Linguistic variation across different groups of translated and non-translated texts

Combined effect and individual contributions of lexico-grammatical features

Tatiana Serbina
RWTH Aachen University
Mario Bisiada
Universitat Pompeu Fabra
Stella Neumann
RWTH Aachen University

Translation in Transition 6, 22.-23.09.2022

Aim of the study

- Interaction of the four explanatory factors
  - language
  - register
  - translation status
  - editorial intervention
- Research questions:
  - Effect of these factors on linguistic profiles of the analyzed texts
  - Individual contributions of the lexico-grammatical features
- Method: Geometric Multivariate Analysis, GMA (Diwersy et al. 2014; Evert & Neumann 2017)

Explanatory factors

Language

- the lexico-grammatical features of texts are determined by the corresponding language

Explanatory factors

Language

- the lexico-grammatical features of texts are determined by the corresponding language
- numerous contrastive studies, such as König & Gast (2018); Hansen-Schirra et al. (2012); Boas (2010) concentrating on the language pair English-German

Explanatory factors

Register

- linguistic features vary in their frequency depending on registers, i.e. situational contexts

Explanatory factors

Register

- linguistic features vary in their frequency depending on registers, i.e. situational contexts
- some empirical findings on register variation in translation (for a review see Neumann (2021)):

Explanatory factors

Register

- linguistic features vary in their frequency depending on registers, i.e. situational contexts
- some empirical findings on register variation in translation (for a review see Neumann (2021)):
  - Neumann (2013) has shown that for culturally comparable registers there are only minor differences in the feature distribution found in translations and the comparable originals

Explanatory factors

Register

- linguistic features vary in their frequency depending on registers, i.e. situational contexts
- some empirical findings on register variation in translation (for a review see Neumann (2021)):
  - Neumann (2013) has shown that for culturally comparable registers there are only minor differences in the feature distribution found in translations and the comparable originals
  - Delaere (2015) suggests that register plays an important role in norm conformity in translated language

Explanatory factors

Translation status

- the uniqueness of translation as a linguistic activity should not be overstated →

Explanatory factors

Translation status

- the uniqueness of translation as a linguistic activity should not be overstated →
  - the same cognitive mechanisms as in language production in general are at play (Halverson & Kotze 2022)

Explanatory factors

Translation status

- the uniqueness of translation as a linguistic activity should not be overstated →
    - the same cognitive mechanisms as in language production in general are at play (Halverson & Kotze 2022)
    - translation as a type of constrained communication (Kotze 2022)

Explanatory factors

Translation status

- the uniqueness of translation as a linguistic activity should not be overstated →
  - the same cognitive mechanisms as in language production in general are at play (Halverson & Kotze 2022)
  - translation as a type of constrained communication (Kotze 2022)
- linguistic profiles depend on language- and register-specific entrenchment but some distributions specific to translations → research on translation properties (overview in de Sutter & Lefer 2020) and translationese (Volansky et al. 2015)

Explanatory factors

Translation status

- the uniqueness of translation as a linguistic activity should not be overstated →
  - the same cognitive mechanisms as in language production in general are at play (Halverson & Kotze 2022)
  - translation as a type of constrained communication (Kotze 2022)
- linguistic profiles depend on language- and register-specific entrenchment but some distributions specific to translations → research on translation properties (overview in de Sutter & Lefer 2020) and translationese (Volansky et al. 2015)
- linguistic behavior shaped by the 'practice of translating texts, of particular kinds, for particular purposes and for particular clients' (Halverson & Kotze 2022: 72)

Explanatory factors

Editorial intervention

- published texts are often edited

Explanatory factors

Editorial intervention

- published texts are often edited
- editorial intervention could be seen as normal and 'not something that the linguist should wish away as noise or change imposed on "authentic data"' (Kruger & van Rooy 2018: 220)

Explanatory factors

Editorial intervention

- published texts are often edited
- editorial intervention could be seen as normal and 'not something that the linguist should wish away as noise or change imposed on "authentic data"' (Kruger & van Rooy 2018: 220)
- → linguistic features present in the final published version contribute to entrenchment of these features (Kruger & van Rooy 2018: 220)

Explanatory factors

Editorial intervention

- published texts are often edited
- editorial intervention could be seen as normal and ’not something that the linguist should wish away as noise or change imposed on “authentic data”’ (Kruger & van Rooy 2018: 220)
- → linguistic features present in the final published version contribute to entrenchment of these features (Kruger & van Rooy 2018: 220)
- still essential to acknowledge and assess the additional workflow stages, particularly for a successful integration of product and process research (Serbina & Neumann 2022: 142)

Interaction between editorial intervention & transl. status

Kruger (2012): ‘mediated discourse’ in Afrikaans–English

- normalisation, explicitation & simplification in ‘mediated’ (translated, edited) and ‘unmediated’ (unedited) text

Interaction between editorial intervention & transl. status

Kruger (2012): ‘mediated discourse’ in Afrikaans–English

- normalisation, explicitation & simplification in ‘mediated’ (translated, edited) and ‘unmediated’ (unedited) text
- no evidence of shared ‘mediation effect’
  - translators favour ‘explicit and standardised language’
  - editors ‘introduce collocational variety’

Interaction between editorial intervention & transl. status

Kruger (2012): 'mediated discourse' in Afrikaans–English

- normalisation, explicitation & simplification in 'mediated' (translated, edited) and 'unmediated' (unedited) text
- no evidence of shared 'mediation effect'
  - translators favour 'explicit and standardised language'
  - editors 'introduce collocational variety'

