GTNC: A Many-To-One Dataset of Google Translations from NewsCrawl

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

Charlotte Pouw
ILLC, University of Amsterdam
c.m.pouw@uva.nl

Abstract

This paper lays the groundwork for initiating research into Source Language Identification; the task of identifying the original language of a machine-translated text. We contribute a carefully-crafted dataset of translations from a typologically diverse spectrum of languages into English and use it to set initial baselines for this novel task. The dataset is publicly available on our GitHub repository: damiaanr/gtnc.

1 Introduction

In an era of globalisation, the world is becoming increasingly reliant on machine translation. But as translation tools find their way into people's daily routines, they spark curiosity about previously unexplored tasks, such as identifying the source language of a machine-translated text. This is an emerging challenge that has been referred to as Source Language Identification (SLI, La Morgia et al. 2023). The task has a relevant application in forensics: knowledge of an individual's native language can offer crucial insights into their identity.

The problem of classifying the original language of a machine-translated text inherently relies on finding markers in the translation that hint at the source (*i.e.*, traces of 'source language interference'). In a first exploration of the field, Reijnaers and Herrewijnen (2023) indicated that such markers can be related to typological differences between the languages involved in the translation process, aligning with theory on human translation (Teich, 2003, pp. 217–20). Typological features contribute to the explainability of SLI models (Kredens et al., 2020, pp. 17–19), a quality essential in forensic contexts (Cheng, 2013, pp. 547–49). However, owing to the novelty of the task, research on SLI is hindered by a lack of sufficiently sized datasets that contain machine translations from a large number of languages into a single language.

This work aims to fill this gap to propel this emerging area of research forward. We introduce **Google Translations from NewsCrawl (GTNC)**: a unique dataset of state-of-the-art machine translations from a diverse set of languages into English, offering a rich typological diversity to facilitate experiments with a wide range of source languages. The dataset spans **50 languages** (listed below), contains **7,500 sentences** per language, and is representative of real-world data given its domain (news articles) and the translation engine used (Google Translate). In addition, we offer initial baselines for future work on SLI and thereby confirm the feasibility of the task.

The next section of this paper will discuss existing datasets that may be used for SLI. In addressing their limitations, we propose a novel dataset in Section 3, which we will then use in a series of experiments in the section that follows. The findings reiterate the value of a typological approach in SLI.

Included languages Amharic, Arabic, Bengali, Bulgarian, Chinese, Croatian, Czech, Dutch, English (untranslated), Estonian, Finnish, French, German, Greek, Gujarati, Hausa, Hindi, Hungarian, Icelandic, Igbo, Indonesian, Italian, Japanese, Kannada, Korean, Kyrgyz, Latvian, Lithuanian, Macedonian, Malayalam, Marathi, Odia, Oromo, Pashto, Persian, Polish, Portuguese, Punjabi, Romanian, Russian, Shona, Spanish, Swahili, Tagalog, Tamil, Telugu, Tigrinya, Turkish, Ukrainian, and Yoruba.

2 Existing datasets

In the realm of *human* translation, several corpora exist that contain translations from multiple languages into a single language, among which the most popular is a collection of proceedings of the European Parliament (Europarl, Koehn 2005). Numerous studies have leveraged this corpus to provide empirical evidence for distinctions between original and translated texts (Koppel and Ordan

Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2024), pages 58–65

2011; Rabinovich and Wintner 2015; Volansky et al. 2013), while some have explicitly aimed to identify the European source language of these documents (Rabinovich et al. 2017; van Halteren 2008). However, *machine* translations are divergent from human translations in a systematic (Fu and Nederhof, 2021) and measurable (van der Werff et al., 2022) way: machine translations often exhibit less morphological and lexical diversity (Vanmassenhove et al., 2021) and adhere more closely to the structure of the source text (Ahrenberg, 2017). Moreover, machine translations are more susceptible to source language interference (Toral, 2019, p. 279), particularly concerning the structural properties of the source language (Bizzoni et al. 2020, p. 288; Popovic et al. 2023). As such, a dataset of purely machine translations is desirable.

