Are Translated Texts Useful for Gradient Word Order Extraction?

Amanda Kann
Stockholm University
Stockholm, Sweden
amanda.kann@su.se

Abstract

Gradient, token-level measures of word order preferences within a language are useful both for cross-linguistic comparison in linguistic typology and for multilingual NLP applications. However, such measures might not be representative of general language use when extracted from translated corpora, due to noise introduced by structural effects of translation. We attempt to quantify this uncertainty in a case study of subject/verb order statistics extracted from a parallel corpus of parliamentary speeches in 21 European languages. We find that word order proportions in translated texts generally resemble those extracted from non-translated texts, but tend to skew somewhat toward the dominant word order of the target language. We also investigate the potential presence of underlying source language-specific effects, but find that they do not sufficiently explain the variation across translations.

1 Introduction

When investigating cross-lingual transfer in multilingual language models, NLP researchers often rely heavily on data from typological databases such as WALS (Dryer and Haspelmath, 2013) for quantitative measures of language distance.[1] These databases typically reduce cross-linguistic variation to a set of categorical binary distinctions, obscuring the intra-linguistic variation present in many features (Wälchli, 2009), including word order.

This type of gradient variation is better captured by continuous token-level measures, such as statistical distributions of specific constructions (e.g. individual word order types) observed in annotated corpora (Levshina et al., 2023; Baylor et al., 2024). Corpus-based measures also allow for greater transparency and reproducibility than manual categorical judgments, and enable cross-linguistic comparisons at the potential scale of thousands of languages with maintained methodological consistency (see e.g. Östling and Kurfalı, 2023).

However, care must be taken to ensure that the selected texts are both sufficiently representative of their respective languages and comparable across languages, in order to control for variation resulting from differences between text types. Using massively parallel texts ensures that text type and pragmatic context will be identical across all analyzed languages, reducing the risk of misleading cross-linguistic comparisons (Ebert et al., 2024).

Parallel texts are also inherently translational, however, and could thus diverge structurally from original (non-translated) texts because of artefacts introduced in the translation process. For instance, translated texts commonly contain less lexical and grammatical variation than original texts in the same language (*regularization*). Structural properties of the source language may also be retained in translation, even when they are marked in the target language (*source language interference*). Thorough descriptions of features theorized to be cross-linguistically typical of translated text can be found in translation studies literature (e.g. Baker, 1993).

Translational artefacts can be strong enough to train reliable classifiers for automatic detection of translated texts (Volansky et al., 2015), and to accurately determine the relative genealogical distance between different source-target language pairs based only on cues in translations (Rabinovich et al., 2017). The cited studies rely heavily on syntactic features (most commonly part-of-speech *n*-grams), suggesting that translational artefacts could have a direct impact on word order proportions – however, word order (particularly of subject, verb and object) is not necessarily well captured by part-of-speech sequences, and the relationship between general and source-language specific translation effects in this domain has yet to

[1] For an overview of common approaches and typological distance measures in cross-lingual transfer research, see Philippy et al. (2023).

Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2025), pages 177–182

be systematically studied.

We therefore conduct an analysis of translational artefacts in gradient subject/verb order extraction from a parallel corpus of transcribed speeches with high-quality human translations in 21 languages. Our aim is to investigate:

- whether gradient word order statistics extracted from translations vary significantly from those extracted from original texts, and
- whether the direction or amplitude of such differences is influenced by word order preferences in the source language.

We expect that observed variation will be stronger in the direction of the dominant word order (as a result of regularization), and that source language interference will pull the word order proportions of translations toward the proportions observed in their source texts.

2 Data

We use *CoStEP* (Graën et al., 2014), a cleaned and turn-level aligned version of the *Europarl* parallel corpus (Koehn, 2005). *Europarl* consists of transcribed speeches and human translations in 21 European languages, obtained from European Parliament proceedings between 1996 and 2011. Since both the original speeches and their translations are present in the corpus, the source language for any given translated sentence is always known – this quality is essential for disambiguating potential source language-specific effects. All 420 possible source-target language pairs occur in the corpus, with data sizes ranging between 21 885 (Estonian–Bulgarian) and 8 738 402 (English–French) tokens. The corpus contains considerably more text (both original and translated) in the 11 languages that already had official EU language status prior to the expansions in 2004 and 2007.

