Lexicogrammatical and Discursive Features in Translation and Interpreting

Ekaterina Lapshinova-Koltunski[a] **and Maria Kunilovskaya**[b], [a] Department of Language and Information Science, University of Hildesheim, Hildesheim, Germany; and [b] Department of Language Science and Technology, Saarland University, Saarbrücken, Germany

Key Points

- Translations have specific linguistic, i.e., lexicogrammatical and discursive features compared to originally-authored texts.
- Linguistic features of translation can be automatically extracted from text corpora and be analyzed quantitatively.
- Linguistic features of translation help to automatically differentiate between translated and non-translated texts.

Glossary

Translationese Specific frequency distribution of linguistic features of translations that make them statistically distinct from non-translations.

Translationese studies Studies of linguistic properties of translations.

Originally-authored texts Texts that were originally written and not translated.

Non-translations texts that were originally written and not translated.

Shining-through tendency of translated texts to reproduce source-language patterns instead of following the target language conventions.

Explicitation a tendency of translations to spell things out rather than leave them implicit.

Nomenclature

PoS Part-of-Speech
TTR Type-to-token ratio
SL Source language
TL Target language
MHD Mean hierarchical distance
MDD Mean dependency distance
NLP Natural Language Processing
UD Universal Dependencies

Abstract

This article provides an overview of lexicogrammatical and discursive features, e.g., infinitives, discourse connectives that can be used to analyze specific properties of translations that distinguish them from nontranslations called translationese. We classify the features according to various categories (e.g., word classes), describe the requirements to these features and summarize the results of using these features in the analysis of translationese.

Introduction

This article describes linguistically motivated and humaninterpretable features that capture the quantitative distinctions between translations and non-translations and are typical examples of translationese, i.e., specific properties of translated language.

Translationese and Lexicogrammar section briefly introduces the main concepts of translationese studies, an area of research focused on linguistic specificity of translations. **Methodology and Feature Design** section surveys theoretical underpinnings of feature selection and design and presents a methodological sketch of translationese studies based on lexicogrammatical and discursive features. In **Lexicogrammatical and Discursive Findings** section we demonstrate the effectiveness of these features in distinguishing translations and non-translations.

Translationese and Lexicogrammar

Translationese refers to the specific frequency distributions of linguistisc features of translations that make them statistically distinct from non-translations in the target language (Baker, 1995; Gellerstam, 1986; Teich, 2003). The foundations of translationese studies can be traced back to the philosophical conceptualizations of translation as a particular type of writing with its differences from originally-authored texts in the TL offered by Frawley (1984). Since the rise of the corpus-based approaches to text analysis in the last decade of the 20th century, this claim has been supported by considerable empirical evidence. The deviations of translations from the expected TL norm are rooted in the specificity of the translation process as well as cognitive and sociolinguistic factors involved with translation. These factors influence the linguistic choices made in this type of cross-lingual communication. The most researched translationese phenomena include: *shining-through* (a tendency to reproduce SL patterns rather than follow the target language conventions, Teich, 2003) *(over-)normalization* (a tendency to conform to the typical patterns of the TL or exaggerate them, Baker, 1993), *explicitation* (a tendency to spell things out rather than leave them implicit, Olohan & Baker, 2000) and *simplification* (a tendency to simplify the language used in translation Toury, 1995). In this article, these phenomena are referred to as translation properties and the term feature is used to mean quantitative characteristics of documents used to represent these properties in empirical analyses.

The general recommendations on the choice of features in translationese studies, originally formulated by Volansky et al. (2015), favor the lexicogrammatical and discursive features. The most effective candidate features should:

- capture relatively frequent linguistic items to minimize sparsity;
- be content-independent to reduce the influence of possible domain differences between translations and non-translations;
- be easy to interpret and yield insights regarding translational deviations from what is expected in non-translation.