A handful of datasets exist that contain machine translations from multiple languages into one. An example is DEMETR (Karpinska et al., 2022), consisting of translations from ten, predominantly Indo-European languages[1] into English. The dataset was constructed to aid models in detecting errors in machine translation output. As a result, a downside is that the authors post-edited the translations to ensure their correctness, thereby potentially eliminating valuable hints that pointed to the source language of these texts. DEMETR is also modest in size, comprising only 100 sentences per language.

Another example is MLQE-PE (Fomicheva et al., 2022), containing the translations of 9,000 sentences for each of five, diverse Indo-European languages[2] into English. Apart from the small number of classes, a drawback of this dataset is that the samples vary widely in length across languages (Figure 2a) and are often noisy (*e.g.*, containing URLs or HTML tags). This could potentially bias SLI models towards relying on spurious features instead of learning linguistic patterns purely governed by typology.

A comparison of the key features mentioned above can be found in Table 1. Notably, all of the above-cited datasets were created to *evaluate* machine translation models. The mentioned limitations thus only come to light when analysing their usefulness for SLI, highlighting the need for a dataset crafted specifically for the task. In the next section, we will introduce such a dataset.

[1]DEMETR includes Chinese, Czech, French, German, Hindi, Italian, Japanese, Polish, Russian, and Spanish.

[2]The relevant languages in MLQE-PE include Estonian, Nepali, Romanian, Russian, and Sinhala.

Table 1: Comparison of dataset size characteristics for usage in a *many-to-one* context.

Dataset	# languages	# sentences/lang.
DEMETR	10	100
MLQE-PE	5	9,000
GTNC	50	7,500

3 A dataset for SLI

In the subsections below, we will describe the steps taken to build GTNC and will provide analyses into its diversity and characteristics. The data and all code used to generate them is available on GitHub.

3.1 Selecting the source texts

To enable a fair comparison of translations across languages, we would ideally obtain a collection of parallel source texts. Yet, while creating a 'one-to-many' corpus is relatively straightforward, building a *many-to-one* variant is practically impossible—it would require the spontaneous utterance of identical content in each language. We therefore aimed to make the data *as parallel as possible*.

On the presumption that the news genre is both universal and relatively consistent worldwide, we selected NewsCrawl (Kocmi et al., 2022) as the repository for the source texts as it contains sentences scraped from news articles in 59 languages and is 'parallel by year'. For GTNC, we sampled from articles that appeared in 2020, '21, and '22, with equal proportions of each year per language.

To enable analyses on the typological level, beyond the prediction of individual language labels, we aimed to include a large number of languages in GTNC. Ultimately, we selected 50 languages.[3] English sentences were naturally left untranslated, allowing for experiments with both translated and original texts. Figure 1 illustrates the data's diversity, based on the World Atlas of Language Structures (WALS, Dryer and Haspelmath 2013). WALS is a resource (wals.info) that contains typological features for over 2,000 languages in tabular format.

3.2 Filtering the samples

The source sentences from NewsCrawl are already shuffled and duplicate-free. We additionally re-

[3]We excluded 9 languages for the following reasons: noise (Kinyarwanda and Somali), similarity to other languages in light of dataset diversity (Bosnian, Serbian, Kazakh), lack of available WALS features to enable effective typological analyses (Afrikaans), lack of data (Tigre and Bambara), and incompatibility with Google Translate (South Ndebele).