Bisiada (2017): replication of Kruger (2012) for English–German

- normalisation: difference between translated and non-translated text → non-translated texts use unconventional/creative language

Interaction between editorial intervention & transl. status

Kruger (2012): 'mediated discourse' in Afrikaans–English

- normalisation, explicitation & simplification in 'mediated' (translated, edited) and 'unmediated' (unedited) text
- no evidence of shared 'mediation effect'
  - translators favour 'explicit and standardised language'
  - editors 'introduce collocational variety'

Bisiada (2017): replication of Kruger (2012) for English–German

- normalisation: difference between translated and non-translated text → non-translated texts use unconventional/creative language
- simplification: difference between manuscripts and published texts → notable editorial influence

Editorial intervention: individual features and multivariate analysis

- some phenomena like sentence splitting are caused by both translators and editors (Bisiada 2016; 2018b)

Editorial intervention: individual features and multivariate analysis

- some phenomena like sentence splitting are caused by both translators and editors (Bisiada 2016; 2018b)
- translators and editors are guided by different goals

Editorial intervention: individual features and multivariate analysis

- some phenomena like sentence splitting are caused by both translators and editors (Bisiada 2016; 2018b)
- translators and editors are guided by different goals
- they both make extensive changes to nominalisations (Bisiada 2018a,c)

Editorial intervention: individual features and multivariate analysis

- some phenomena like sentence splitting are caused by both translators and editors (Bisiada 2016; 2018b)
- translators and editors are guided by different goals
- they both make extensive changes to nominalisations (Bisiada 2018a,c)
- editors eliminate passive constructions from translations, especially when the verb is in the past tense (Bisiada 2019)

Editorial intervention: individual features and multivariate analysis

- some phenomena like sentence splitting are caused by both translators and editors (Bisiada 2016; 2018b)
- translators and editors are guided by different goals
- they both make extensive changes to nominalisations (Bisiada 2018a,c)
- editors eliminate passive constructions from translations, especially when the verb is in the past tense (Bisiada 2019)
- previous multivariate analysis considering only German originals and translations did not indicate a profound effect of editorial intervention (Serbina et al. 2021) – calling for more extensive analysis across languages

Corpus-based methodologies

Corpus-based research

- The frequency approach: Comparison of frequencies of a feature across two or more data sets, e.g. different registers or varieties etc.
- The regression approach: Prediction of a single quantity from multiple explanatory factors, e.g. alternation studies (what drives the choice for one feature over another?)
- The multivariate approach: Exploration of complex relationships between multiple features and multiple factors, e.g. Biber's multi-dimensional analysis

Corpus-based methodologies

Multivariate studies

- Determine latent, i.e. hidden, factors characterised by a set of features with the help of correlation analysis

Corpus-based methodologies

Multivariate studies

- Determine latent, i.e. hidden, factors characterised by a set of features with the help of correlation analysis
    - Reduction of dimensionality of the data set

Corpus-based methodologies

Multivariate studies

- Determine latent, i.e. hidden, factors characterised by a set of features with the help of correlation analysis
  - Reduction of dimensionality of the data set
  - One dimension: a factor reflected in co-occurring features of data points

Corpus-based methodologies

Multivariate studies

- Determine latent, i.e. hidden, factors characterised by a set of features with the help of correlation analysis
  - Reduction of dimensionality of the data set
  - One dimension: a factor reflected in co-occurring features of data points
  - Interpreted as ‘causes’ of variation

Corpus-based methodologies

Multivariate studies

- Determine latent, i.e. hidden, factors characterised by a set of features with the help of correlation analysis
  - Reduction of dimensionality of the data set
  - One dimension: a factor reflected in co-occurring features of data points
  - Interpreted as ‘causes’ of variation
  - Datasets are often characterised by more than one dimension

Corpus-based methodologies

Multivariate studies

- Determine latent, i.e. hidden, factors characterised by a set of features with the help of correlation analysis
  - Reduction of dimensionality of the data set
  - One dimension: a factor reflected in co-occurring features of data points
  - Interpreted as ‘causes’ of variation
  - Datasets are often characterised by more than one dimension
- Outcome: clusters of features a group of data points share

Corpus-based methodologies

Multivariate studies

- Determine latent, i.e. hidden, factors characterised by a set of features with the help of correlation analysis
  - Reduction of dimensionality of the data set
  - One dimension: a factor reflected in co-occurring features of data points
  - Interpreted as 'causes' of variation
  - Datasets are often characterised by more than one dimension
- Outcome: clusters of features a group of data points share
- Text as the unit capturing the combined effect of factors

Corpus-based methodologies

Multivariate studies

- Determine latent, i.e. hidden, factors characterised by a set of features with the help of correlation analysis
  - Reduction of dimensionality of the data set
  - One dimension: a factor reflected in co-occurring features of data points
  - Interpreted as 'causes' of variation
  - Datasets are often characterised by more than one dimension
- Outcome: clusters of features a group of data points share
- Text as the unit capturing the combined effect of factors
  - Each text (= data point) characterised by a set of quantitative linguistic variables (a 'feature vector')

Data: Overview of the data sample

Corpora: Harvard Business Corpus, HBC (Bisiada 2018a) and CroCo Corpus (Hansen-Schirra et al. 2012)

Corpus	Translation Status	Register	Size in words	No. of texts
HBC	ST EN	Business	106,035	26
HBC	Manuscript T DE	Business	112,810	26
HBC	Published T DE	Business	106,958	26
CroCo	ST EN	Share, Popsci	62,952	24
CroCo	Published T DE	Share, Popsci	61,791	24
CroCo	ST DE	Share, Popsci, Speech, Essay	124,926	62