The term lexicogrammatical is reminiscent of the ideas from Systemic Functional Grammar (Halliday & Hasan, 1989), which links language patterns to the main components of a communicative event (field, tenor, mode) to explain the language variation across the functional registers. This approach is particularly relevant to the study of translations viewed as resulting from a specific form of cross-lingual communication, where the particular aspects of this situation shape translated language. The most influential fact, i.e., the typological distance between the source and the target language (Dutta Chowdhury et al., 2020; Nikolaev et al., 2020), the translator's competence (De Sutter et al., 2017; Kunilovskaya & Lapshinova-Koltunski, 2019; Lapshinova-Koltunski, 2022; Rubino et al., 2016), the mode of communication (Bernardini et al., 2016; Kunilovskaya & Lapshinova-Koltunski, 2019; Lapshinova-Koltunski et al., 2021) and the register (Evert & Neumann, 2017; Neumann, 2013) of the source text. Discursive features mostly refer to such phenomena as cohesion and information density. In essence, lexicogrammatical and discursive features are measured by normalized frequencies of specific linguistic items such as passive voice constructions, relative clauses, personal pronouns or contrastive discourse markers. Note that a feature value often reflects the cumulative frequency of all items (e.g., the frequency of all items annotated as subordinate conjunctions).

Translationese indicators are typically revealed by comparing translated and non-translated texts of the same register in the TL. This comparison is often framed as an automatic text classification task supported by statistical analyses. The key methodological decision is the choice of the features to be extracted and tested. Most studies start with an extensive feature set and demonstrate that some features differentiate between translations and non-translations better than others (Baroni & Bernardini, 2006; Hu & Kuebler, 2021; Volansky et al., 2015). Although there is a demand for easily extractable and scalable features that can be used in NLP applications, this article focuses on human-interpretable features that shed light on the peculiarities of the translated language. Translationese studies that are interested in explanation for the observed peculiarities require extended corpus resources including parallel corpora where SL segments are aligned with their targets (see for example, Evert and Neumann (2017); Kunilovskaya and Lapshinova-Koltunski (2020)) or translations from several SLs into the same TL (Koppel & Ordan, 2011; Kunilovskaya et al., 2021; Rabinovich et al., 2017).

Methodology and Feature Design

The choice of features in translationese studies depends on the research methodology and the research goal. Some studies rely on manual or semi-automatic analysis of (parallel) concordances and use a few theoretically well-motivated linguistic items or groups

of them. In these studies, the feature selection is commonly guided by **contrastive analysis** of the source and target languages and **translation theory-backed expectations** of translational behavior. For example, based on contrastive knowledge about grammatical structures in English and Portuguese, Santos (1995) explores the use of tense and aspect verb forms in clearly defined translationally interesting types of contexts. The usual suspects for translationese deviations are cross-linguistic contrasts, where a specific item is obligatory in the SL and optional in the TL. A typical example is the category of discourse connectives which have become a fruitful object for studies of both human and machine translation (Becher, 2011; Castagnoli, 2009; Kunilovskaya & Kutuzov, 2017; Lapshinova-Koltunski, 2015, 2017a; Olohan, 2001). The invariably specific distributions of connectives in the translated language (regardless of the language pair and register) are linked to explicitation, shining-through or simplification (the latter trend is particularly visible in simultaneous interpreting (Przybyl et al., 2022)).

Other studies start with a large pool of potential translationese indicators and look at their ability to distinguish translations and non-translations. Their methods include machine learning techniques such as Principal Component Analysis, Linear Discriminant Analysis, text classification or clustering. In understandingoriented studies, the feature sets usually count several dozens of carefully hand-engineered items that are expected to be effective based on SL/TL contrastive analysis (see, for example, the implications of English-German contrasts for translation in Kunz et al. (2017)) and findings from corpus-based translationese studies. It is typical for these studies to consider language-pair-specific features, sometimes further limited to those that are shared between the source and target languages (Evert & Neumann, 2017). Apart from contrastive analysis and translation theory, candidate features are often inspired by **variational linguistics and register studies** (Biber, 1995; Neumann, 2013). It is known that register is as important a factor for shaping properties of translations as are the typological relations between the source and target languages. Translated registers are treated as language varieties and their distinctions can be effectively captured by features used to differentiate non-translated registers (Evert & Neumann, 2017; Kunilovskaya & Corpas Pastor, 2021; Lapshinova-Koltunski, 2017b; Poltorak, 2022).