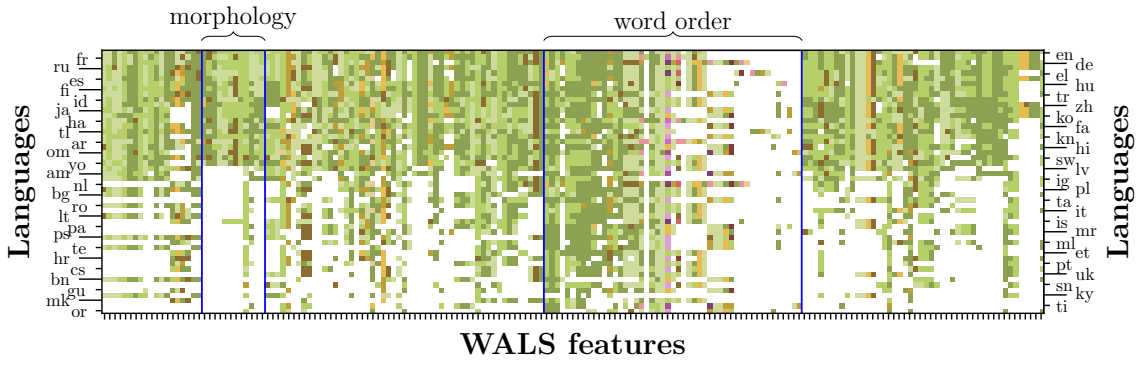


Figure 1: Visualisation of language diversity in GTNC. Columns represent typological features and rows correspond to languages (tagged by ISO 639-1 codes). Hues in columns denote different classes for each feature (overlap in colour thus hints at languages being similar). White boxes indicate unset features. Blue lines act as indicators of feature clusters, of which two are exemplified. Note that each row can intuitively be viewed as a language's unique 'signature', with SLI involving the identification of these signatures through the artifacts they leave in a translation.

moved sentences that contained either of:

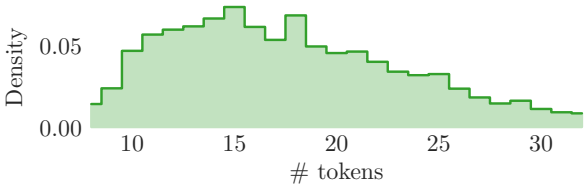
- A total of < 30 or > 400 characters.
- Non-alphanumerics, excluding `;:()!?` and equivalences in other languages.
- Characters that directly followed a period (`.`) and were not a white space, a digit, a question mark, another period, or an exclamation mark.
- Four consecutive, identical characters.
- Not `.!?` or equivalences in other languages as the last character of the sentence.
- Latin alphabet characters for non-Latin-script languages and *vice-versa*.
- Russian: Ukrainian-specific characters (as the Russian corpus also contained Ukrainian text).

Furthermore, samples that were deemed 'short' or 'bad' by JusText (Pomikálek, 2011) were also left out.[4] JusText is a tool for removing boilerplate content (*i.e.*, frequently-used and non-unique text).

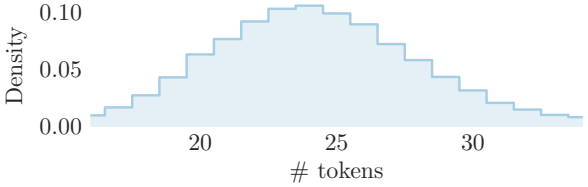
3.3 Translating and aligning by length

The samples were obtained by using Google Translate.[5] To avoid a spurious correlation between sentence length and language class—which an SLI-model could potentially exploit—we aimed at maintaining a consistent average and median length of 125 characters across all resulting English translations. This was accomplished by selecting sentences of specific lengths, determined by pre-computed, frequentist character-to-character ratios for every translation pair. Ultimately, $42,667,664$ source characters were translated into $46,460,290$ English characters ($\mu \approx 126.42$ characters per sample; not including the original English sentences from NewsCrawl). The data over all languages is normal (Figure 2b). As the resulting ratios might be of interest to other studies in machine translation, we included them as appendix material in Table 2.

Finally, all translated samples were scored by Monocleaner (Sánchez-Cartagena et al., 2018) to denote their 'fluency'. These annotations are essentially language-model scores, calculated as the normalised perplexity of character 7-grams.