Methods

POS tagging

- German: TreeTagger (Schmid 1994) using STTS tagset (Schiller et al. 1999)
- English: CLAWS tagger (Garside & Smith 1997) using the CLAWS 7 tagset

Methods

POS tagging

- German: TreeTagger (Schmid 1994) using STTS tagset (Schiller et al. 1999)
- English: CLAWS tagger (Garside & Smith 1997) using the CLAWS 7 tagset

Feature extraction

- cqp script (Fest et al. 2019; Neumann & Evert 2021) based on the lexico-grammatical features developed by Neumann (2013)

Methods

POS tagging

- German: TreeTagger (Schmid 1994) using STTS tagset (Schiller et al. 1999)
- English: CLAWS tagger (Garside & Smith 1997) using the CLAWS 7 tagset

Feature extraction

- cqp script (Fest et al. 2019; Neumann & Evert 2021) based on the lexico-grammatical features developed by Neumann (2013)
- normalized to an appropriate unit of measurement (e.g. nominalizations/word or passive/finite verb) and standardized to mitigate different ranges of variation

Methods

POS tagging

- German: TreeTagger (Schmid 1994) using STTS tagset (Schiller et al. 1999)
- English: CLAWS tagger (Garside & Smith 1997) using the CLAWS 7 tagset

Feature extraction

- cqp script (Fest et al. 2019; Neumann & Evert 2021) based on the lexico-grammatical features developed by Neumann (2013)
- normalized to an appropriate unit of measurement (e.g. nominalizations/word or passive/finite verb) and standardized to mitigate different ranges of variation
- after inspecting for excessive correlations: 36 features

Methods

Geometric Multivariate Analysis

- procedure combining statistical analysis and visualization-based interpretation (Diwersy et al. 2014; Evert & Neumann 2017); due to size of the data set focus on PCA

Methods

Geometric Multivariate Analysis

- procedure combining statistical analysis and visualization-based interpretation (Diwersy et al. 2014; Evert & Neumann 2017); due to size of the data set focus on PCA
- data pre-processing and analysis – R scripts (based on scripts by Stephanie Evert, FAU Erlangen-Nürnberg)

Methods

Geometric Multivariate Analysis

- procedure combining statistical analysis and visualization-based interpretation (Diwersy et al. 2014; Evert & Neumann 2017); due to size of the data set focus on PCA
- data pre-processing and analysis – R scripts (based on scripts by Stephanie Evert, FAU Erlangen-Nürnberg)
- every text represented as a feature vector in multi-dimensional feature space

Methods

Geometric Multivariate Analysis

- procedure combining statistical analysis and visualization-based interpretation (Diwersy et al. 2014; Evert & Neumann 2017); due to size of the data set focus on PCA
- data pre-processing and analysis – R scripts (based on scripts by Stephanie Evert, FAU Erlangen-Nürnberg)
- every text represented as a feature vector in multi-dimensional feature space
- Euclidean distances between feature vectors assumed to represent meaningful differences between texts

Methods

Multi-dimensional space

- Variation influenced by multiple factors –> multiple dimensions

Methods

Multi-dimensional space

- Variation influenced by multiple factors –> multiple dimensions
- Distances between data points influenced in complex ways by these factors

Methods

Multi-dimensional space

- Variation influenced by multiple factors –> multiple dimensions
- Distances between data points influenced in complex ways by these factors
  - Position in space reflects this

Methods

Multi-dimensional space

- Variation influenced by multiple factors –> multiple dimensions
- Distances between data points influenced in complex ways by these factors
    - Position in space reflects this
- Our visual concept of dimensionality is geared towards three dimensions of perceptual space

Methods

Multi-dimensional space

- Variation influenced by multiple factors –> multiple dimensions
- Distances between data points influenced in complex ways by these factors
    - Position in space reflects this
- Our visual concept of dimensionality is geared towards three dimensions of perceptual space
    - In higher-dimensional space, there are simply more perspectives

Methods

Multi-dimensional space

- Variation influenced by multiple factors –> multiple dimensions
- Distances between data points influenced in complex ways by these factors
  - Position in space reflects this
- Our visual concept of dimensionality is geared towards three dimensions of perceptual space
  - In higher-dimensional space, there are simply more perspectives
  - Multi-dimensional space is not a cube, but a polygon

Methods

Multi-dimensional space

- Variation influenced by multiple factors –> multiple dimensions
- Distances between data points influenced in complex ways by these factors
    - Position in space reflects this
- Our visual concept of dimensionality is geared towards three dimensions of perceptual space
    - In higher-dimensional space, there are simply more perspectives
    - Multi-dimensional space is not a cube, but a polygon
- In scatter plot one feature vector, i.e. one text, represented by one symbol

