca. 1.5 M)
  3. Exit fiction (only available for FR>NL; n = 116.178)
  4. Exit data with unknown source language (n = 313.774)

Data, method, analysis

	Non-translated Dutch	Translated Dutch (< EN)	Translated Dutch(< FR)
ADMIN	428,391	237,579	339,826
JOURNAL	483,714	295,039	272,429
INSTR	106,640	0	45,371
EXTERNAL	371,154	311,493	261,640
NON-FIC	412,712	0	96,688

Total: n = 3.662.676

Data, method, analysis

How to measure formality adequately?

Profile-based method (Speelman et al. 2003)

- Central idea: lexical variation implies that language users have different options to express a given concept
- Profile = the set of synonyms designating a concept (e.g. *underground vs. subway*)
- Crucial to use a profile-based method when studying lexical variation

Data, method, analysis

Let us illustrate what can go wrong

– When studying a formal lexeme as *tewerkstelling* (job employment) in isolation

	Admin	Extern	Instr	Journal	Non_Fic
Abs. freq.	79	37	2	8	11
Norm freq.	0.78	0.52	0.35	0.07	0.21

Difficult to interpret

High frequency of *tewerkstelling* in ADMIN could mean either 'this is a formal register' or 'this register is often about job employment'

Data, method, analysis

Profile-based method

– If you want to know how lexically formal a register is, the proportion of neutral words should be taken into account too

	Admin	Extern	Instr	Journal	Non_Fic
Tewerkstelling	26,9	33,9	66,7	11,1	54,2
Werk-gelegenheid	73,1	66,1	33,3	88,9	45,8

Data, method, analysis

- Extracting 10 profiles
  - Profile = set of synonymous onomasiological variants to express a certain concept (i.c. formal vs. neutral)
  - 10 is more than 1
- Profile selection
  1. Based on dictionaries and style guides (without contradiction)
  2. Only difference in formality allowed

Data, method, analysis

Profile number	Formal variant	Neutral variant
1	*Echter* (n = 1452)	*Maar* (n = 9328)
2	*Alsook; alsmede* (n = 74; 299)	*Evenals* (n = 171)
3	*Trachten* (n = 133)	*Proberen* (n = 552)
4	*Thans* (n = 170)	*Nu* (n = 2902)
5	*Bekomen* (n = 199)	*Verkrijgen* (n = 409)
6	*Indien* (n = 163)	*Als* (n = 720)
7	*Te* + plaats (n = 264)	*In* + plaats (n = 3208)
8	*Reeds* (n = 778)	*Al* (n = 3774)
9	*Dienen* + Inf (n = 826)	*Moeten* + Inf (n = 1238)
10	telwoord + *maal* (n = 97)	telwoord + *keer* (n = 411)

Total: n = 27.166

Data, method, analysis

	Admin	Extern	Instr	Journal	Non-fic
Al	296	254	67	731	634
Reeds	161	118	33	37	87
Dienen+inf	251	64	180	18	32
Moeten+inf	193	47	36	137	98

How to discern patterns in this table?