Translationese studies that use lexicogrammatical and discursive features rely on corpora annotated with morphological and syntactic categories. Most recently the parsers like UDPipe (Straka & Straková, 2017) and Stanza (Qi et al., 2020) equipped with language models trained on the Universal Dependencies (UD) treebanks[1] have become popular. This framework offers consistent annotation of similar constructions across languages and uses universal tags for similar structural categories across many languages. The extraction procedures can include the straightforward counting of PoS tags or listed surface forms (e.g., connectives grouped according to their meaning into additive, contrastive, etc.) and more heavily engineered features that take into account the context, the morphological characteristics of a word or of its dependents/heads. Such features as TTR or MHD can be implemented not as frequency-based values but as metrics, calculated from the syntactic parse of each sentence. Examples of featureextracting pipelines and tools can be found in previous work in translationese studies and variational linguistics (Evert & Neumann, 2017; Katinskaya & Sharoff, 2015; Kunilovskaya & Lapshinova-Koltunski, 2020; Nini, 2015)[2].

The frequency features are typically estimated and normalized. The normalization basis can be selected with regard to the feature type. For example, modal predicates and relative clauses can be normalized per number of sentences, verbal forms per number of verbs, and morphological categories per tokens in the document. It is a good practice to exclude highly correlated features. Most computational studies use documents or chunks of text of over 400 tokens as their samples.

The features can be categorized in different ways depending on the research task pursued. An example categorization may include (*i*) word forms, (*ii*) word classes, (*iii*) sentence structure and types of clauses, (*iv*) other types of dependencies, (*v*) semantic types of discourse markers and (*vi*) textual measures. While translationese is predominantly a text-level rather than sentence-level phenomenon, the features counted at the level of individual forms and lexicogrammatical categories aggregate to characterize the discursive structure of translated language. For example, the frequency of modal predicates and parenthetical phrases cansignal the author's epistemic stance (e.g., *perhaps, of course, to my mind*).

Lexicogrammatical and Discursive Findings

The question of the relevance of a particular feature is commonly linked to its contribution to the translation detection task, which shows how effective a feature is in differentiating translations from non-translations. This can be done through either significance testing (see the studies on translated Chinese, for example Xiao (2010); Dai and Xiao (2011)) and/or on the basis of the performance of the feature sets in a monolingual binary classification task (translations vs. non-translations). In the case of an automatic classification of translations versus non-translations, the accuracy of such a classification measured with automatic scores plays a role. For instance, using the F1-score (a harmonic mean between precision and recall), classification results can be represented from zero to 100, with the latter being the best achievable result. We summarize the results from selected studies that use lexicogrammatical and discursive features to automatically tease apart translations and non-translations across several registers and language pairs in **Table 1**.

We also excluded studies of translations from multiple source languages. **Table 1** illustrates that F1-score varies depending on the language pair, with the worst score for Ukrainian-Russian in classification of translated and nontranslation fictional texts and the best result for the English-Spanish pair for technical texts. This means that lexicogrammatical and discursive features serves as good

[1]https://universaldependencies.org/introduction.html.
[2]https://sites.google.com/site/multidimensionaltagger, https://github.com/Askinkaty/MDRusanalyser, https://github.com/kunilovskaya/translationese45/tree/master.

Table 1 Selected text-level translationese classification results in % on lexicogrammatical and discursive features across translation directions and registers.

lang.pair3	*register*	*F1*	*reference*
FR-EN	Europarl	83.6	Rabinovich and Wintner (2016)
DE-EN	Europarl	80.00	Kunilovskaya et al. (2024)
EN-DE	Europarl	88.8	
DE-EN	multiregister	77.00	Evert and Neumann (2017)
EN-ES	Europarl fiction	96.2	Poltorak (2022)
EN-ES		77.3	
EN-ES	technical	97.6	Ilisei et al. (2010)
EN-RU	mass media	90.2	Kunilovskaya (2023)
EN-RU	fiction	75.6	
DE-RU	fiction	84.4	
ES-RU	fiction	74.4	Kunilovskaya et al. (2021)
SV-RU	fiction	71.1	
UK-RU	fiction	59.4	

Note that these studies report results based on a set of features only and not for each feature separately (see, for example, Volansky et al. (2015); Hu and Kuebler (2021)).