Principal Component Analysis

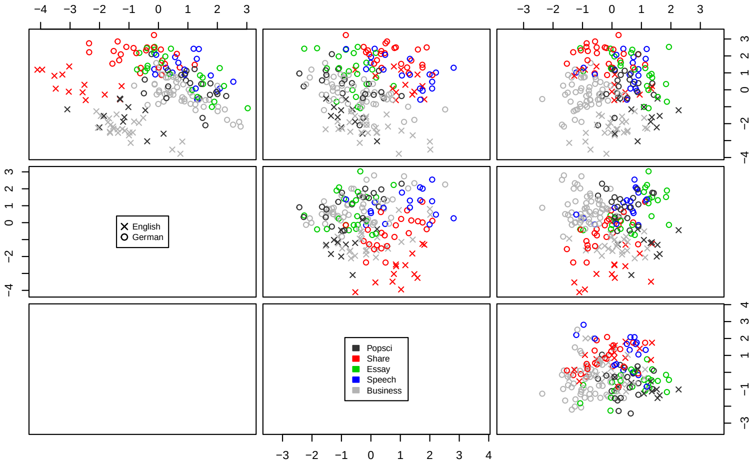

Figure: Scatterplot matrix of the first four PCA dimensions

Principal Component Analysis

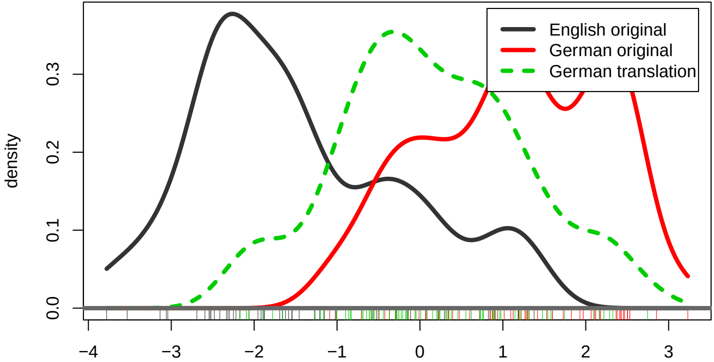

Figure: Discriminant plot of the 1st PCA dim according to language and translation status

Principal Component Analysis

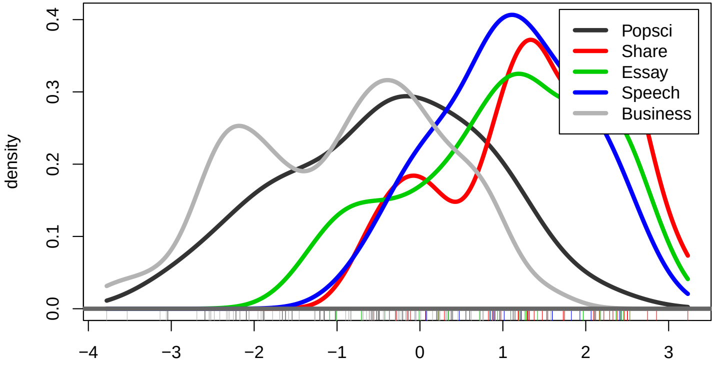

Figure: Discriminant plot of the 1st PCA dim according to register

Principal Component Analysis

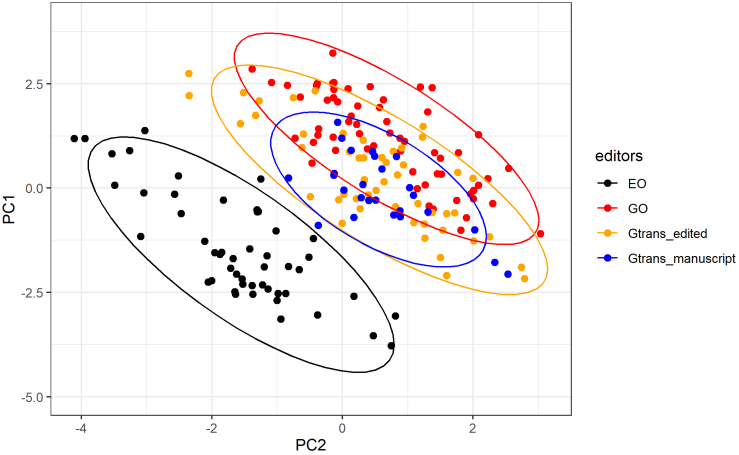

Figure: Scatterplot matrix of the first two PCA dimensions showing editorial intervention

Principal Component Analysis

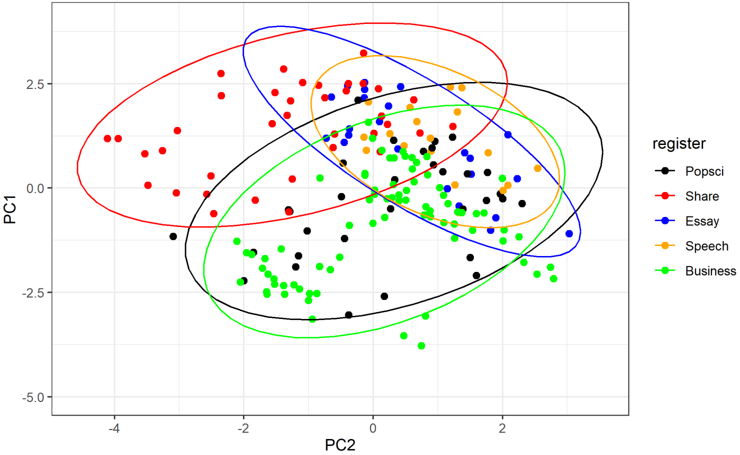

Figure: Scatterplot matrix of the first two PCA dimensions showing registers

PCA Dim 1: Feature weights

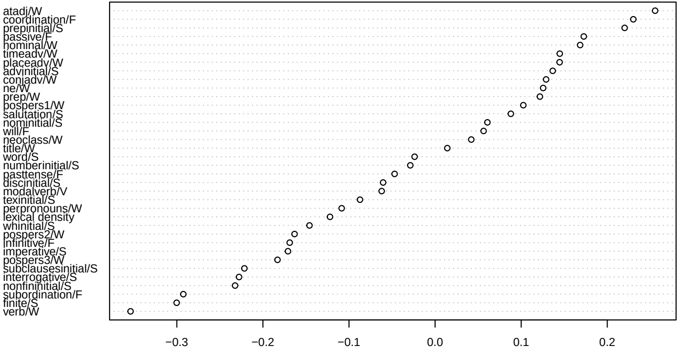

Figure: Dot chart of the 1st PCA dim

Distribution of attributive adjectives (AtAdj/W)