To enable syntactic analysis, all texts (both original and translated, across all 21 languages) have been automatically tokenized, part-of-speech tagged and dependency parsed using the monolingual *Universal Dependencies* (Nivre et al., 2020) models availale through `Stanza` (Qi et al., 2020). While the parsing accuracy of these models varies somewhat across languages, noise from automatic annotation appears to have a minimal impact on word order proportions extracted from larger corpora (Levshina et al., 2023) – in addition, cross-linguistic performance differences do not directly affect comparisons between translations into the same language (regardless of source language).

3 Word order extraction

Subject/verb order can be defined and delimited in several ways, capturing different constructions and patterns of variation. We use a combination of part-of-speech and dependency tags on a given token and its direct head, operationalizing the relative order of nominal subject and verb as [NOUN|PROPN] $\overleftarrow{\texttt{nsubj}}$ [VERB] (i.e. a nominal subject relation between a noun or proper noun and a verb). Following Ebert et al. (2024), we only consider main clauses, and in auxiliary constructions we use the position of the finite verb (which may be an auxiliary) rather than the lexical verb. We include both transitive and intransitive verbs, and both declaratives and interrogatives; however, we distinguish these categories in extraction so that they can be analyzed separately.

We split the corpus by target language and compute the relative frequencies of both possible word orders (subject-verb and verb-subject) separately per source language.[2] The resulting word order proportions for each source-target pair are then compared to the reference proportion extracted from original texts in the target language.

4 General translation effects

Figure 1 displays the distributions of verb-subject (VS) order proportions per language pair, grouped by target language and sorted by VS proportion in original texts in the target language. All languages in the corpus prefer subject-verb (SV) order[3], to varying degrees. The highest VS proportions are found in German (`de`), Estonian (`et`), Swedish (`sv`) and Dutch (`nl`); this is expected, as their dominant word order in main clauses is typically analyzed as *verb-second* (or, for spoken Estonian, *verb-third*) rather than SV (Vihman and Walkden, 2021).

Overall, the proportions observed in translated texts are similar to original texts – the mean difference across language pairs is -0.017. However, there is also variation between translated texts with different source languages. Even for French (`fr`), which has the lowest dispersion across translations

[2] Following Levshina et al. (2023), we set a minimum total frequency threshold of 500 occurrences of the construction of interest – 412 of 420 language pairs in the corpus meet this threshold for nominal subject/verb constructions.

[3] This preference is expected for all languages in the *Europarl* sample; see section 6 for further discussion.