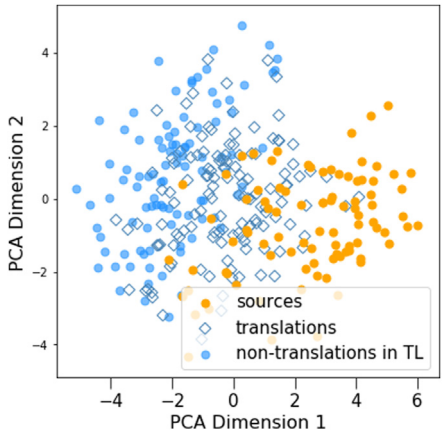

Fig. 1 The difference of English-to-German translations from non-translations captured by the lexicogrammatical and discursive shining-through indicators. Figure 3 from Kunilovskaya and Lapshinova-Koltunski (2020).

distinctive indicators of translated language. Fig. 1 demonstrates that translated texts (shown by empty square markers in the graph) form a separate category of texts located between their sources (orange circle markers to the right of the graph) and comparable non-translations in the TL (blue filled markers in the left part of the graph) based on their lexicogrammatical and discursive features.

Conclusion and Outlook

This article aimed at an account of lexicogrammatical and discursive features of translated language that can be linked to translationese. An important aspect of these features is their empirical character—they can be (semi-)automatically extracted from collections of texts and quantified according to their distributions. These features proved to be effective in identifying translated texts, i.e., in differentiating translated texts from non-translations.

However, it is important to note that existing feature lists are not exhaustive and can be extended in future works. Moreover, we did not go into details on which of the features is linked to which translation property, such as simplification, explicitation, shining through and normalization, as this can also be dependent on the specific register and language pair involved. Moreover, these features can be combined with the features from other levels of linguistic description, e.g. those based on probabilistic approaches, such as Information theory.

Acknowledgments

The research presented here was partly funded by DFG (German Research Foundation) — Project-ID 232722074 — SFB 1102.

References

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In G. F, M. Baker, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 233—250). Amsterdam: Benjamins.

Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target, 7*(2), 223—243.

Baroni, M., & Bernardini, S. (2006). A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing, 21*(3), 259—274.

Becher, V. (2011). *Explicitation and implicitation in translation. A corpus-based study of English-German and German-English translations of business texts*. University of Hamburg (Unpublished doctoral dissertation).

Bernardini, S., Ferraresi, A., & Miličević, M. (2016). From EPIC to EP-TIC — Exploring simplification in interpreting and translation from an intermodal perspective. *Target, 28*(1), 61—86.

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

Castagnoli, S. (2009). *Regularities and variations in learner translations: A corpus-based study of conjunctive explicitation*. University of Pisa (Unpublished doctoral dissertation).

Dai, G., & Xiao, R. (2011). "SL shining through" in translational language: A corpus-based study of Chinese translation of English passives. *Translation Quarterly, 62*(1), 85—108.

De Sutter, G., Cappelle, B., De Clercq, O., Loock, R., & Plevoets, K. (2017). Towards a corpus-based, statistical approach of translation quality : Measuring and visualizing linguistic deviance in student translations. *Linguistica Antverpiensia, New Series — Themes in Translation Studies, 16*(1), 25—39.

Dutta Chowdhury, K., España-Bonet, C., & van Genabith, J. (2020, December). Understanding translationese in multi-view embedding spaces. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th international conference on computational linguistics* (pp. 6056—6062).

Evert, S., & Neumann, S. (2017). The impact of translation direction on characteristics of translated texts. A multivariate analysis for English and German. In G. De Sutter, M.-A. Lefer, & I. Delaere (Eds.), *Empirical translation studies: New methodological and theoretical traditions* (Vol. 300). Mouton de Gruyter.