Complex tables with different variables included ask for multivariate statistics

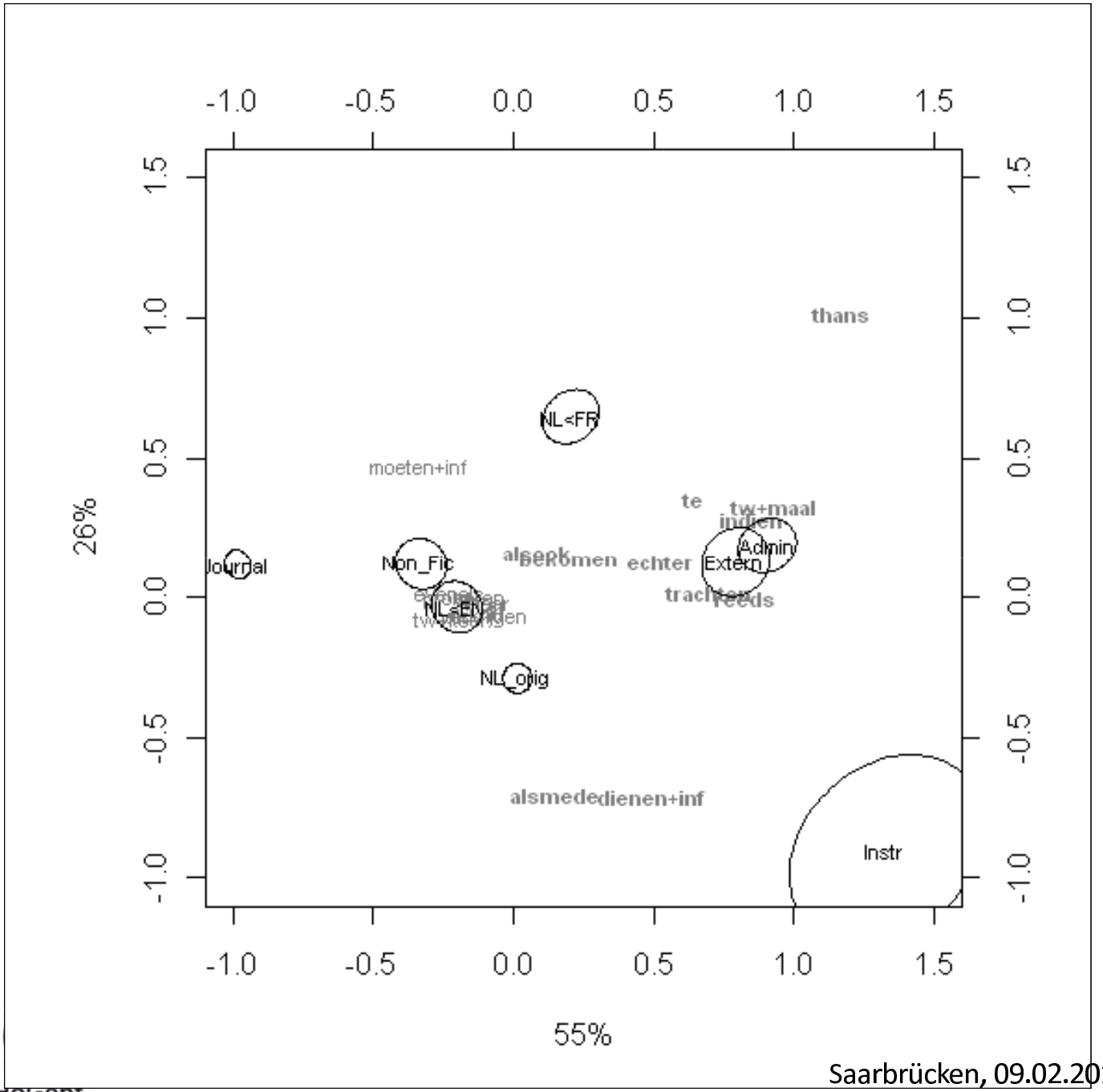
Conservatism, part 1

Lectometric research goal

- Measure formality distances between different lects of translated and non-translated Dutch texts
  - Lects = varieties (text types, translations vs. non-translations)
  - Underlying idea: the more lexical use differs between lects, the bigger the distance between these lects

