Machine-Translated Texts from English to Polish Show a Potential for Typological Explanations in Source Language Identification

Damiaan J. W. Reijnaers
University of Amsterdam
info@damiaanreijnaers.nl

Elize Herrewijnen
Utrecht University
National Police Lab AI
e.herrewijnen@uu.nl

Abstract

This work examines a case study that investigates (1) the achievability of extracting typological features from Polish texts, and (2) their contrastive power to discriminate between machine-translated texts from English. The findings indicate potential for a proposed method that deals with the explainable prediction of the source language of translated texts.

1 Introduction

In the modern-day world of global interconnectedness, *the act of translation* has evolved into an indispensable part of daily life as a result of the growing availability of ever more advanced translation engines (Vieira et al., 2021, pp. 1515–16). This trend has been further amplified by the increasing accessibility of such tools; *e.g.*, through their integration into messaging services or social media platforms (Xinxing (2023); Turovsky (2016)). The recently acquired prominent position of translation tools in human society brings attention to the value of comparatively un(der)explored areas in machine translation (MT) that are more ethical in nature.

Found within this space is the task of determining the source language of a machine-translated text, also referred to as Source Language Identification[1] (SLI). This task may not only contribute to the qualitative improvement of translation engines, it further has a practical application in the field of forensics—bad actors use translation tools too.

The viability of SLI relies on the premise that, despite considerable progress in MT as a result of 'the transformer revolution' (Zhang and Zong, 2020, p. 2229), machine translations may still be imperfect or *unnatural*: they may contain artifacts that indicate that a text originated in a different language, even when being grammatically and semantically sound. Nida and Taber (1969) coined the term 'translationese' to refer to this phenomenon that is now widely recognised in the literature. Translationese may arise from various causes, with the characteristics of the *source language* (*i.e.*, source language interference) playing a prominent role (Rabinovich et al., 2017). An understanding of the features leading to such perceived linguistic unnaturalness might therefore allow the source language to be identified from translated text alone. We propose a methodology that exploits these *symptoms of translationese* to approach the task of SLI.

Although research on translationese has mostly concerned *human* translation, an emerging line of work specifically focuses on MT, with the majority of effort in this area dealing with evaluating MT models (*e.g.*, Graham et al. (2020); Kurokawa et al. (2009)). While some authors presuppose that translationese across humans and machines is similar (Riley et al., 2020, p. 7738), empirical evidence suggests that *structural* properties of the source language (*i.e.*, grammar) 'shine through' more overtly in text translated by machines (Bizzoni et al., 2020, p. 288). As MT is ultimately trained on human translation, of particular interest is therefore how structural features of human translationese can be identified. A common strategy for this involves training a classifier to leverage *surface features* indicative of structural translationese: surface features comprise easily observable attributes of texts, such as parts of speech (PoS). Particularly the latter have demonstrated encouraging performance in the form of n-grams (Baroni and Bernardini (2005, p. 268); Rabinovich et al. (2017, p. 534), Pylypenko et al. (2021, p. 8603)). Drawing on these results, we employ such features in our method.

While SLI remains largely unexplored, many have studied its equivalent in (human) second-language acquisition: *Native* Language Identification (NLI[2]). This field was initially researched

[1] To our knowledge, the task was mentioned only once in a recent paper by La Morgia et al. (2023), coincidentally aligning with our own formulation of the novel task.

[2] Not to be confused with Natural Language Inference,

by Koppel et al. (2005) and gained momentum through a shared task (Tetreault et al., 2013). The inherent reliance of both NLI and SLI on leveraging features of language interference makes the former a highly relevant field. A shortcoming of many NLI approaches is their lack of explainability (Berti et al., 2022, p. 8); a quality that is naturally demanded by the field of forensics (Cheng, 2013, pp. 547–49), and a quality that could provide useful insight into the limitations of current approaches in MT. As we exploit the structural properties of a source language that hint at the origin through its artifacts, correspondingly, explanations ought to be in terms of the *structural differences* between the source and target language of a translated text. We aim to achieve this by reformulating SLI as a typological feature prediction task. Such features are the products of the field of linguistic typology and serve to *distinguish between the structural properties* of languages (Daniel, 2010, pp. 1–2). Consequently, they have the capacity to provide human-interpretable explanations that are linguistically grounded. Berzak et al. (2014) show that the typology of native languages are predictable within the context of NLI, providing further ground to our approach. In a paper published in the field of law, Kredens et al. (2020) similarly advocate the need for typology-based explanations in SLI-like contexts, indicating a convergence of ideas. Our paper contributes by presenting a practical implementation, while also adding to its theoretical foundation.