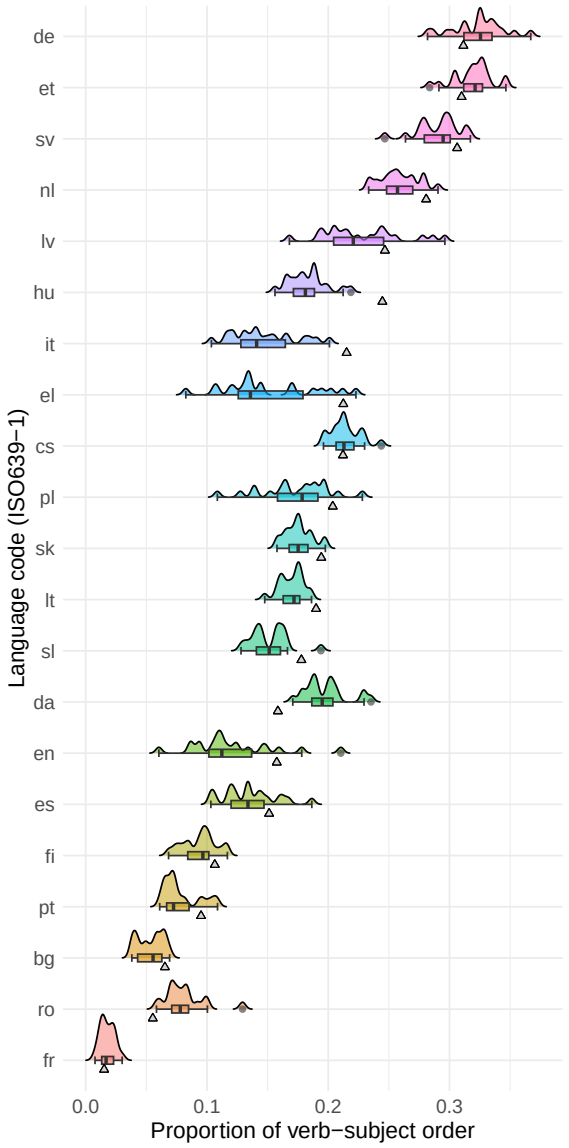

Figure 1: Distributions of VS order proportions in translated *Europarl* texts with different source languages, grouped by target language. The box plots display the median, interquartile range and whiskers extending to 1.5 IQR, with outliers plotted as individual points. The gray triangles indicate the VS order proportion in original texts in the respective target language.

($n_t = 20; \sigma_t = 0.0057; IQR_t = [0.013, 0.023]$), grouping the original French data by year of production (as a non-translational reference variable) results in a distribution with slightly lower dispersion ($n_{year} = 17; \sigma_{year} = 0.0027; IQR_{year} = [0.013, 0.017]$). Similar results are found for German (de), suggesting the presence of some unexplained variation specific to translated texts.

It should be noted that this variation is of a similar scale to the differences resulting from operationalizing the word order of interest differently; for instance, including only intransitive sentences results in higher dispersion for both the translations ($\sigma_{t_{Intr}} = 0.0077$) and the reference population ($\sigma_{year_{Intr}} = 0.0038$).

For 15 of 21 languages in the sample, the VS proportion in original texts is higher than both the median and upper quartile of VS proportions in the population of translations into that language; several original texts (e.g. Italian (it) and Hungarian (hu) would be outliers in their respective populations. The overall population of differences in VS proportion between translations and original texts (across all target languages) is approximately normally distributed, with a slight negative skew ($\tilde{x} = -0.015, IQR = -0.034, 0.004$). This tendency toward SV order in translations aligns with our hypothesis, and may be a reflection of the regularization effects described in section 1.

5 Source language-specific effects

To examine the potential effects of source language interference, VS order proportions from the set of translated turns in a given source-target language pair are also compared to the proportions extracted from the same turn set in the source language. Figure 2 plots this relationship for all source languages, into three target languages with different mean VS order proportions and dispersions across translations. We find no signifi-

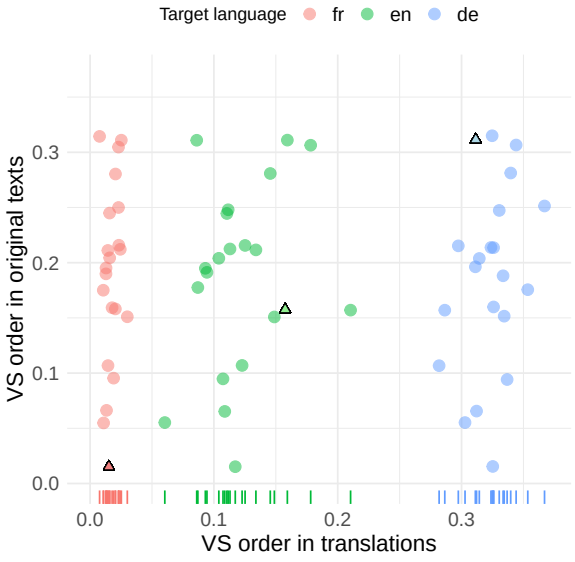

Figure 2: VS order proportions in translations (into French (fr), English (en) and German (de)) and in original texts for each language. The triangles indicate the proportions in original texts in the target languages.

cant correlations between source and target proportions for any individual target language, as would have been expected from the hypothesized source language interference model. Across the entire population, we find a weak positive correlation between source proportion and translation effect (the difference between proportions in translations and original texts in the target language), but it explains very little of the variation in translation effects ($\beta = 0.08; CI_{95\%,\beta} = [0.04, 0.12]$; $R^2 = 0.04; CI_{95\%,R^2} = [0.01, 0.08]$).