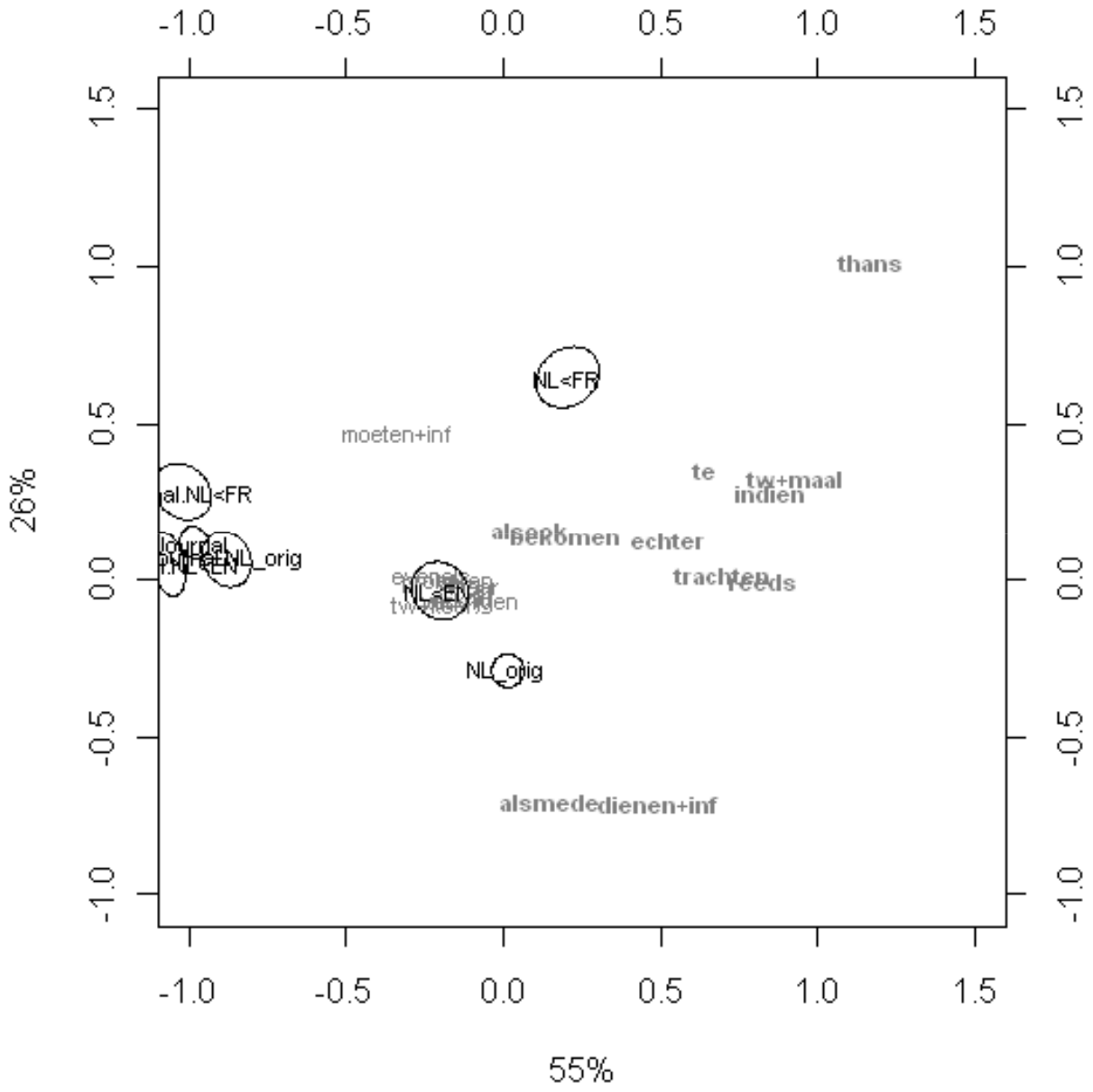
Multivariate statistics, part 1

- Profile-based correspondence analysis
  - Explorative technique for data reduction
  - Conditional analysis of associations between rows and columns (taking into account the profile structure)
  - Visual representation of these associations in a low-dimensional plot