Frawley, W. (1984). Prolegomenon to a theory of translation. In W. Frawley (Ed.), *Translation: Literary, linguistic and philosophical perspectives* (pp. 159—175). London: Associated University Press.

Gellerstam, M. (1986). Translationese in Swedish novels translated from English. In L. Wollin, & H. Lindquist (Eds.), *Translation studies in scandinavia* (pp. 88—95). Lund: CWK Gleerup.

Halliday, M., & Hasan, R. (1989). *Language, context, and text: Aspects of language in a social-semiotic perspective* (2nd ed.). Oxford: Oxford University Press.

Hu, H., & Kuebler, S. (2021). Investigating translated Chinese and its variants using machine learning. *Natural Language Engineering, 5*(3), 339—372.

Ilisei, I., Inkpen, D., Pastor, G. C., & Mitkov, R. (2010). Identification of translationese : A machine learning approach. In *International conference on intelligent text processing and computational linguistics* (pp. 503—511). Berlin Heidelberg: Springer.

Katinskaya, A., & Sharoff, S. (2015, 5-11 September). Applying multi-dimensional analysis to a Russian webcorpus: Searching for evidence of genres. In J. Piskorski, L. Pivovarova, J. Šnajder, H. Tanev, & R. Yangarber (Eds.), *Proceedings of the 5th workshop on balto-slavic natural language processing* (pp. 65—74).

Koppel, M., & Ordan, N. (2011). Translationese and its dialects. In *Proceedings of the 49th acl* (Vol. 1, pp. 1318—1326). ACL.

Kunilovskaya, M. (2023). *Translationese indicators for human translation quality estimation (based on english-to-russian translation of mass-media texts)*. University of Wolverhamptopn (Unpublished doctoral dissertation).

Kunilovskaya, M., & Corpas Pastor, G. (2021). Translationese and register variation in english-to-russian professional translation. In V. X. Wang, D. Li, & L. Lim (Eds.), *New frontiers in translation studies* (pp. 133—180). Singapore: Springer Nature.

Kunilovskaya, M., & Kutuzov, A. (2017). Testing target text fluency: A machine learning approach to detecting syntactic translationese in English-Russian translation. In K. Menzel, E. Lapshinova-Koltunski, & K. Kunz (Eds.), *New perspectives on cohesion and coherence: Implications for translation* (pp. 75—104). Language Science Press.

Kunilovskaya, M., & Lapshinova-Koltunski, E. (2019). Translationese features as indicators of quality in English-Russian human translation. In *Proceedings of the HiT-IT 2019 workshop* (pp. 47—56). Varna, Bulgaria: Incoma Ltd. (Shoumen, Bulgari).

Kunilovskaya, M., & Lapshinova-Koltunski, E. (2020). Lexicogrammatic translationese across two targets and competence levels. In *Proceedings of the 12th lrec* (pp. 4102—4112). Marseille, France: ELRA.

Kunilovskaya, M., Lapshinova-koltunski, E., & Mitkov, R. (2021, 7—11 November). Translationese in Russian literary texts. In *Proceedings of the 5th joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature* (pp. 101—112). ACL.

Kunilovskaya, M., Chowdhury, K. D., Przybyl, H., España i Bonet, C., & Van Genabith, J. (2024, 24-27 June). Mitigating translationese with GPT-4: Strategies and performance. In *Proceedings of the 25th EAMT*. Sheffield, UK: ACL.

Kunz, K., Degaetano-Ortlieb, S., Lapshinova-Koltunski, E., Menzel, K., & Steiner, E. (2017). English-german contrasts in cohesion and implications for translation. In G. De Sutter, M.-A. Lefer, & I. Delaere (Eds.), *Empirical translation studies: New methodological and theoretical traditions* (Vol. 300). Mouton de Gruyter.

Lapshinova-Koltunski, E. (2015, September). Exploration of interand intralingual variation of discourse phenomena. In B. Webber, M. Carpuat, A. Popescu-Belis, & C. Hardmeier (Eds.), *Proceedings of the 2nd discomt* (pp. 158—167). Lisbon, Portugal: Association for Computational Linguistics.

