Register-specific interference in translation

Stella Neumann  RWTH Aachen University
Stefan Evert  FAU Erlangen-Nürnberg
Gert De Sutter  Ghent University
Register defined

- Functional language variation according to use, determined by the context of situation (Halliday 1978)
- Central role in language
  - Language system: virtual collection of all possible linguistic features, no existence outside of linguistic theorising
  - Register: actually available linguistic features in a given situational context
  - Register determines the distribution of linguistic features and their concrete specification → Probabilistic perspective on features
- No language use outside of register: Halliday & Hasan (1989, 40): we are never selecting with complete freedom from all resources of our linguistic system

Translators actively collect register-specific corpora to adapt their translation to the informally perceived features of the register

(Matthiessen 1993, Halliday 1991)
The multi-dimensional character of registers

Texts characterised by features on all levels of linguistic description → all of them to be included in register analysis

→ Biber’s bottom-up approach: $n$ dimensions as the result of multivariate analysis of corpora
→ SFL-inspired top-down approach: Set of latent parameters that can be operationalised in terms of linguistic indicators

→ Both require a holistic approach

Translationese

- **Translation properties**
  - Effects such as normalisation/sanitisation, simplification, explicitation etc. that make translations distinct from non-translated texts (e.g. Baker 1996)

- **Disputed in translation studies, but**
  - Effect not uniform across language pairs (i.e. universal), but probabilistic (Toury 2004), playing out specifically per language pair
  - Multifactorial: Contradictory effect of different features
  - Register-specific (Neumann 2013, Delaere 2015)

- **Machine learning discriminates with high accuracy**
  - Text classification: orig. vs. trans. (e.g. Volansky et al. 2015)
  - Indirect evidence for translationese, i.e. language use specific to translations
Interference / shining through

- In SLA research: transfer of features from L1 into L2
- Also observable in translations
  - **but**: translations by default from the L2 into the L1, i.e. opposite direction from using a foreign language
  - Why should the foreign language interfere with the native language?

- Possible explanations
  - Writing under the influence: Individual features of the ST so salient that translator uses them in the TT
  - Parallel activation of both language systems: since translation takes place between two languages, both languages must be activated and more than just the ST triggers might be at work ("genuine shining through")
Lines of research our approach draws on

- Traditional approach (e.g. Neumann to 2013)
  - comparison of translated and original texts (or texts and their translations) with respect to individual features
  - theoretical interpretation in terms of linguistic functions

- Machine learning: identification of translationese
  - e.g. Baroni & Bernardini (2006), Koppel & Ordan (2011)
  - usually based on low-level features (words, POS, n-grams)

- Multivariate statistical analysis of variation (Biber 1988, …)
  - latent (register) dimensions = groups of correlated features
  - we use principal component analysis (PCA) instead of Biber's factor analysis (for mathematical reasons)
  - But cannot show fine distinctions between originals and translations
Our approach
(Diwersy, Evert & Neumann 2014, Evert & Neumann 2017)

- Theory-driven choice of features
- Distance between feature vectors = (dis)similarity of texts
  \( \rightarrow \) wrt. theoretical framework / research question
- Exploratory multivariate analysis
  \( \rightarrow \) identifies latent dimensions with PCA
  \( \rightarrow \) orthogonal projection interpreted as “perspective” on data set
- Visualisation
  \( \rightarrow \) view shape of data set from different perspectives
- Minimally supervised intervention
  \( \rightarrow \) introduce theory-neutral information (DE/EN and transl./orig.)
  \( \rightarrow \) linear discriminant analysis (LDA) identifies best ‘perspective’
  for discrimination of these categories
- Interpretation
  \( \rightarrow \) characteristic features of latent dimensions can be interpreted in
  terms of theoretical background (here: underlying functions)
Previous work

- 150 EN-DE and DE-EN translation pairs from the CroCo Corpus (Hansen-Schirra et al. 2012): focus on 5 relatively similar registers ESSAY, POPSCI, SHARE, SPEECH, WEB

- 27 lexico-grammatical indicators of underlying functions proposed by Neumann (2013) in the context of register theory
  - Only comparable indicators
  - E.g. nouns/tokens, finites/sentences, passives/verbs, imperatives/sentences, adverbial themes/themes, contracted forms/tokens, lexical density, tokens/sentences

- Each text represented as a feature vector in multi-dimensional space characterized by the 27 indicators (as z-scores)

- “Shining through” perspective: LDA for EN vs. DE originals
  - Complemented by PCA dimensions capturing register variation
Multivariate analysis: EN / DE discriminant
Multivariate analysis: Translations vs. originals
Distribution along the language discriminant

significant*** shining through

d = 1.5

d = -1.1

![Graph showing distributions along the language discriminant with significant*** increases at d = 1.5 and d = -1.1.](graph.png)
Evidence of direction-specific shining through

- Separation of EN vs. DE original texts, with translations filling in space between the originals
  → Clear evidence of shining-through based on comparable register features

- Directionality effect contradicts simple parallel activation assumption
  → If shining through was due to parallel activation of both language systems, the effect should be the same in both directions and independent of register