To further underline the potential of a typological approach to achieve explainable SLI, we present a case study that examined the feasibility of extracting typological features from Polish texts, and their capacity to discriminate between translated texts from English. The Slavic language family poses an interesting testbed for such analyses, as it exhibits unique features in contrast to English (§2.1), while still being in relative linguistic proximity. Our preliminary experiments indicate that structural features reminiscent of the origin language display significant promise for typology prediction to warrant further research that implements the methodology proposed in this work. We are currently examining the effectiveness of our method in practice.

In the following section, we provide the aforementioned experiments. The paper then proceeds by explicating our proposed methodology. It concludes by discussing the findings and limitations.

which is also commonly abbreviated as NLI in the literature.

2 Experimenting with Polish and English

To gauge the exploitability of features specific to the Polish language, we conducted two experiments. The first experiment analyses Polish word order to gain an intuition on the practical utility of the features. Experiment 2 then compares the applicability of *all* features listed in the following subsection. All code relating to the experiments and the scraping of the data can be found on GitHub.[3]

2.1 Language-specific features of Polish

Word order Polish is a strongly inflected language and therefore exhibits a relatively flexible word order (Kuh, 1990). This manifests at the level of the constituent, *i.e.*, '**S**ubject–**O**bject–**V**erb order' (Kubon et al., 2016, p. 16), but also at the level of parts of speech (PoS); *e.g.*, adjectives may be placed both before *and* after a noun (Bielec (2012, p. 211); Siewierska and Uhlířová (1998, pp. 109, 168, 134–37)). These differences may lead to errors in machine-translated texts to and from English (Popović and Arčan, 2015, p. 98, 100).

Verbal aspect Polish explicitly marks verbal aspect, which may be a source of error (Kupsc (2003, p. 17); Zangenfeind and Sonnenhauser (2014)).

Negation The Slavic double negation may cause error in translations from English (Hossain et al. (2020); Popović and Arčan (2015, p. 101)).

Cases Polish morphologically marks words by seven cases. English translations may show unnatural case distributions (Wolk and Marasek, 2019).

2.2 Dataset and preprocessing

The data was scraped from Vinted (a marketplace platform tailored towards second-hand clothing) in two locales: Polish (.pl) and English (.co.uk). Samples were translated via Google Translate.[4] Each language (pair) forms a category, resulting in 4 categories of 7,500 samples. Texts are typically short in length and 'in nature' (*e.g.*, skipping conventional words: *"Brand new boxed excellent condition"*), and are often ungrammatical, presenting an additional challenge. This allows for a realistic assessment, as it accommodates real-world use cases. Surface features were assigned using `SpaCy`.[5]

[3] `https://github.com/damiaanr/xai-srclangid`

[4] API endpoint of *Google Dictionary*. This endpoint is less accurate than the live version on translate.google.com.

[5] The `en_core_web_trf` and `pl_core_news_lg` models were used for English and Polish respectively: spacy.io.

Table 1: Entropy of conditional probability distributions of relevant PoS tags. The $\rightarrow$ symbol denotes translation. A number closer to one means a higher level of 'uncertainty', *i.e.*, a more flexible word order.

unigram	**EN**	**PL**	**PL→EN**	**EN→PL**
$t =$PROPN	.70	.78	.63	.62
$t =$NOUN	.77	.82	.69	.69
$t =$ADJ	.56	.67	.49	.60
$t =$DET	.41	.51	.33	.45
$t =$PRON	.72	.77	.59	.82
$t =$AUX	.68	.71	.70	.66
$t =$PART	.55	.78	.51	.76
$t =$X (oth.)	.56	.68	.50	.59
$t =$SCONJ	.60	.76	.56	.78

2.3 Experiment 1: Word order freedom

A measure for 'word order freedom' is computed for all four categories of samples in the dataset. Similarly to Kubon et al. (2016, p. 15) and Nikolaev et al. (2020), the scores are calculated by measuring the entropy H, here for PoS bigrams given their unigrams (*i.e.*, the entropy of the next PoS tag given a current tag; see Equation 1, where $\mathcal{T}$ is the set of all tags). The results are reported in Table 1. Tags with higher entropy in English than in Polish are excluded as these are deemed irrelevant in light of this experiment. SPACE and SYM were also omitted.