(a) MLQE-PE: Separate length distributions per class.



(b) GTNC: Normally distributed length across all classes.

Figure 2: Sample length across many-to-one datasets.

[4]This step was not available for Amharic, Hausa, Japanese, Oromo, Odia, Punjabi, Pashto, Shona, Tigrinya, and Chinese.

[5]The data were translated on June 20th, 2023, using the v3 Translation API. To support the creation of this dataset, Google granted the equivalent of USD $1,000 in API credits.

4 Preliminary Experiments

In this section, we will use the English translations from GTNC to predict the source language of both individual samples and combinations of sentences.

4.1 Input representation

It is through parts of speech that hints about a language's structure—and thereby its typology—may be obtained (Cutting et al., 1992, p. 133). Given the structural nature of source language interference in machine translation, we opted for part of speech (PoS) tags as input features for our models.

When discerning between human-translated and original texts, many studies have achieved good performance by representing input texts as sequences of PoS tags; generally by training an SVM (Hearst et al., 1998) on frequency counts of PoS n-grams (Baroni and Bernardini 2005, p. 268; Rabinovich et al. 2017, p. 534; Pylypenko et al. 2021, p. 8603). In a recent study, Popovic et al. (2023) did so for *machine* translations and likewise indicated the efficacy of PoS tags, affirming their relevance in SLI.

Of the above, the work closest to ours is the paper by Rabinovich et al. (2017). Its authors perform source language identification on human-translated texts and report an accuracy of 75.61% when considering samples from 14 source languages.

4.2 Model architecture and training

To enable the model to capture structural patterns over longer distances, we conducted our experiments using a bidirectional LSTM (Graves and Schmidhuber, 2005) of 64 hidden dimensions. In correspondence with the granularity of the data, the LSTM operates on the sentence level, classifying one sentence at a time. At every timestep, it takes in a 70-dimensional input vector, consisting of a one-hot encoding of a token's PoS tag (fine-grained), concatenated with a multi-hot vector that additionally encodes grammatical number, tense, and person, pronominal type, definiteness, verb form, and whether a word is possessive and/or reflexive. All PoS and morphological tags were assigned by SpaCy (spacy.io) and adhere to the Universal Dependencies standard (Nivre et al., 2017).

The final hidden state of the LSTM is concatenated for every direction and subsequently processed by a single feed-forward layer that directly maps to output classes (*i.e.*, possible source languages). To obtain a prediction for a *group* of sentences, the logits of this layer are averaged over all individually classified samples within the group.

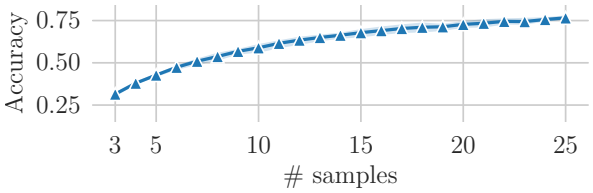


Figure 3: Performance over number of sentences.

We trained all LSTMs for 20 epochs and averaged 'best epoch'-results over three runs. Optimisation was done using Adam (Kingma and Ba, 2017) with a learning rate of 1e-3, a weight decay parameter of 1e-4, and a batch size of 16. Data was split into train and test fractions of 0.9 and 0.1 respectively.

4.3 Initial baseline results for SLI

Figure 3 illustrates the model's accuracy as a function of the number of sentences.[6] Individual samples were classified with an accuracy of 15.68%. Note that we work with samples of only a few tens of tokens in size (Figure 2b), while Rabinovich et al. (2017) use samples of 1,000 tokens. Naturally, the longer the document, the more opportunity the source language has to leave its fingerprints. The positive correlation between document length and accuracy, as shown in Figure 3, provides evidence that supports this tendency.

