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

Gert De Sutter

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

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 acknowledged 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

Saarbrücken, 09.02.2013
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

<table>
<thead>
<tr>
<th>Category</th>
<th>Non-translated Dutch</th>
<th>Translated Dutch (&lt; EN)</th>
<th>Translated Dutch (&lt; FR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADMIN</td>
<td>428,391</td>
<td>237,579</td>
<td>339,826</td>
</tr>
<tr>
<td>JOURNAL</td>
<td>483,714</td>
<td>295,039</td>
<td>272,429</td>
</tr>
<tr>
<td>INSTR</td>
<td>106,640</td>
<td>0</td>
<td>45,371</td>
</tr>
<tr>
<td>EXTERNAL</td>
<td>371,154</td>
<td>311,493</td>
<td>261,640</td>
</tr>
<tr>
<td>NON-FIC</td>
<td>412,712</td>
<td>0</td>
<td>96,688</td>
</tr>
</tbody>
</table>

Total: n = 3,662,676
Data, method, analysis

How to measure formal*ity* 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

<table>
<thead>
<tr>
<th></th>
<th>Admin</th>
<th>Extern</th>
<th>Instr</th>
<th>Journal</th>
<th>Non_Fic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abs. freq.</td>
<td>79</td>
<td>37</td>
<td>2</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>Norm freq.</td>
<td>0.78</td>
<td>0.52</td>
<td>0.35</td>
<td>0.07</td>
<td>0.21</td>
</tr>
</tbody>
</table>

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'

Saarbrücken, 09.02.2013
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.

<table>
<thead>
<tr>
<th></th>
<th>Admin</th>
<th>Extern</th>
<th>Instr</th>
<th>Journal</th>
<th>Non_Fic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tewerkstelling</td>
<td>26,9</td>
<td>33,9</td>
<td>66,7</td>
<td>11,1</td>
<td>54,2</td>
</tr>
<tr>
<td>Werkgelegenheid</td>
<td>73,1</td>
<td>66,1</td>
<td>33,3</td>
<td>88,9</td>
<td>45,8</td>
</tr>
</tbody>
</table>
Data, method, analysis

- **Extracting 10 profiles**
  - Profile = set of synonymous onomasiological variants to express a certain concept (i.e. 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

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

**Total:** n = 27.166
## Data, method, analysis

<table>
<thead>
<tr>
<th></th>
<th>Admin</th>
<th>Extern</th>
<th>Instr</th>
<th>Journal</th>
<th>Non-fic</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Al</em></td>
<td>296</td>
<td>254</td>
<td>67</td>
<td>731</td>
<td>634</td>
</tr>
<tr>
<td><em>Reeds</em></td>
<td>161</td>
<td>118</td>
<td>33</td>
<td>37</td>
<td>87</td>
</tr>
<tr>
<td><em>Dienen+inf</em></td>
<td>251</td>
<td>64</td>
<td>180</td>
<td>18</td>
<td>32</td>
</tr>
<tr>
<td><em>Moeten+inf</em></td>
<td>193</td>
<td>47</td>
<td>36</td>
<td>137</td>
<td>98</td>
</tr>
</tbody>
</table>

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
  - Exploratory 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
Multivariate statistics, part 1

Adding interactions to the profile-based correspondence analysis
Journal
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

Saarbrücken, 09.02.2013
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

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Odds ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reference value: instr</strong></td>
<td></td>
</tr>
<tr>
<td>ADMIN</td>
<td>2.04 ***</td>
</tr>
<tr>
<td>EXTERN</td>
<td>2.56 ***</td>
</tr>
<tr>
<td>NON-FICTION</td>
<td>6.22 ***</td>
</tr>
<tr>
<td>JOURNAL</td>
<td>22.27 ***</td>
</tr>
<tr>
<td><strong>Reference value: non-trans</strong></td>
<td></td>
</tr>
<tr>
<td>Translations from French</td>
<td>0.79 ***</td>
</tr>
<tr>
<td>Translations from English</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

Model L.R. = 3056.27, df = 6, p < .0001  
C = 0.75  

Saarbrücken, 09.02.2013
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

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Odds ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reference value: instr</strong></td>
<td></td>
</tr>
<tr>
<td>ADMIN</td>
<td>1.63***</td>
</tr>
<tr>
<td>EXTERN</td>
<td>1.73***</td>
</tr>
<tr>
<td>NON-FICTION</td>
<td>4.01***</td>
</tr>
<tr>
<td>JOURNAL</td>
<td>15.33***</td>
</tr>
<tr>
<td><strong>Reference value: non-trans</strong></td>
<td></td>
</tr>
<tr>
<td>Translations from French</td>
<td>0.86***</td>
</tr>
<tr>
<td>Translations from English</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

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