$$H(t \in \mathcal{T}) = - \sum_{t' \in \mathcal{T}} \Big(P(t, t'|t) \cdot \log_{|\mathcal{T}|} P(t, t'|t) \Big) \tag{1}$$

As expected, the results show a relative freedom of word order in Polish, while all translations seem to be less free than original texts. A plausible explanation for this phenomenon is that MT models tend to stick to fixed constructions 'that it learned to be valid,' therefore indirectly allowing less variance in word order (Bizzoni et al., 2020, p. 280). As Polish allows for a high degree of variation in word order, the translations from English are not necessarily invalid; they might just be *unnatural*—precisely what it means for a model to 'suffer' from translationese.

2.4 Experiment 2: Detecting translation

We now put forward an array of hand-crafted features, designed to capture characteristics of Polish, to train a vanilla SVM to discriminate between original and translated Polish texts. Each feature corresponds to a set of classes (listed below), the frequency counts of which are concatenated into a single vector for every sample (except for PoS-entropy classes, which are qualitative values). Each test set comprised 1,500 samples (train 6,000). The following categories of classes were considered:

Table 2: Accuracy of a linear SVM trained on features extracted from 12K Polish texts from the Vinted platform, half of them translated from English. Tested on a balanced set of 3K samples vs. a random baseline of $\frac{1}{2}$.

feature	**acc.**	**Δ baseline**	**# classes**
constit. order	.553	+.053	10
verbal aspect	.597	+.097	2
negations	.519	+.019	1
cases	.556	+.056	7
A–N order	.636	+.136	2
PoS entropy	.645	+.145	14

1. **Constituent order** Two- or three-component orderings (*e.g.*, SVO, or SV). 10 of 12 occurred.
2. **Verbal aspect** Imperfective or perfective.
3. **Negations** Contains only the word *nie*.
4. **Cases** Seven grammatical cases (*e.g.*, dative).
5. **Adjective–Noun order** Either A–N or N–A.
6. **PoS entropy** §2.3. No SYM, PUNCT, X, SPACE.

The results for each independent set of features are reported in Table 2. In part, features that may grasp more subtle 'unnaturalities' (*i.e.*, translationese) appear to outperform those that seem effective at capturing errors (*i.e.*, large semantic shifts or ungrammatical forms), indicating that translations have fewer of the latter (*e.g.*, incorrect case markers or wrong negations). This is not surprising—as MT train sets contain human translation, they inevitably exhibit translationese (Riley et al., 2020, pp. 7337–38), while presumably having few 'plain errors.'

Given the nature of the dataset and the sparse number of employed features, we judge the performance to be surprisingly well above a random baseline. A closer look at ADJ–NOUN orderings (Figure 1) shows that observations align with expectations.

3 A methodology for explainable SLI

We now put forth a two-part methodology that places an intermediary map to typological features in between the definitional map of SLI from translated text to source language. This effectively elevates the problem of SLI to a 'typology prediction-like' task that is unique in that it aims not to grasp

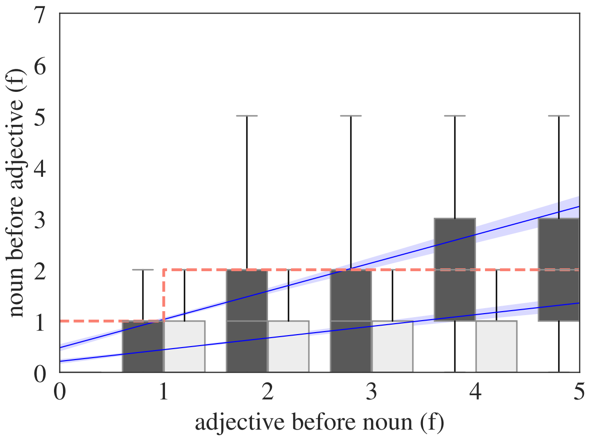

Figure 1: Density comparison of ADJ–NOUN vs. NOUN–ADJ usage in orig. Polish (dark) and translated from English (light). Samples below the discriminator (dashed red) were classified as translations. Regressors in blue.

the typology of the language in which a translated text is given, but rather to identify the typological features of the origin language of the text. The first mapping, from translated text to a prediction of the typological features of its source language, may be realised by an appeal to surface features. These lend themselves well for verifying whether the model exploiting these surface features has truly 'learned' to identify the artifacts of the typology characteristic to the source language. Moreover, surface features show promise for typology prediction in the first place, as established in the previous sections. The incentive to identify source *features* instead of source *languages* is motivated by the fact that the second mapping, from predicted typological features to a set of possible languages carrying these features, subsequently becomes a more trivial component in the pipeline that can be addressed by traditionally interpretable models, such as SVMs. The typological features then become the 'building blocks' of the explanations for predictions.