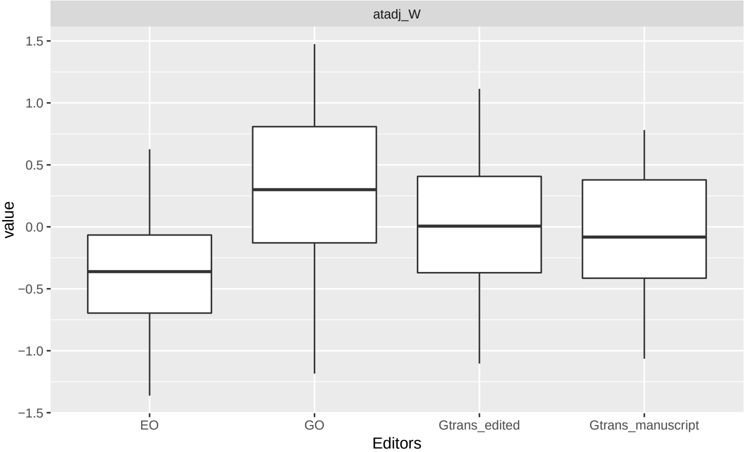

Figure: Boxplot for attributive adjectives

Distribution of verbs (Verb/W)

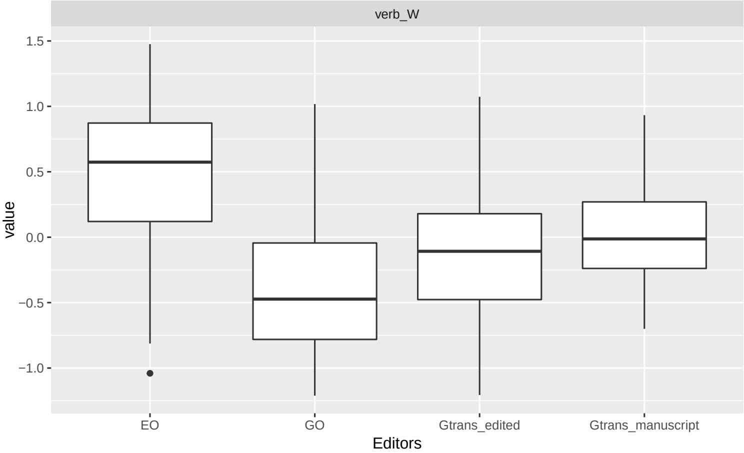

Figure: Boxplot for verbs

Distribution of attributive adj-s and verbs

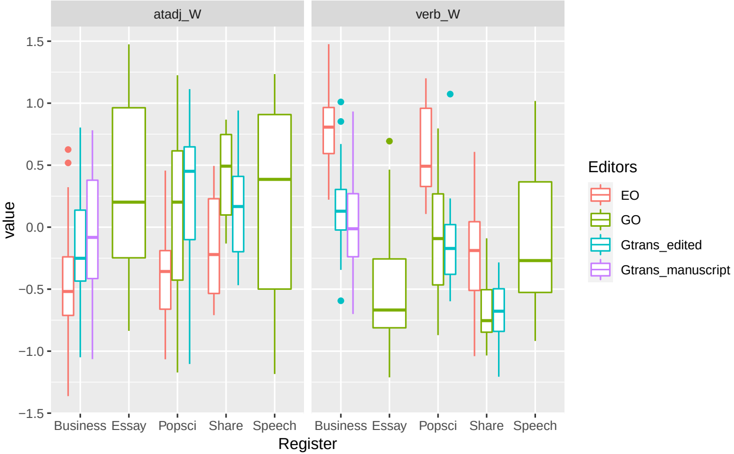

Figure: Boxplot for attributive adjectives and verbs showing editorial intervention and registers

Interaction of the four explanatory variables

- language and translation status seem to account for most of the variation (see also previous GMA studies on the role and interaction of these two factors (Diwersy et al. 2014; Evert & Neumann 2017))

Interaction of the four explanatory variables

- language and translation status seem to account for most of the variation (see also previous GMA studies on the role and interaction of these two factors (Diwersy et al. 2014; Evert & Neumann 2017))
- in contrast to Serbina et al. (2021), register effect is less profound

Interaction of the four explanatory variables

- language and translation status seem to account for most of the variation (see also previous GMA studies on the role and interaction of these two factors (Diwersy et al. 2014; Evert & Neumann 2017))
- in contrast to Serbina et al. (2021), register effect is less profound
- considering the whole range of linguistic features and other explanatory variables, editorial intervention does not appear to contribute much to linguistic variation

Possible reasons

- influence of the languages unsurprising

Possible reasons

- influence of the languages unsurprising
- unlike Bisiada's analyses, our data set contains a mixture of registers

Possible reasons

- influence of the languages unsurprising
- unlike Bisiada's analyses, our data set contains a mixture of registers
    - differences between registers are so strong that they even interact with language

Possible reasons

- influence of the languages unsurprising
- unlike Bisiada's analyses, our data set contains a mixture of registers
  - differences between registers are so strong that they even interact with language
- changes introduced by editors much subtler