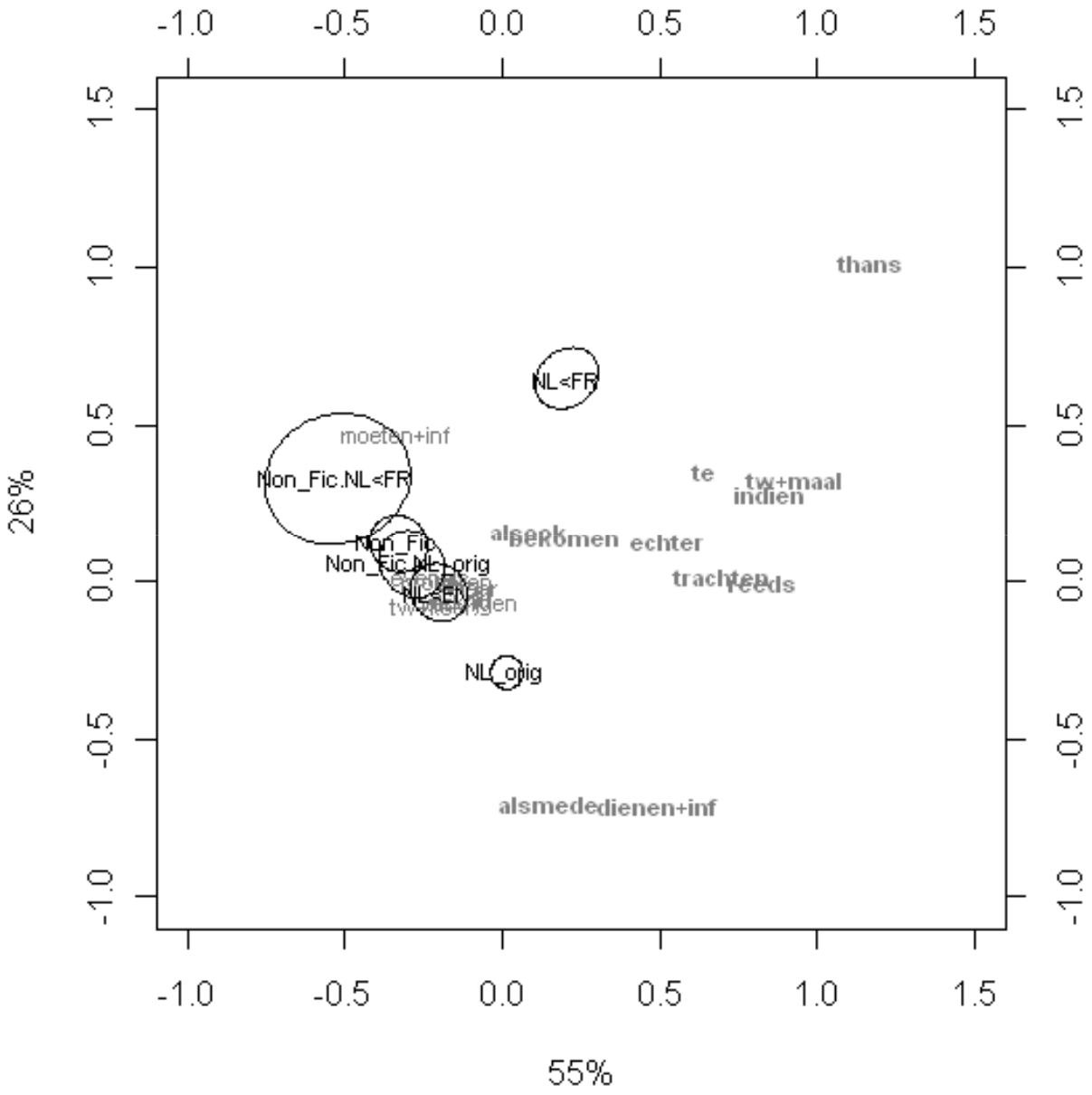

-1.0
-0.5
0.0
0.5
1.0
1.5
26%
55%
thans
NL<FR
moeten+inf
te
tw+maal
indien
Admin
Journal
Non_Fic
als
echter
Extern
NL<EN
NL_orig
alsmededienen+inf
Instr