Inspired by the same work, we reconstructed a phylogenetic tree using hierarchical agglomerative clustering applied to the averaged confusion scores for all Indo-European languages in GTNC. The tree is shown in Figure 4. 'Ward's method' was used as linkage criterion (Ward, 1963). The model was trained only on the 24 Indo-European source languages present in GTNC (excluding English). The tree provides intuitive evidence that the model tends to confuse genetically similar languages, indicating that the model exploits language-specific patterns that align with their typology. This, in turn, implies that the sentences in GTNC do indeed carry typological features of their source counterparts, rendering it a well-suited dataset for SLI. The tree-figure additionally provides insight into the kind of errors that the model makes. For example, when contrasted with the frequently referenced 'gold tree' by Serva, M. and Petroni, F. (2008), Greek is being misclassified as being part of a branch with

[6] Ideally, the samples would have appeared in natural sequence, however, due to lack of data, they were drawn *i.i.d.*

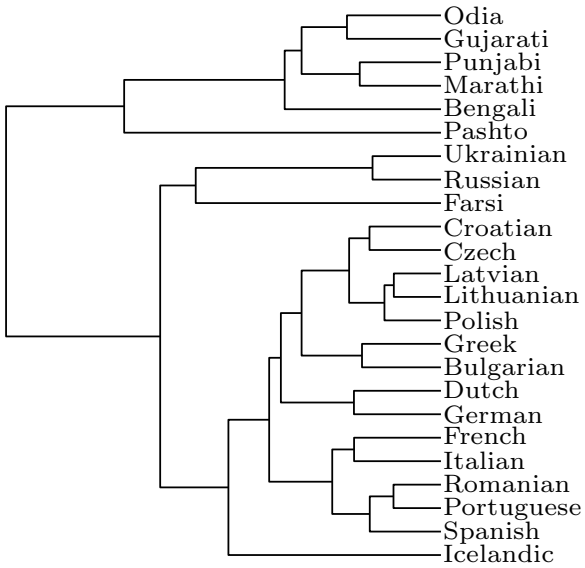


Figure 4: Reconstructed phylogenetic tree.

Bulgarian, among other Slavic languages in higher branches. This suggests that the model mistakenly relies on similar markers to discern between English translations from Greek and those from Slavic languages. A logical next step for future research would therefore involve a detailed analysis of the specific markers employed by such models. Based on the figure, a similar argument could be made for Farsi, or the East Slavic sub-branch.

5 Conclusion and discussion

We showcased GTNC: a thoughtfully designed dataset of Google-Translated news articles from diverse languages into English. Our experiments provide compelling evidence attesting to the feasibility of SLI and emphasise the dataset's suitability for typological approaches—a quality that holds significant promise on the path to *explainable* SLI.

As our goal was to introduce a dataset, we deliberately avoided a lengthy discussion on the underlying phenomena that enable the identification of a source language in the first place; *i.e.*, a more in-depth analysis of how 'artifacts' of the typology of the original language are left behind in a translation. An exploration of the involved scientific concepts, particularly 'translationese' (Nida and Taber, 1969) and 'interlanguage' (Selinker, 1972), and how they relate to machine translation, would demand a thorough examination that is beyond the scope of this short paper.

While GTNC encompasses a wide array of languages, the number of samples per language remains limited. We encourage the community to improve our dataset using the tools that we have made available. Beyond SLI, the data may also help other applications, such as evaluating Google Translate's performance across languages. We hope that GTNC will additionally foster exploration in new directions.

6 Acknowledgements

The authors are grateful for the useful feedback provided by Ella Rabinovich, Wilker Aziz, Willem Zuidema, and the anonymous reviewers. Charlotte Pouw's research is funded by the NWA-ORC grant 'InDeep', no. 1292.19.399, provided by the Netherlands Organisation for Scientific Research.

References

Lars Ahrenberg. 2017. Comparing machine translation and human translation: A case study. In *Proceedings of the Workshop Human-Informed Translation and Interpreting Technology*, pages 21–28, Varna, Bulgaria. Association for Computational Linguistics, Shoumen, Bulgaria.