Lapshinova-Koltunski, E. (2017a). Cohesion and translation variation: Corpus-based analysis of translation varieties. In K. Menzel, E. Lapshinova-Koltunski, & K. Kunz (Eds.), *New perspectives on cohesion and coherence: Implications for translations* (Vol. 6, pp. 103—128). Berlin: Language Science Press.

Lapshinova-Koltunski, E. (2017b). Exploratory analysis of dimensions influencing variation in translation: The case of text register and translation method. In G. De Sutter, M.-A. Lefer, & I. Delaere (Eds.), *Empirical translation studies: New methodological and theoretical traditions* (Vol. 300, pp. 207—234). Mouton de Gruyter (TILSM series).

Lapshinova-Koltunski, E. (2022). Detecting normalization and shining-through in novice and professional translations. In *Extending the scope of corpus-based translation studies* (pp. 182—206). Bloomsbury Academic.

Lapshinova-Koltunski, E., Bizzoni, Y., Przybyl, H., & Teich, E. (2021, May). Found in translation/interpreting: Combining data-driven and supervised methods to analyse cross-linguistically mediated communication. In Y. Bizzoni, E. Teich, C. España-Bonet, & J. van Genabith (Eds.), *Proceedings for the 1st motra* (pp. 82—90). ACL. online.

Neumann, S. (2013). *Contrastive register variation. A quantitative approach to the comparison of English and German*. Berlin, Boston: Mouton de Gruyter.

Nikolaev, D., Karidi, T., Kenneth, N., Mitnik, V., Saeboe, L., & Abend, O. (2020). Morphosyntactic predictability of translationese. *Linguistics Vanguard, 6*(1).

Nini, A. (2015). *Multidimensional analysis tagger*.

Olohan, M. (2001). Spelling out the optionals in translation: A corpus study. In P. Rayson, A. Wilson, T. McEnery, A. Hardie, & S. Khoja (Eds.), *Proceedings of the corpus linguistics* (pp. 423—432). Lancaster University.

Olohan, M., & Baker, M. (2000). Reporting that in translated English: Evidence for subconscious processes of explicitation? *Across Languages and Cultures, 1*(2), 141—158.

Poltorak, K. (2022). Computational approaches to register as a factor in english-to-spanish translation. In *New trends in translation and technology* (pp. 82—90).

Przybyl, H., Lapshinova-Koltunski, E., Menzel, K., Fischer, S., & Teich, E. (2022, 20-25 June). EPIC-UdS - Creation and applications of a simultaneous interpreting corpus. In *Proceedings of the 13th LREC* (pp. 1193—1200). Marseille, France: ELDA.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th ACL: System demonstrations*.

Rabinovich, E., & Wintner, S. (2016). Unsupervised identification of translationese. *TACL, 3*, 419—432. https://doi.org/10.1162/tacl_a_00148

Rabinovich, E., Ordan, N., & Wintner, S. (2017, 30 July—4 August). Found in translation: Reconstructing phylogenetic language trees from translations. In *Proceedings of the 55th acl* (pp. 530—540). ACL.

Rubino, R., Lapshinova-Koltunski, E., & van Genabith, J. (2016). Information density and quality estimation features as translationese indicators for human translation classification. In *Proceedings of NAACL-HT-2006* (pp. 960—970) (San Diego, California).

Santos, D. (1995, 29—30 May). On grammatical translationese. In *Proceedings of the 10th nodalida* (pp. 59—66). University of Helsinki.

Straka, M., & Straková, J. (2017, 3—4 August). Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In D. Zeman, et al. (Eds.), *Proceedings of the conll 2017 shared task: Multilingual parsing from raw text to universal dependencies* (pp. 88—99). ACL.

Teich, E. (2003). *Cross-linguistic variation in system and text. A methodology for the investigation of translations and comparable texts*. Berlin: Mouton de Gruyter.

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

Volansky, V., Ordan, N., & Wintner, S. (2015). On the features of translationese. *Digital Scholarship in the Humanities, 30*(1), 98—118.

Xiao, R. (2010). How different is translated Chinese from native Chinese? *International Journal of Corpus Linguistics, 15*(1), 5—35.