A potentially confounding source language effect is the proportion of different clause types in the source texts (assuming that they are carried over to the target language), since word order preferences for different clause types vary across languages. For instance, SV order in transitive clauses is stricter than in intransitive clauses in some European languages, such as Spanish and Latvian (Dryer, 2013) – in these languages, the proportion of transitive clauses should have a greater impact on the extracted word order proportion than in languages where the two clause types pattern similarly. The *Europarl* data supports this claim: there is a positive correlation between the proportion of intransitive clauses and VS order proportion in both Spanish and Latvian. However, we unexpectedly find no correlation between intransitive clause proportions in source texts and translations, either for these languages or across the entire sample. While the turn-level alignment of *CoStEP* is too coarse to meaningfully investigate this further, individual clause-level comparison in a word-aligned parallel corpus could verify to what extent properties of source language clauses which may influence word order proportions are preserved in translation.

6 Conclusions

In this study, we analyzed the general and source language-specific effects of translation on verb/subject order statistics extracted from *Europarl*. We observed a general tendency toward rigid SV order in translations compared to original texts, in line with the broader *regularization* effect discussed in translation studies literature. Unexpectedly, we found that word order proportions in the source texts do not sufficiently explain this tendency, at least when averaged at corpus level. This suggests that controlling for source language factors will not reliably reduce uncertainty when using translated texts to approximate word order distributions in original texts.

Crucially, the issue of translational artefacts should not disqualify good-quality translations from use in the extraction of gradient word order typology, assuming that the uncertainty in the extracted proportions is properly taken into account in interpretation – as is good practice for any parameter by which syntactic properties of a text may vary. As with text genres, including multiple different source languages in a corpus of translations may reduce the risk of unrepresentativity. A well-motivated theoretical definition (and operationalization) of the word order feature of interest is also necessary in order to make valid cross-linguistic comparisons based on extracted word order proportions. With these aspects in mind, even an uncertain estimate of gradient word order proportions will encode considerably more fine-grained and useful comparative information than the customary binary word order classifications.

It is important to note the restricted scope of this case study. We only investigate one word order feature, which is particularly prone to pragmatically motivated variation in many languages. Additionally, the language sample in *Europarl* is highly areally and genealogically skewed. Most languages in the sample are members of the Standard Average European *Sprachbund*, and are thus likely to share some cross-linguistically marked syntactic features – for instance, inverted subject/verb order in polar questions (Haspelmath, 2001). *Europarl* is also unusual in other aspects, such as text genre (formal speeches, with higher average sentence and utterance length than spontaneous informal speech) and the purpose of translation (accurate representation of the original speeches, likely prioritizing clear language). These properties should be kept in mind when applying our findings to other contexts.

We hope that this study can serve as a framework for further cross-lingual investigations of the effects of translation on word order. In addition to analyzing more word order features, future work could cover a larger and more diverse language sample by making use of machine translations, which are an interesting object of analysis in their own right. Machine translations appear to produce different translational artefacts to human translations (Bizzoni et al., 2020), and – not least because of the prevalence of machine translated text in large text datasets – a comparison between word order extractions from human and machine translations would be very useful.

Limitations

In addition to the areal and genealogical bias discussed in section 6, the sample in *Europarl* consists entirely of high-resource languages. Accurate pre-trained parsing models are only available for a fraction of the world's languages (`Stanza` provides UD models for fewer than 100 languages), and high quality training data for PoS tagging and dependency parsing is similarly scarce.

Our word order extraction method is simple, and the per-text average measure obscures the various underlying causes of potential word order variation. Subject/verb order preferences can vary structurally across clause types or nominal categories, or pragmatically for information structure or discourse reasons – this method can only disambiguate between the structural variation sources which are accounted for in the chosen word order operationalization.