Possible reasons

- influence of the languages unsurprising
- unlike Bisiada's analyses, our data set contains a mixture of registers
    - differences between registers are so strong that they even interact with language
- changes introduced by editors much subtler
- the multivariate exploration puts things into perspective

Possible reasons

- influence of the languages unsurprising
- unlike Bisiada's analyses, our data set contains a mixture of registers
    - differences between registers are so strong that they even interact with language
- changes introduced by editors much subtler
- the multivariate exploration puts things into perspective
    - frequency analysis controlling for factors and individual features brings out subtle phenomena

Possible reasons

- influence of the languages unsurprising
- unlike Bisiada's analyses, our data set contains a mixture of registers
    - differences between registers are so strong that they even interact with language
- changes introduced by editors much subtler
- the multivariate exploration puts things into perspective
    - frequency analysis controlling for factors and individual features brings out subtle phenomena
    - "zooming out" to a more complex analysis of a large set of features and different factors shifts focus to other factors

Conclusion and outlook

- in the current data set and combination of factors, the respective contribution of each factor to the overall linguistic profile appears to be the following:

Conclusion and outlook

- in the current data set and combination of factors, the respective contribution of each factor to the overall linguistic profile appears to be the following:
    - language – high

Conclusion and outlook

- in the current data set and combination of factors, the respective contribution of each factor to the overall linguistic profile appears to be the following:
  - language – high
  - register – medium

Conclusion and outlook

- in the current data set and combination of factors, the respective contribution of each factor to the overall linguistic profile appears to be the following:
    - language – high
    - register – medium
    - editorial intervention – low

Conclusion and outlook

- in the current data set and combination of factors, the respective contribution of each factor to the overall linguistic profile appears to be the following:
  - language – high
  - register – medium
  - editorial intervention – low
- however, as shown in the previous studies, each factor in its own right can have an effect on individual linguistic features

Conclusion and outlook

- in the current data set and combination of factors, the respective contribution of each factor to the overall linguistic profile appears to be the following:
  - language – high
  - register – medium
  - editorial intervention – low
- however, as shown in the previous studies, each factor in its own right can have an effect on individual linguistic features
- future research should

Conclusion and outlook

- in the current data set and combination of factors, the respective contribution of each factor to the overall linguistic profile appears to be the following:
  - language – high
  - register – medium
  - editorial intervention – low
- however, as shown in the previous studies, each factor in its own right can have an effect on individual linguistic features
- future research should
  - consider larger datasets with rich meta-data (similar to the MUST corpus and DPC 2.0)

Conclusion and outlook

- in the current data set and combination of factors, the respective contribution of each factor to the overall linguistic profile appears to be the following:
  - language – high
  - register – medium
  - editorial intervention – low
- however, as shown in the previous studies, each factor in its own right can have an effect on individual linguistic features
- future research should
  - consider larger datasets with rich meta-data (similar to the MUST corpus and DPC 2.0)
  - include further explanatory factors

Conclusion and outlook

- in the current data set and combination of factors, the respective contribution of each factor to the overall linguistic profile appears to be the following:
    - language – high
    - register – medium
    - editorial intervention – low
- however, as shown in the previous studies, each factor in its own right can have an effect on individual linguistic features
- future research should
    - consider larger datasets with rich meta-data (similar to the MUST corpus and DPC 2.0)
    - include further explanatory factors
    - study a greater variety of languages

Thank you for your attention!

References I

Bisiada, Mario. 2016. "Lösen Sie Schachtelsätze möglichst auf": The impact of editorial guidelines on sentence splitting in German business article translations. *Applied Linguistics* 37(3). 354–376. https://doi.org/10.1093/applin/amu035.

Bisiada, Mario. 2017. Universals of editing and translation. In Silvia Hansen-Schirra, Oliver Czulo, Sascha Hofmann & Bernd Meyer (eds.), *Empirical modelling of translation and interpreting*, 241–275. Berlin: Language Science Press.

Bisiada, Mario. 2018a. Editing nominalisations in English–German translation: When do editors intervene? *The Translator* 24(1). 35–49. https://doi.org/10.1080/13556509.2017.1301847.

Bisiada, Mario. 2018b. The editor's invisibility: Analysing editorial intervention in translation. *Target* 30(2). 288–309. https://doi.org/10.1075/target.16116.bis.

Bisiada, Mario. 2018c. Translation and editing: A study of editorial treatment of nominalisations in draft translations. *Perspectives: Studies in Translation Theory and Practice* 26(1). 24–38. https://doi.org/10.1080/0907676X.2017.1290121.

Bisiada, Mario. 2019. Translated language or edited language? A study of passive constructions in translation manuscripts and their published versions. *Across Languages and Cultures* 20(1). 35–56. https://doi.org/10.1556/084.2019.20.1.2.

Boas, Hans C. (ed.). 2010. *Contrastive studies in construction grammar*. Amsterdam: Benjamins.

References II

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

Diwersy, Sascha, Stefan Evert & Stella Neumann. 2014. A weakly supervised multivariate approach to the study of language variation. In Benedikt Szmrecsanyi & Bernhard Wälchli (eds.), *Aggregating dialectology, typology, and register analysis: linguistic variation in text and speech*, 174–204. Berlin: de Gruyter.

Evert, Stefan & Stella Neumann. 2017. The impact of translation direction on characteristics of translated texts: a multivariate analysis for English and German. In Gert de Sutter, Marie-Aude Lefer & Isabelle Delaere (eds.), *Empirical translation studies: new theoretical and methodological traditions*, 47–80. Berlin: Mouton de Gruyter.