Multivariate statistics, part 1

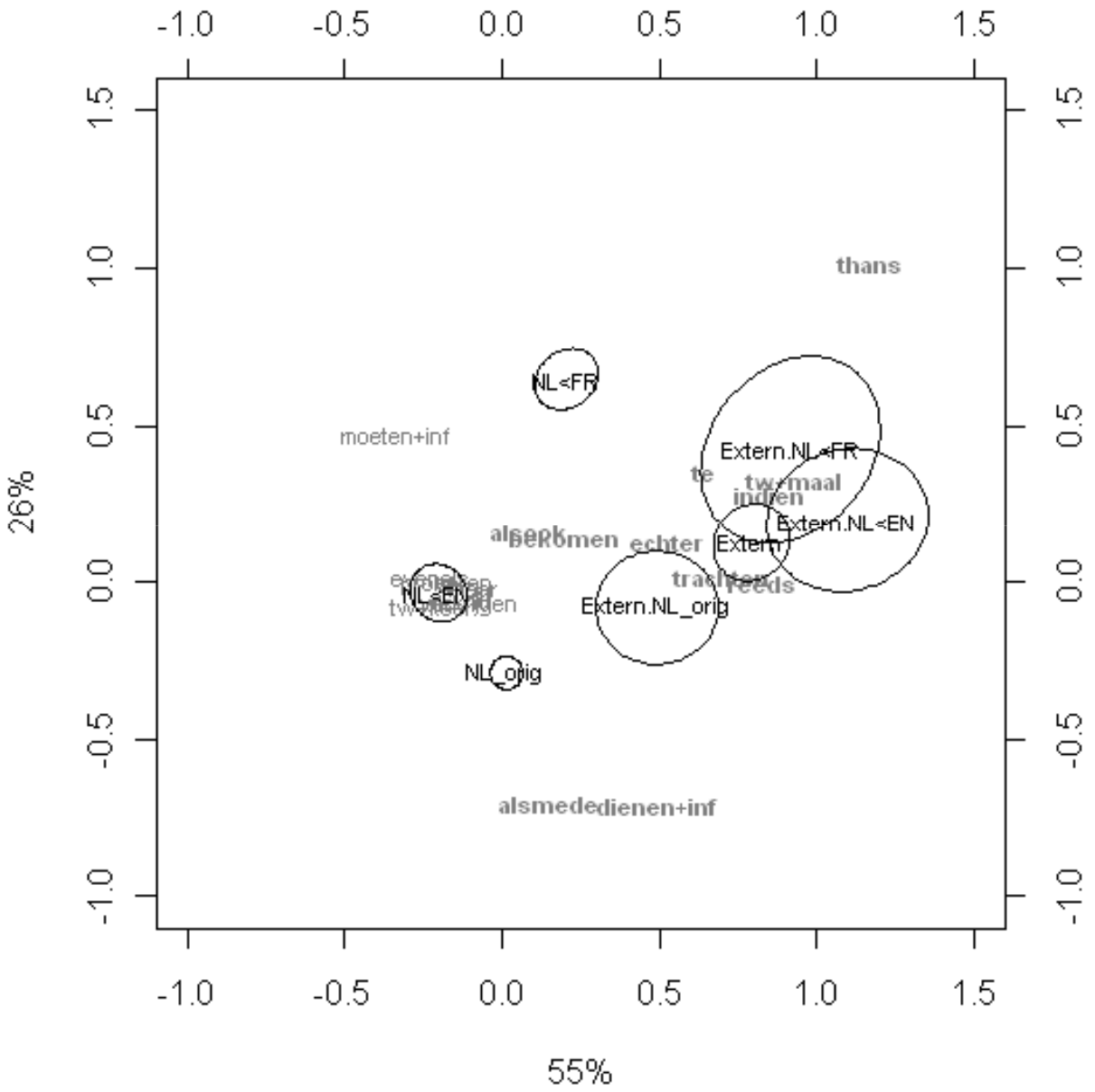
Adding interactions to the profile-based correspondence analysis


