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

Damiaan J. W. Reijnaers[1] Charlotte Pouw[2]

[1]University of Amsterdam

[2]ILLC, University of Amsterdam

The 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (**SIGTYP**)

Table of Contents

GTNC is a ‘many-to-one’ dataset

- Contains translations from **50 languages** into English

GTNC is a 'many-to-one' dataset

- Contains translations from **50 languages** into English

GTNC is a 'many-to-one' dataset

- Contains translations from **50 languages** into English *(including English original texts)*

GTNC is a 'many-to-one' dataset

- Contains translations from **50 languages** into English
- Great typological diversity

GTNC is a ‘many-to-one’ dataset

- Contains translations from **50 languages** into English
- Great typological diversity

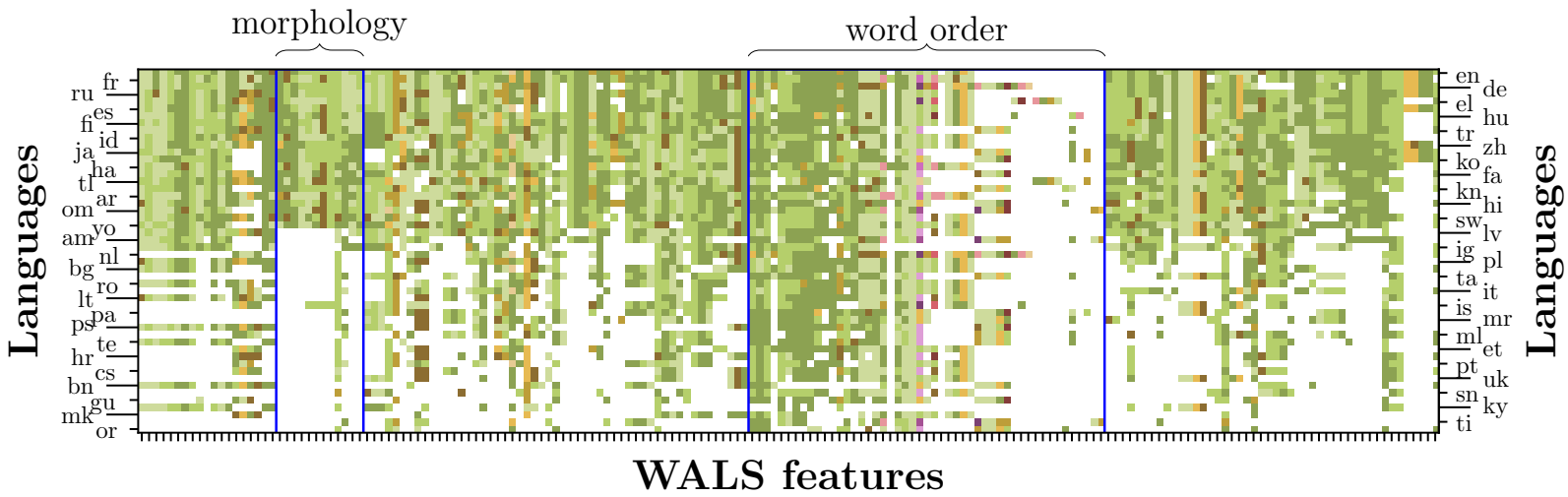

Figure: Visualisation of language diversity in GTNC.

General dataset characteristics

- Source data originates from **NewsCrawl**[1]

[1] Tom Kocmi et al. "Findings of the 2022 Conference on Machine Translation (WMT22)". In: *Proceedings of the Seventh Conference on Machine Translation (WMT)*. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 1–45. URL: https://aclanthology.org/2022.wmt-1.1.

General dataset characteristics

- Source data originates from **NewsCrawl**[1]
- Translated by a recent version of **Google Translate**

[1] Kocmi et al., "Findings of the 2022 Conference on Machine Translation (WMT22)".

General dataset characteristics

- Source data originates from **NewsCrawl**[1]
- Translated by a recent version of **Google Translate**
- **7,500 sentences** *per* language

[1] Kocmi et al., "Findings of the 2022 Conference on Machine Translation (WMT22)".

General dataset characteristics

- Source data originates from **NewsCrawl**[1]
- Translated by a recent version of **Google Translate**
- **7,500 sentences** *per* language
- $\approx$ 125 characters per sentence

[1] Kocmi et al., "Findings of the 2022 Conference on Machine Translation (WMT22)".