As brought up in the introduction, the choice for typological features is justified by their tendency to capture the structural elements of a source language, which are especially pronounced in machine translationese. As an additional consequence, they tend to be more robust than, *e.g.*, lexical features, for language change manifests slowest in the core structural elements of a language (Trudgill, 2020, p. 1)—precisely those grasped by high-level typological descriptions. For example, word order features are "diachronically stable" (Ponti et al., 2019, p. 579). We therefore hypothesise that SLI approaches based in typology perform more consistently across MT models, linguistic (sub)communities, and genres. Moreover, typological features naturally reflect phylogenetic relationships between languages (Berzak et al., 2014); this paves the way for 'fuzzy' classifications that align with the historical development of languages along geographical lines, as predictions may not necessarily be restricted to a single source language, but (branches of) a language family instead. They have the additional capacity to *transcend* these genealogical boundaries where overlapping typology challenges traditional linguistic classification, as is, *e.g.*, the case with Ukrainian in relation to Russian and Polish (Shevelov, 1980). This and the previous implication may especially prove beneficial in forensic contexts. Lastly, an approach that is rooted in a robust body of linguistic research offers a ground for verification of the internal reasoning of a resulting model. It moreover keeps open a dialogue between linguistics and AI: developments in linguistic typology may inform work in SLI, and possibly vice-versa.

The World Atlas of Language Structures online (WALS) appears to be a natural fit as a basis from which to draw the reasoning underlying the prediction of source languages and the corresponding explanations; it is a rich, freely available resource of typological features in a table-like format for over 2,000 languages (Dryer and Haspelmath, 2013).

A diagram of the method is given in Figure 2.

4 Discussion and conclusion

Discussion A perceived limitation of the method stems from the presumption that typology prediction is more challenging than SLI, as this could harm the performance of a model that implements the suggested methodology. However, we argue that this is only of secondary concern to a work that primarily focuses on explainability. An aim for *trustworthy* explanations requires the internal reasoning of a model to align with the reasoning conveyed in *explanations* for the model's behaviour. Our method thus needs to incorporate human-understandable concepts (*e.g.*, typological features) that are potentially less sophisticated than those developed by more 'naive', 'black-box' methods. Furthermore, although introducing typology likely complicates the task in the general case, the complexity may be reduced in a multilingual setting, for the ability to predict a fixed set of typological features provides access to a wide prediction