Fest, Jennifer, Arndt Heilmann, Oliver Hohlfeld, Stella Neumann, Helge Reelfs, Marco Schmitt & Alina Vogelgesang. 2019. Determining response-generating contexts on microblogging platforms. In *Proceedings of the 15th conference on natural language processing (konvens)*, 171–182.

Garside, Roger & Nicholas Smith. 1997. A hybrid grammatical tagger: claws4. In Roger Garside, Geoffrey Leech & Anthony McEnery (eds.), *Corpus annotation: linguistic information from computer text corpora*, 102–121. London: Longman.

Halverson, Sandra L. & Haidee Kotze. 2022. Sociocognitive constructs in translation and interpreting studies (tis): do we really need concepts like norms and risk when we have a comprehensive usage-based theory of language? In Sandra L. Halverson & Álvaro Marín García (eds.), *Contesting epistemologies in cognitive translation and interpreting studies*, 51–79. New York: Routledge.

References III

Hansen-Schirra, Silvia, Stella Neumann & Erich Steiner. 2012. *Cross-linguistic corpora for the study of translations: insights from the language pair english-german*. Berlin: de Gruyter.

König, Ekkehard & Volker Gast. 2018. *Understanding English-German contrasts*. 4th edn. Berlin: Erich Schmidt Verlag.

Kotze, Haidee. 2022. Translation as constrained communication: principles, concepts and methods. In Sylviane Granger & Marie-Aude Lefer (eds.), *Extending the scope of corpus-based translation studies*, 67–97. London: Bloomsbury.

Kruger, Haidee. 2012. A corpus-based study of the mediation effect in translated and edited language. *Target* 24(2). 355–388. https://doi.org/10.1075/target.24.2.07kru.

Kruger, Haidee & Bertus van Rooy. 2018. Register variation in written contact varieties: a multidimensional analysis. *English World-Wide: A Journal of Varieties of English* 39(2). 214–242.

Neumann, Stella. 2013. *Contrastive register variation: a quantitative approach to the comparison of english and german*. Berlin: de Gruyter.

Neumann, Stella. 2021. Register and translation. In Mira Kim, Jeremy Munday, Zhenhua Wang & Pin Wang (eds.), *Systemic functional linguistics and translation studies*, 65–82. London: Bloomsbury.

Neumann, Stella & Stefan Evert. 2021. A register variation perspective on varieties of english. In Elena Seoane & Douglas Biber (eds.), *Corpus based approaches to register variation*, 143–178. Berlin: de Gruyter.

References IV

Schiller, Anne, Simone Teufel, Christine Stöckert & Christine Thielen. 1999. *Guidelines für das tagging deutscher textcorpora mit stts*. Universität Stuttgart: Universität Stuttgart. http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*.

Serbina, Tatiana, Mario Bisiada & Stella Neumann. 2021. Linguistic profiles of translation manuscripts and edited translations. In *Proceedings for the first workshop on modelling translation: translatology in the digital age*, 34–45. Association for Computational Linguistics. https://aclanthology.org/2021.motra-1.4.

Serbina, Tatiana & Stella Neumann. 2022. Translation product and process data: a happy marriage or worlds apart? In Sandra L. Halverson & Álvaro Marín García (eds.), *Contesting epistemologies in cognitive translation and interpreting studies*, 131–152. New York: Routledge.

de Sutter, Gert & Marie-Aude Lefer. 2020. On the need for a new research agenda for corpus-based translation studies: a melti-methodological, multifactorial and interdisciplinary approach. *Perspectives* 28(1). 1–23.

Volansky, Vered, Noam Ordan & Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities* 30(1). 98–118.