General dataset characteristics

- Source data originates from **NewsCrawl**[1]
- Translated by a recent version of **Google Translate**
- **7,500 sentences** *per* language
- $\approx$ 125 characters per sentence

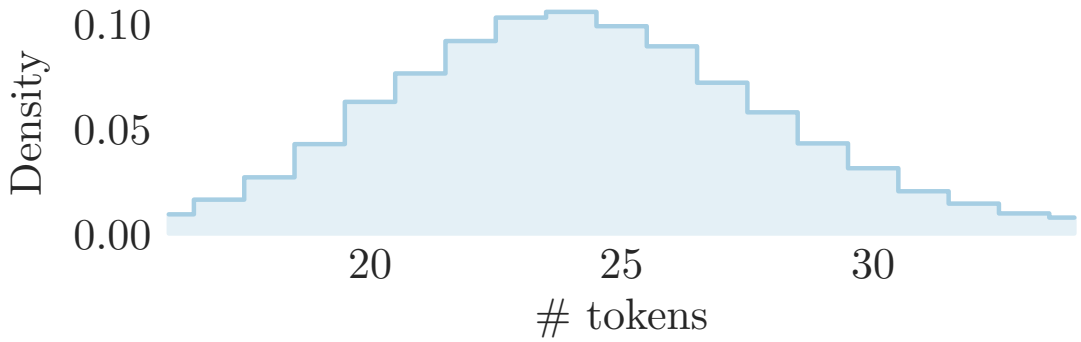

Figure: Normally distributed length across all classes.

[1] Kocmi et al., "Findings of the 2022 Conference on Machine Translation (WMT22)".

Table of Contents

Table of Contents

Source Language Identification: a novel task

Source Language Identification (SLI)

The task of inferring the origin language of machine-translated texts *using only the translated text*.

- First mentioned by La Morgia et al. (2023).[2]

[2] Massimo La Morgia et al. "Translated Texts Under the Lens: From Machine Translation Detection to Source Language Identification". In: *Advances in Intelligent Data Analysis XXI*. ed. by Bruno Crémilleux, Sibylle Hess, and Siegfried Nijssen. Cham: Springer Nature Switzerland, 2023, pp. 222–235.

Source Language Identification: a novel task

Source Language Identification (SLI)

The task of inferring the origin language of machine-translated texts *using only the translated text.*

- First mentioned by La Morgia et al. (2023).[2]
- Relevant in **forensics**: *knowledge of an individual's native language can offer crucial insights into their identity.*

[2] La Morgia et al., "Translated Texts Under the Lens: From Machine Translation Detection to Source Language Identification".

Source Language Identification: a novel task

Source Language Identification (SLI)

The task of inferring the origin language of machine-translated texts *using only the translated text.*

- First mentioned by La Morgia et al. (2023).[2]
- Relevant in **forensics**: *knowledge of an individual's native language can offer crucial insights into their identity.*
- Inherently relies on finding markers in the translation that hint at the source.

[2] La Morgia et al., "Translated Texts Under the Lens: From Machine Translation Detection to Source Language Identification".

About those "markers that hint at the source."

- Such markers can be related to **typological differences** between the languages involved in the translation process (Reijnaers and Herrewijnen 2023).[3]

[3] Damiaan Reijnaers and Elize Herrewijnen. "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)*. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 40–46. URL: https://aclanthology.org/2023.bsnlp-1.6.

About those "markers that hint at the source."

- Such markers can be related to **typological differences** between the languages involved in the translation process (Reijnaers and Herrewijnen 2023).[3]
  - Aligns with theory on human translation!

[3] Reijnaers and Herrewijnen, "Machine-translated texts from English to Polish show a potential for typological explanations in Source Language Identification".

About those "markers that hint at the source."

- Such markers can be related to **typological differences** between the languages involved in the translation process (Reijnaers and Herrewijnen 2023).[3]
- Classifying in terms of **typological features** contributes to the explainability of SLI models.

[3] Reijnaers and Herrewijnen, "Machine-translated texts from English to Polish show a potential for typological explanations in Source Language Identification".

About those "markers that hint at the source."

- Such markers can be related to **typological differences** between the languages involved in the translation process (Reijnaers and Herrewijnen 2023).[3]
- Classifying in terms of **typological features** contributes to the explainability of SLI models.
  - A quality essential in forensic contexts!

[3] Reijnaers and Herrewijnen, "Machine-translated texts from English to Polish show a potential for typological explanations in Source Language Identification".

About those "markers that hint at the source."

- Such markers can be related to **typological differences** between the languages involved in the translation process (Reijnaers and Herrewijnen 2023).[3]
- Classifying in terms of **typological features** contributes to the explainability of SLI models.
- GTNC is optimised for typology-oriented approaches.

[3] Reijnaers and Herrewijnen, "Machine-translated texts from English to Polish show a potential for typological explanations in Source Language Identification".

Table of Contents

Teaser!

Read our paper *or listen in on March 22nd*

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