- Potential explanation: diverging prestige of the languages involved (Toury 2012)
  → Prestige could be modulated by register
Hypothesis

- Translators are aware of and react to register-specific translation requirements
  - This includes the specific amount of shining through required/ permissible in the given register
- CroCo data suggests register effect, but too small for more detailed analysis
  - Moreover, the robustness of the shining through effect would be corroborated if found for other language pairs with similar constellations in terms of prestige
The Dutch Parallel Corpus

- Bi-directional corpus in the language pairs English-Dutch and French-Dutch (Macken et al. 2011)
  - 5017 texts, 10+ million word tokens
  - Six domains
  - Existing alignment not used for this study

- Language-specific PoS tagging
  - English sub-corpus re-tagged with CLAWS tagger (Garside 1987) for more specific coverage, higher accuracy

- Currently focus on EN-NL and NL-EN translations
Preprocessing

- Re-classification of text types drawing on Biber and Conrad’s (2009) situational characteristics (Delaere 2015, 61)
  - New text types: Broad Commercial Texts, Instructive Texts, Journalistic Texts, Legal Texts, Political Speeches, Specialized Communication, Tourist Information, Fiction

- Data: 504 pairs of original and matching translated texts
  - Text types with \(\leq 10\) texts removed (Legal, Tourist, Fiction)
  - At most 75 pairs per text type & translation direction
  - Short texts with < 500 tokens removed (CroCo: texts of 500 – 5,000 tokens)

- 37 lexico-grammatical indicators as log-transformed z-scores
  - Logarithmic transformation reduces skew & outliers, especially if short texts are not excluded
Multivariate analysis: EN / NL discriminant
Multivariate analysis: Translations vs. originals

EN

NL
Distribution along the language discriminant

shining through $d = 0.7^{***}$

n.s.

Distribution of texts along the language discriminant:
- NL ori
- NL trans
- EN ori
- EN trans

Legend:
- NL ori
- NL trans
- EN ori
- EN trans

Y-axis: Distribution of texts
X-axis: Language discriminant

Significance levels:
- $d = 0.7^{***}$
- n.s.
Discriminant distribution across text types

- **Broad**
  - EN ori
  - EN trans
  - NL ori
  - NL trans

- **Special**
  - EN ori
  - EN trans
  - NL ori
  - NL trans

- **Journalistic**
  - EN ori
  - EN trans
  - NL ori
  - NL trans

- **Political**
  - EN ori
  - EN trans
  - NL ori
  - NL trans
Discriminant distribution across text types

![Graph showing distribution of texts across language discriminants]

- EN ori
- EN trans
- NL ori
- NL trans

- Significance levels: *** (significant) and n.s. (not significant)
Discriminant distribution across text types

![Distribution graph showing differences in discriminant values for different text types.](image_url)
Discriminant distribution across text types

Journalistic

- EN ori
- EN trans
- NL ori
- NL trans

language discriminant

distribution of texts

*** ***
Discriminant distribution across text types

Political

n.s. n.s.

distribution of texts

language discriminant

EN ori
EN trans
NL ori
NL trans
EN / NL discriminant: Feature weights

The graph represents the normalized feature weights for English (EN) and Dutch (NL) languages. Each feature is plotted along the x-axis, with positive weights indicated by green bars and negative weights by red bars. The y-axis shows the normalized feature weights ranging from -0.25 to 0.25. The diagram is color-coded to indicate the weight values, with darker shades representing higher weights.

Features include:
- token/S
- nm/T
- np/T
- nominal/T
- prep/T
- pronouns/T
- p1pronoun/T
- p2pronoun/T
- p3pronoun/T
- verb/T
- willzellen/F
- infinitive/F
- finite/S
- passiveN/S
- imperative/S
- timeadey/T
- lexical density/S
- discourse/S
- predady/T
EN / NL discriminant: Contributions

language discriminant

[Diagram showing box plots for various language features and contributions, with labels for different groups such as 'NL ori', 'NL trans', 'EN trans', and 'EN ori'.]
Discussion

- Register distribution clearly visible
- Smaller shining through effect than for EN-DE and only for Dutch
  - Question of the language pair or of the corpus design/compilation?
  - Plausible: (over) normalisation in English
- Register-specific shining through
  - Broad commercial and Specialised texts blur the distinction of languages AND display strongest shining through effect
  - Unusual effect in Journalistic texts possibly due to incomparability of registers → methodological issue
  - Target language orientation of Political speeches
Summary & outlook

- Analysis rests on the comparability of features
  - Flaws in the choice of features and their counting will create artefacts
- Corpus design matters
- Include French data from the DPC
  - More complex relationship between languages: interaction between French and Dutch
- Methodological considerations
  - How stable are the LDA and PCA dimensions?
  - Idea: systematic bootstrapping of texts and features to determine which patterns are “real” in the plots
Thank you for your attention!

Stella Neumann  
stella.neumann@ifaar.rwth-aachen.de

Stefan Evert  
stefan.evert@fau.de

Gert De Sutter  
gert.desutter@ugent.be
References


References 2