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.

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.

Doug Cutting, Julian Kupiec, Jan Pedersen, and Penelope Sibun. 1992. A practical part-of-speech tagger. In *Third Conference on Applied Natural Language Processing*, pages 133–140, Trento, Italy. Association for Computational Linguistics.

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

Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2022. MLQE-PE: A multilingual quality estimation and post-editing dataset. In *Proceedings of the Thirteenth Language Resources*

and Evaluation Conference, pages 4963–4974, Marseille, France. European Language Resources Association.

Yingxue Fu and Mark-Jan Nederhof. 2021. Automatic classification of human translation and machine translation: A study from the perspective of lexical diversity. In *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pages 91–99, online. Association for Computational Linguistics.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610. IJCNN 2005.

M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28.

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022. DEMETR: Diagnosing evaluation metrics for translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9540–9561, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA. Association for Computational Linguistics.

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.

Massimo La Morgia, Alessandro Mei, Eugenio Nerio Nemmi, Luca Sabatini, and Francesco Sassi. 2023. Translated texts under the lens: From machine translation detection to source language identification. In *Advances in Intelligent Data Analysis XXI*, pages 222–235, Cham. Springer Nature Switzerland.

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

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Jan Pomikálek. 2011. *Removing boilerplate and duplicate content from web corpora*. Ph.D. thesis, Masaryk university, Faculty of informatics, Brno, Czech Republic.

Maja Popovic, Ekaterina Lapshinova-Koltunski, and Maarit Koponen. 2023. Computational analysis of different translations: by professionals, students and machines. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 365–374, Tampere, Finland. European Association for Machine Translation.

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.

Ella Rabinovich and Shuly Wintner. 2015. Unsupervised identification of translationese. *Transactions of the Association for Computational Linguistics*, 3:419–432.

Damiaan Reijnaers and Elize Herrewijnen. 2023. Machine-translated texts from English to Polish show a potential for typological explanations in source language identification. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 40–46, Dubrovnik, Croatia. Association for Computational Linguistics.

Víctor M. Sánchez-Cartagena, Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez-Sánchez. 2018. Prompsit's submission to wmt 2018 parallel corpus filtering shared task. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, Brussels, Belgium. Association for Computational Linguistics.

Larry Selinker. 1972. Interlanguage. 10(1-4):209–232.

Serva, M. and Petroni, F. 2008. Indo-european languages tree by levenshtein distance. *EPL*, 81(6):68005.

Elke Teich. 2003. *Cross-Linguistic Variation in System and Text*. De Gruyter Mouton, Berlin, Boston.

Antonio Toral. 2019. Post-editese: an exacerbated translationese. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 273–281, Dublin, Ireland. European Association for Machine Translation.

Tobias van der Werff, Rik van Noord, and Antonio Toral. 2022. Automatic discrimination of human and neural machine translation: A study with multiple pretrained models and longer context. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 161–170, Ghent, Belgium. European Association for Machine Translation.

Hans van Halteren. 2008. Source language markers in EUROPARL translations. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 937–944, Manchester, UK. Coling 2008 Organizing Committee.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. Association for Computational Linguistics.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2013. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Joe H. Ward. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

A Character-to-character ratios

See Table 2 on the next page.

Table 2: Character-to-character ratios of languages in GTNC, relative to English. $n = 100$.