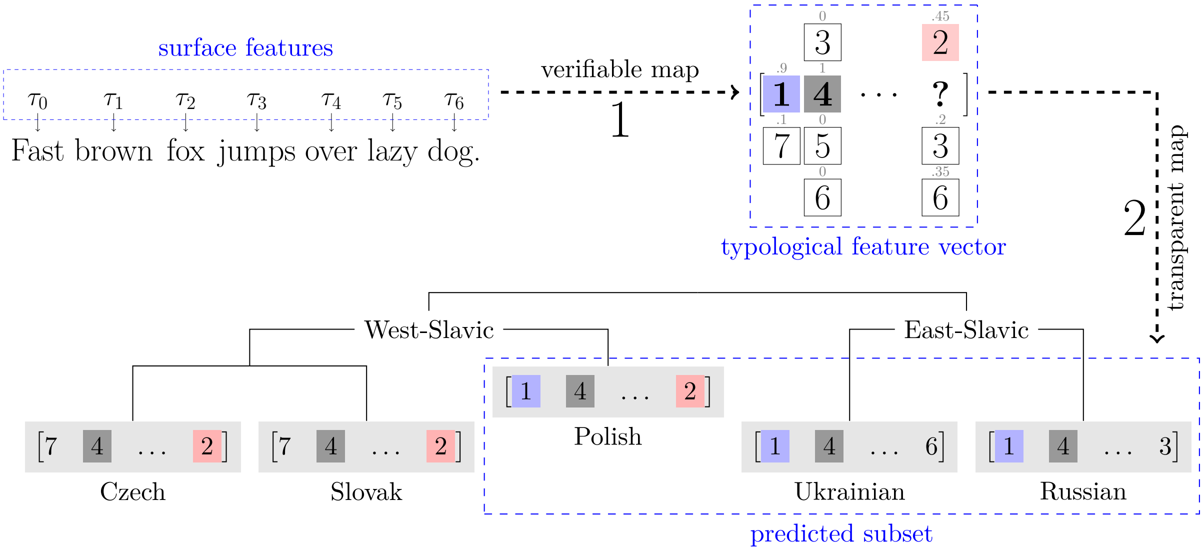

Figure 2: Intuitive diagram of a two-segment pipeline that (1) maps from surface features of a translated text to a typological feature vector w.r.t. the source language of the translated text, to (2) a (subset of) source language(s). Every element in the typological feature vector is predicted independently, resulting in a probability distribution over classes *per feature*. In this example, the source is Polish: *"Szybki brazowy lis przeskakuje nad leniwym psem."*

range of languages (namely, all that have these features set). Moreover, surface features are required only for the target language—this is assumed to always hold, as the availability of a language in MT usually means that tools to assign surface features are also available. The latter points actually testify to the assumed strengths of the method.

The experimental results in section 2 indicate that features specifically designed to accommodate *known* differences within a certain language pair may be fruitfully used for the methodology proposed in section 3. However, the method is likely to be limited by its reliance on WALS, which contains much more generally described features than those introduced in our case study (Ponti et al., 2019, p. 571). For example, Polish is classified as an 'ADJ–NOUN language' (feature 87A), placing English and Polish in the same category, while, clearly, the latter language is more permitting in its word order for adjectives and nouns, as was also observed in our case study. The broad nature of WALS may limit the ability of the method to exploit surface features in the way that was manually done in experiment 2. Moreover, it may pose an additional challenge to define a subset of WALS features that is relevant for pointing to the source language of (small) texts in the first place. Kredens et al. (2020, pp. 17–19) come to a similar conclusion about the usefulness of WALS for this task. Their 2020 paper puts forward a framework for providing different types of explanations for SLI-like tasks, with those informed by typology comprising only one tier among the multiple levels elaborated on by the authors, which ultimately lessens the effect of this issue on the overall picture, in which our methodology takes on only a part of the solution.

In conclusion, while it is impossible to make hard assertions, the experimental findings indicate promising potential for further development of the proposed methodology. A natural progression of this work is to implement the method and to qualitatively analyse its performance by evaluating it on language pairs including Slavic languages. Especially in light of the latter, future approaches may additionally consider more fine-grained differences between Slavic languages, such as tendencies for nominal or verbal constructions between Ukrainian, Polish, and Russian (Pchelintseva, 2022, p. 168).

5 Limitations

The present work was limited in that it did not assess to what extent other sources of translationese (*e.g.*, the translation model) impact the feasibility of the suggested SLI approach. The study further lacked a comparative analysis of different translation engines to test the robustness of the considered features. Moreover, although the work posits the Slavic family as a tool for evaluating explainable SLI, it did not consider in detail the appropriate procedure for conducting such evaluation.

References

Marco Baroni and Silvia Bernardini. 2005. 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.

Barbara Berti, Andrea Esuli, and Fabrizio Sebastiani. 2022. Unravelling interlanguage facts via explainable machine learning.

Yevgeni Berzak, Roi Reichart, and Boris Katz. 2014. Reconstructing native language typology from foreign language usage. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 21–29, Ann Arbor, Michigan. Association for Computational Linguistics.

Dana Bielec. 2012. *Polish: An Essential Grammar*, 2nd edition. Routledge, Oxfordshire.

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.

Edward K. Cheng. 2013. Being pragmatic about forensic linguistics. *Journal of Law and Policy*, 21(2):541–550.

Michael Daniel. 2010. 43 Linguistic Typology and the Study of Language. In *The Oxford Handbook of Linguistic Typology*. Oxford University Press.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.3)*. Zenodo.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online. Association for Computational Linguistics.