Finally, the analysis of source language-specific effects is complicated by the potential presence of indirect translations (where an intermediate language is used in the translation process). Ustaszewski (2021) reports that translations in *Europarl* produced after the official EU language expansion in 2004 more likely use an intermediate language (most commonly English), while earlier translations are more likely direct. The general impact of an intermediate language on the presence of source language artefacts in translations is unclear and warrants further investigation.

Acknowledgments

This work was made possible by individual PhD student funding from the Department of Linguistics at Stockholm University. We thank Bernhard Wälchli and Robert Östling for their valuable comments on prior versions of this paper.

Supplementary materials

The code used to produce the results and figures presented in this paper is available at `https://github.com/amandakann/sigtyp2025`, under the GPL-3.0 license.

References

Mona Baker. 1993. Corpus Linguistics and Translation Studies — Implications and Applications. In Mona Baker, Gill Francis, and Elena Tognini-Bonelli, editors, *Text and Technology*, pages 9–24. John Benjamins Publishing Company, Amsterdam.

Emi Baylor, Esther Ploeger, and Johannes Bjerva. 2024. Multilingual Gradient Word-Order Typology from Universal Dependencies. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 42–49, St. Julian's, Malta. Association for Computational Linguistics.

Yuri Bizzoni, Tom S Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How Human is Machine Translationese? Comparing Human and Machine Translations of Text and Speech. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 280–290, Online. Association for Computational Linguistics.

Matthew S. Dryer. 2013. Order of Subject and Verb (v2020.4). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo. Type: Data set.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Christian Ebert, Balthasar Bickel, and Paul Widmer. 2024. Areal and phylogenetic dimensions of word order variation in Indo-European languages. *Linguistics*, 62(5):1085–1116.

Johannes Graën, Dolores Batinić, and Martin Volk. 2014. Cleaning the Europarl Corpus for Linguistic Applications. In *Konvens 2014*, Hildesheim. Stiftung Universität Hildesheim.

Martin Haspelmath. 2001. The European linguistic area: Standard Average European. In Martin Haspelmath, Ekerhard König, Wulf Oesterreicher, and Wolfgang Raible, editors, *Language Typology and Language Universals*, number 20/2 in Handbücher zur Sprach- und Kommunikationswissenschaft [HSK], pages 1492–1510. De Gruyter Mouton, Berlin, New York.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Natalia Levshina, Savithry Namboodiripad, Marc Allassonnière-Tang, Mathew Kramer, Luigi Talamo, Annemarie Verkerk, Sasha Wilmoth, Gabriela Garrido Rodriguez, Timothy Michael Gupton, Evan Kidd, Zoey Liu, Chiara Naccarato, Rachel Nordlinger, Anastasia Panova, and Natalia Stoynova. 2023. Why we need a gradient approach to word order. *Linguistics*, 61(4):825–883.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of the Twelfth Language Resources*

and Evaluation Conference (LREC'20), pages 4034–4043, Marseille, France. European Language Resources Association (ELRA).

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Towards a Common Understanding of Contributing Factors for Cross-Lingual Transfer in Multilingual Language Models: A Review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5877–5891, Toronto, Canada. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Ella Rabinovich, Noam Ordan, and Shuly Wintner. 2017. Found in Translation: Reconstructing Phylogenetic Language Trees from Translations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 530–540, Vancouver, Canada. Association for Computational Linguistics.

Michael Ustaszewski. 2021. Towards a machine learning approach to the analysis of indirect translation. *Translation Studies*, 14(3):313–331.

Virve-Anneli Vihman and George Walkden. 2021. Verb-second in spoken and written Estonian. *Glossa: a journal of general linguistics*, 6(1):15.

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

Bernhard Wälchli. 2009. Data reduction typology and the bimodal distribution bias. *Linguistic Typology*, 13(1):77–94.

Robert Östling and Murathan Kurfalı. 2023. Language Embeddings Sometimes Contain Typological Generalizations. *Computational Linguistics*, 49(4):1–49.