Distribution of coordinated clauses

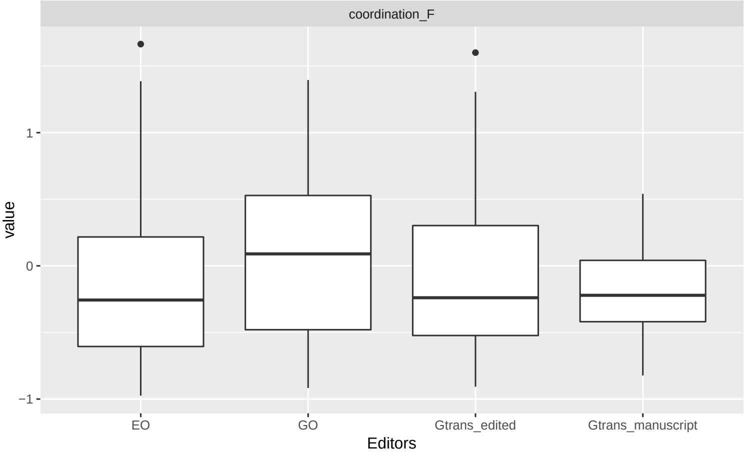

Figure: Boxplot for coordinated clauses

Distribution of finite clauses

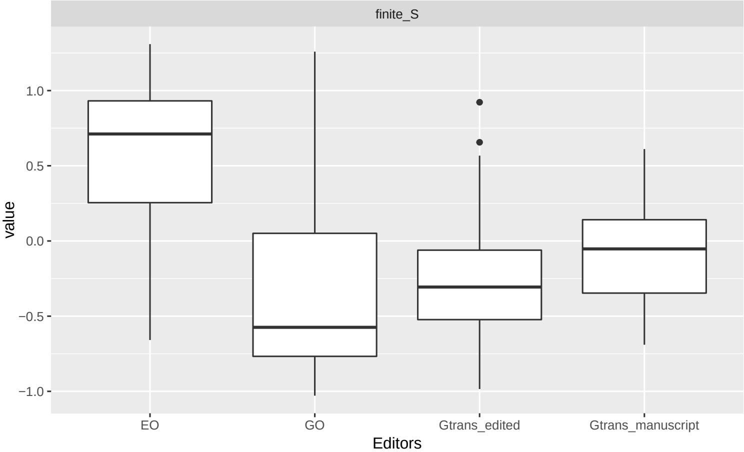

Figure: Boxplot for finite clauses

Distribution of PPs as Themes

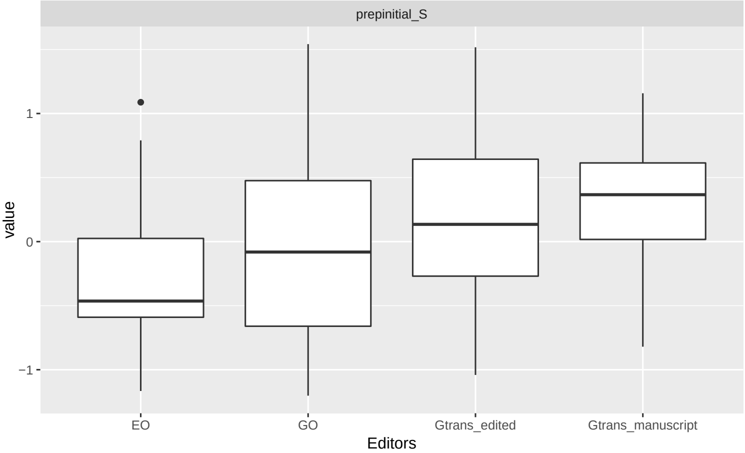

Figure: Boxplot for PPs as Theme

Distribution of subordinated clauses

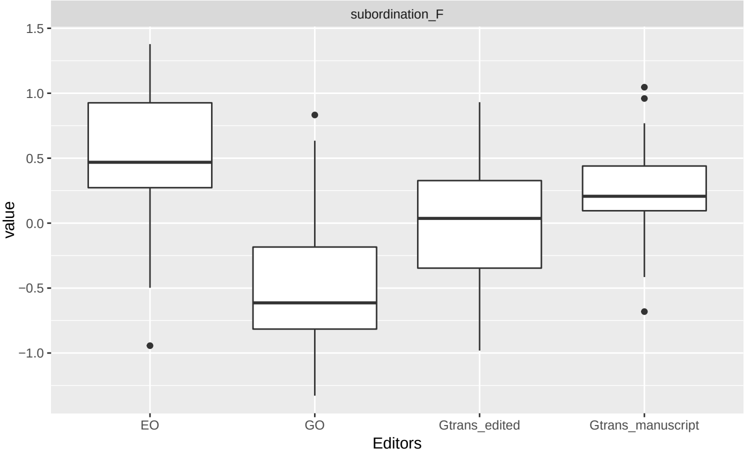

Figure: Boxplot for subordinated clauses

Linear Discriminant Analysis

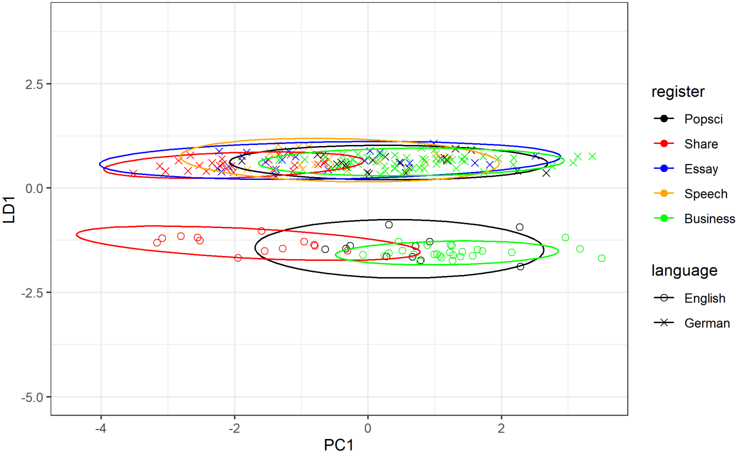

Figure: LDA with language as discriminant

Linear Discriminant Analysis

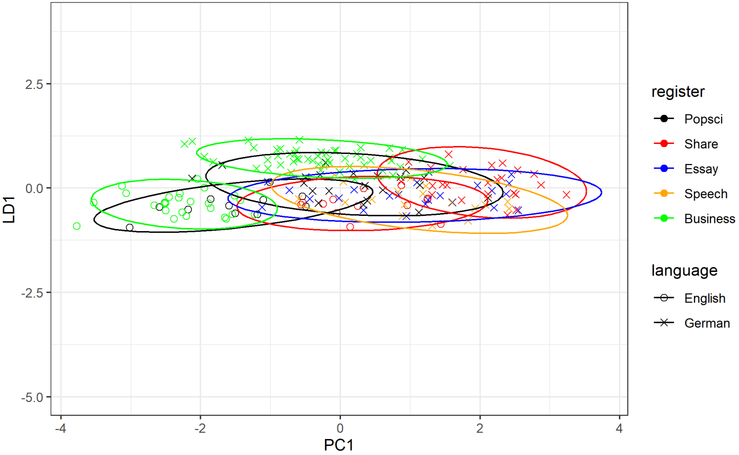

Figure: LDA with translation status as discriminant