Md Mosharaf Hossain, Antonios Anastasopoulos, Eduardo Blanco, and Alexis Palmer. 2020. It's not a non-issue: Negation as a source of error in machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3869–3885, Online. Association for Computational Linguistics.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Automatically determining an anonymous author's native language. In *Intelligence and Security Informatics*, pages 209–217, Berlin, Heidelberg. Springer Berlin Heidelberg.

Krzysztof Kredens, Ria Perkins, and Tim Grant. 2020. Developing a framework for the explanation of interlingual features for native and other language infuence detection. *Language and Law/Linguagem e Direito*, 6(2):10–23.

Vladislav Kubon, Markéta Lopatková, and Tomás Hercig. 2016. Searching for a measure of word order freedom. In *Proceedings of the 16th ITAT: Slovenskočeský NLP workshop (SloNLP 2016)*, volume 1649 of *CEUR Workshop Proceedings*, pages 11–17, Bratislava, Slovakia. Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, CreateSpace Independent Publishing Platform.

Hakan Kuh. 1990. *Correlation between inflection and word order*. Ph.D. thesis, The Ohio State University.

Anna Kupsc. 2003. Two approaches to aspect assignment in an English-Polish machine translation system. In *Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resource and tools for building MT at EACL 2003*.

David Kurokawa, Cyril Goutte, and Pierre Isabelle. 2009. Automatic detection of translated text and its impact on machine translation. In *Proceedings of Machine Translation Summit XII: Papers*, Ottawa, Canada.

Massimo La Morgia, Alessandro Mei, Eugenio Nerio, and Francesco Sassi. 2023. Translated texts under the lens: From machine translation detection to source language identification. In *Proceedings of the symposium on Intelligent Data Analysis (forthcoming)*.

Eugene A. Nida and Charles R. Taber. 1969. *The theory and practice of translation.* Helps for translators. E. J. Brill, Leiden.

Dmitry Nikolaev, Taelin Karidi, Neta Kenneth, Veronika Mitnik, Lilja Saeboe, and Omri Abend. 2020. Morphosyntactic predictability of translationese. *Linguistics Vanguard*, 6(1):20190077.

Olena Pchelintseva. 2022. Aspectual properties of ukrainian verbal action nouns. *Russian Linguistics*, 46(3):167–180.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing. *Computational Linguistics*, 45(3):559–601.

Maja Popović and Mihael Arčan. 2015. Identifying main obstacles for statistical machine translation of morphologically rich South Slavic languages. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 97–104, Antalya, Turkey.

Daria Pylypenko, Kwabena Amponsah-Kaakyire, Koel Dutta Chowdhury, Josef van Genabith, and Cristina España-Bonet. 2021. Comparing feature-engineering and feature-learning approaches for multilingual translationese classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural*

Language Processing, pages 8596–8611, Online and Punta Cana, Dominican Republic. 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.

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2020. Translationese as a language in "multilingual" NMT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7737–7746, Online. Association for Computational Linguistics.

George Y. Shevelov. 1980. Ukrainian. In Alexander M. Schenker and Edward Stankiewicz, editors, *The Slavic Literary Languages: Formation and Development*. Yale Concilium on International and Area Studies, New Haven.

Anna Siewierska and Ludmila Uhlířová. 1998. *An overview of word order in Slavic languages*, pages 105–150. De Gruyter Mouton, Berlin, New York.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 48–57, Atlanta, Georgia. Association for Computational Linguistics.

Peter Trudgill. 2020. Sociolinguistic typology and the speed of linguistic change. *Journal of Historical Sociolinguistics*, 6(2):20190015.

Barak Turovsky. 2016. Ten years of Google Translate. The Keyword. Accessed on March 3, 2023.

Lucas Nunes Vieira, Minako O'Hagan, and Carol O'Sullivan. 2021. Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases. *Information, Communication & Society*, 24(11):1515–1532.

Krzysztof Wolk and Krzysztof Marasek. 2019. Survey on neural machine translation into polish. In *Multimedia and Network Information Systems*, pages 260–272, Cham. Springer International Publishing.

Gu Xinxing. 2023. New features make Translate more accessible for its 1 billion users. The Keyword. Accessed on March 3, 2023.

Robert Zangenfeind and Barbara Sonnenhauser. 2014. Russian verbal aspect and machine translation. In *Dialog-21*, pages 743–752. s.n.

JiaJun Zhang and ChengQing Zong. 2020. Neural machine translation: Challenges, progress and future. *Science China Technological Sciences*, 63(10):2028–2050.