	Source = 100 chrs.				Target = 125 chrs.	
	Length		**Ratio**		**Realised length**	
	μ	σ/μ	$\rightarrow$	$\leftarrow$	μ	σ/μ
Amharic (am)	155.7	.16	1.56	0.64	**124.7**	.15
Arabic (ar)	129.9	.15	1.30	0.77	**128.1**	.15
Bengali (bn)	111.4	.14	1.11	0.90	**127.5**	.14
Bulgarian (bg)	103.5	.12	1.04	0.97	**124.8**	.11
Chinese (zh)	421.7	.16	4.22	0.24	**123.4**	.19
Croatian (hr)	111.1	.12	1.11	0.90	**122.6**	.11
Czech (cs)	112.9	.15	1.13	0.89	**126.3**	.12
Dutch (nl)	94.7	.11	0.95	1.06	**125.7**	.10
English (en)	100.0	⊥	1.00	1.00	**125.0**	⊥
Estonian (et)	111.0	.15	1.11	0.90	**123.9**	.12
Finnish (fi)	104.6	.12	1.05	0.96	**125.1**	.12
French (fr)	92.0	.11	0.92	1.09	**125.9**	.10
German (de)	93.6	.11	0.94	1.07	**126.3**	.11
Greek (el)	91.9	.13	0.92	1.09	**126.1**	.11
Gujarati (gu)	108.1	.16	1.08	0.92	**127.5**	.14
Hausa (ha)	100.1	.16	1.00	1.00	**123.4**	.14
Hindi (hi)	117.1	.16	1.17	0.85	**122.2**	.13
Hungarian (hu)	106.8	.12	1.07	0.94	**122.6**	.12
Icelandic (is)	103.1	.12	1.03	0.97	**126.8**	.12
Igbo (ig)	109.3	.14	1.09	0.92	**121.4**	.16
Indonesian (id)	100.0	.14	1.00	1.00	**125.3**	.13
Italian (it)	97.1	.12	0.97	1.03	**125.1**	.10
Japanese (ja)	236.9	.28	2.37	0.42	**142.9**	.23
Kannada (kn)	102.3	.17	1.02	0.98	**126.8**	.15
Korean (ko)	229.0	.24	2.29	0.44	**170.9**	.18
Kyrgyz (ky)	101.9	.15	1.02	0.98	**126.5**	.14
Latvian (lv)	110.1	.11	1.10	0.91	**124.9**	.12
Lithuanian (lt)	107.9	.14	1.08	0.93	**125.3**	.12
Macedonian (mk)	100.2	.11	1.00	1.00	**127.1**	.11
Malayalam (ml)	87.6	.17	0.88	1.14	**129.5**	.14
Marathi (mr)	103.1	.13	1.03	0.97	**122.5**	.14
Odia (or)	106.6	.16	1.07	0.94	**123.6**	.14
Oromo (om)	82.3	.17	0.82	1.22	**121.0**	.17
Pashto (ps)	114.7	.13	1.15	0.87	**127.7**	.13
Persian (fa)	119.5	.13	1.19	0.84	**127.4**	.15
Polish (pl)	102.8	.13	1.03	0.97	**124.5**	.12
Portuguese (pt)	100.6	.11	1.01	0.99	**122.4**	.10
Punjabi (pa)	102.2	.14	1.01	0.99	**125.0**	.13
Romanian (ro)	97.3	.12	0.97	1.03	**124.7**	.11
Russian (ru)	106.7	.14	1.07	0.94	**125.5**	.13
Shona (sn)	100.5	.14	1.01	0.99	**122.7**	.14
Spanish (es)	95.2	.11	0.95	1.05	**125.4**	.10
Swahili (sw)	100.5	.14	1.00	1.00	**124.1**	.13
Tagalog (tl)	90.6	.12	0.91	1.10	**127.0**	.11
Tamil (ta)	89.0	.18	0.89	1.12	**125.8**	.15
Telugu (te)	105.9	.16	1.06	0.94	**123.8**	.14
Tigrinya (ti)	130.9	.20	1.31	0.76	**127.9**	.18
Turkish (tr)	104.7	.17	1.05	0.96	**127.0**	.13
Ukrainian (uk)	111.0	.13	1.11	0.90	**123.3**	.13
Yoruba (yo)	107.8	.16	1.08	0.93	**124.8**	.14