-1.0
-0.5
0.0
0.5
1.0
1.5
26%
55%
thans
NL<FR
moeten+inf
te
tw+maal
indien
al.NL<FR
als ook
berömen
echter
trachten
teeds
NL_orig
alsmededienen+inf

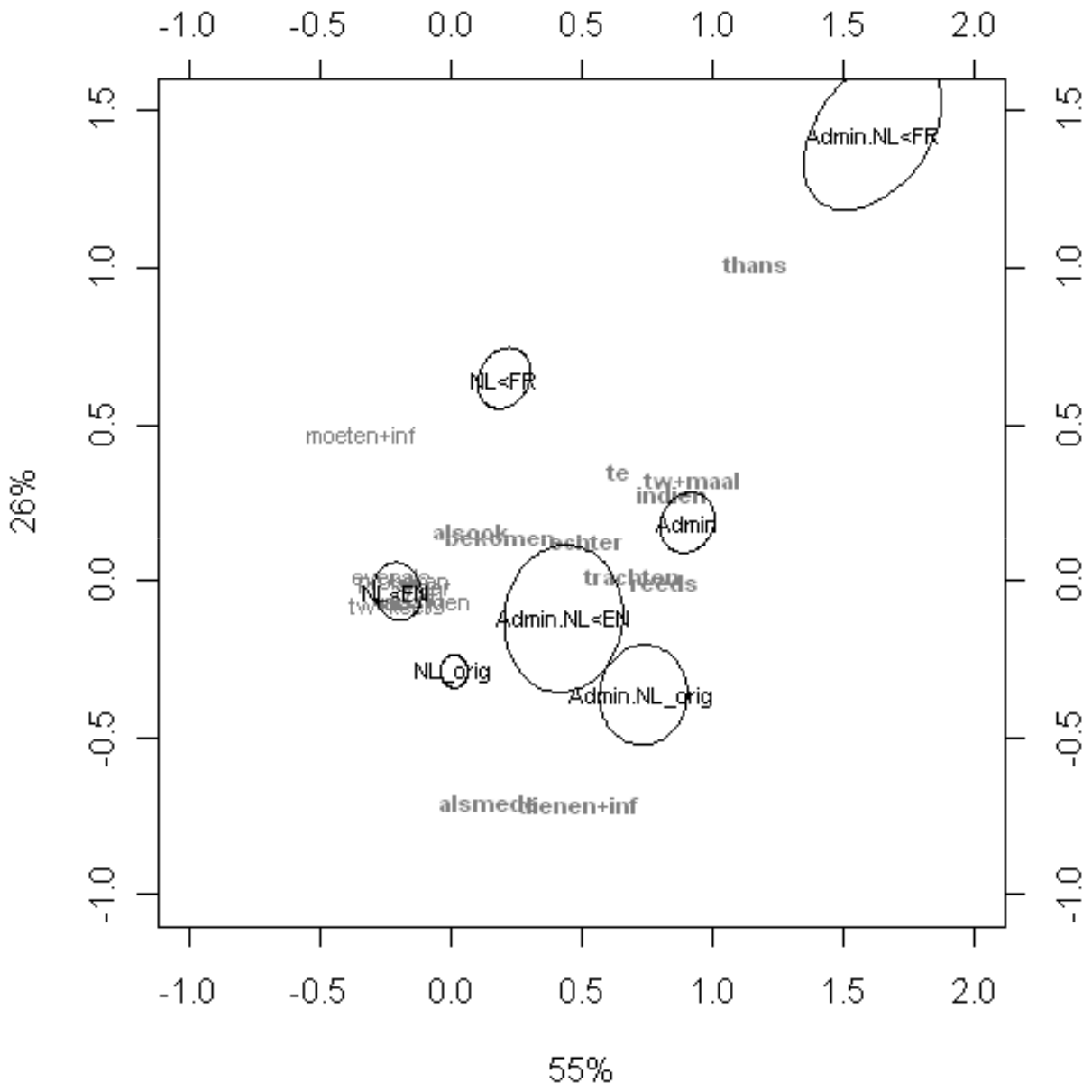
Journal


-1.0
-0.5
0.0
0.5
1.0
1.5
26%
55%
thans
NL<FR
moeten+inf
Non_Fic.NL<FR
te
tw+maal
indien
als
echter
Non_Fic
Non_Fic.NL-orig
NL<EN
NL_orig
alsmededienen+inf

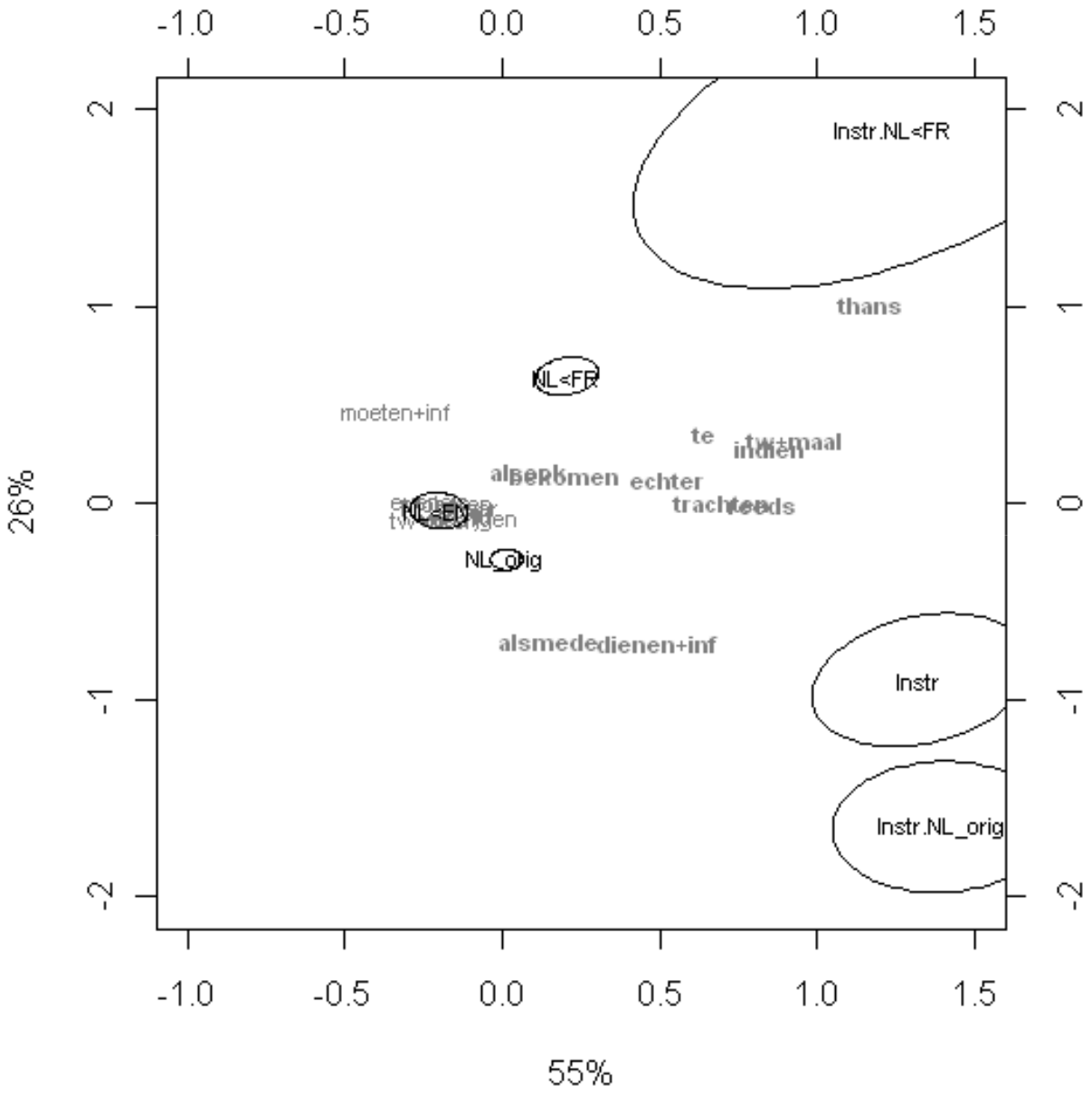
Non-fiction


-1.0
-0.5
0.0
0.5
1.0
1.5
26%
55%
thans
NL<FR
moeten+inf
Extern.NL<FR
te
twemaal
indien
Extern.NL<EN
als
echter
Extern
NL<EN
Extern.NL_orig
NL_orig
alsmededienen+inf

Extern


-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
26%
55%
Admin.NL<FR
thans
NL<FR
moeten+inf
te
.tw+maal
indien
Admin
alsook
trachten
reeds
NL<EN
Admin.NL<EN
NL_orig
Admin.NL_orig
alsmede
dienen+inf

Admin



Instruction

Summing up

- Observations:
  1. Not all text types show formality differences between translated and non-translated Dutch
  2. Conservatism hypothesis confirmed for external communication only
- Conclusion:
  - Features of translation, like conservatism, are not text type and source language dependent (as suggested in Baker 1993)
  - The observed variation in the dataset cannot be explained satisfactorily by translators' assumed conservative behaviour

Conservatism, part 2

Onomasiological research goal

- Measure the exact effect sizes of the two factors on the choice between neutral and formal lexemes
- Measure the explanatory and predictive power of such a model

Multivariate statistics, part 2

Binary logistic regression analysis

- Do all factors have a significant effect on the choice of word order?
- What is the relative impact of each factor?
- How do the different factors relate to each other?
- What is the collective effect of all factors?
- What is the predictive power of the model?

Results

Predictor	Odds ratio
Reference value: instr	
ADMIN	2.04 ***
EXTERN	2.56 ***
NON-FICTION	6.22 ***
JOURNAL	22.27 ***
Reference value: non-trans	
Translations from French	0.79 ***
Translations from English	n.s.

Model L.R. = 3056.27, df = 6, p < .0001

C = 0.75

Summing up

Impact hierarchy of the different factors

- What stimulates the use of formal (vs. neutral) lexemes the most, is not the difference between translations and non-translation, but the difference between text types

Conservatism, part 3

Onomasiological research goal (extd.)

- Measure the exact effect sizes of the two factors on the choice between neutral and formal lexemes, while checking a potential influence of individual lexemes (random variation)
- Measure the explanatory and predictive power of such a model

Multivariate statistics, part 3

Generalized mixed-effect models

– Answers the same questions as logistic regression, but it explicitly takes into account the random variation by the individual lexemes

Results

Predictor	Odds ratio
Reference value: instr	
ADMIN	1.63 ***
EXTERN	1.73***
NON-FICTION	4.01 ***
JOURNAL	15.33 ***
Reference value: non-trans	
Translations from French	0.86 ***
Translations from English	n.s.

C = 0.79

Summing up

- Relative sizes of the factors remain stable, but the mixed model shows that the random effect does play a role, too
  - Random effects are useless when one wants to inform theories
  - Introduce new fixed effects to decrease the effect size of the random effect

Conclusion

- Many techniques and tools in corpus linguistics are ready to be used in CBTS
  - Correspondence analysis, logistic regression, mixed-effect
  - Cf. also Diwersy et al.: PCA
- Be careful when using normalised frequencies, and consider using a profile-based approach (cf. Bernardini & Ferraresi 2011, Chesterman 2010)
  - Drawbacks of profile-based method: a lot of manual annotation (checking function stability), decision on what is synonymous is not always easy

Conclusion

- Multivariate statistics should be used only as a means to answer translation-relevant research questions (it is not a goal in itself)
- Multivariate statistics should be used with caution
  - Check model diagnostics (e.g. multicollinearity)

More information?

gert.desutter@hogent.be
http://webs.hogent.be/gertdesutter