LREC 2026
Palma de Mallorca, Spain

The 19th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC 2026

Workshop Proceedings

Editors
Reinhard Rapp, Ayla Rigouts Terryn,
Serge Sharoff, Pierre Zweigenbaum

11 May 2026

Proceedings of the 19th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC 2026

ISBN 978-2-493814-75-3

Preface

The 19th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC 2026

In the language engineering and linguistics communities, research in comparable corpora has been motivated by two main reasons. In language engineering, on the one hand, it is chiefly motivated by the need to use comparable corpora as training data for data-driven NLP applications such as statistical and neural machine translation, or cross-lingual retrieval. In linguistics, on the other hand, comparable corpora are of interest because they enable cross-language discoveries and comparisons. It is generally accepted in both communities that comparable corpora consist of documents that are comparable in content and form in various degrees and dimensions across several languages. Parallel corpora are on the one end of this spectrum, and unrelated corpora are on the other. Increasingly, these resources are not only collected, but also augmented or even created synthetically, which raises new questions about how to define and measure comparability.

In recent years, the use of comparable corpora for pre-training Large Language Models (LLMs) has led to their impressive multilingual and cross-lingual abilities, which are relevant to a range of applications, including information retrieval, machine translation, cross-lingual text classification, etc. The linguistic definitions and observations related to comparable corpora are crucial to improve methods to mine such corpora, to assess and document synthetic data, and to improve cross-lingual transfer of LLMs. Therefore, it is of great interest to bring together builders and users of such corpora. The aim of the workshop series on "Building and Using Comparable Corpora" (BUCC) is to promote progress in this field.

The previous editions of the workshop took place in Africa (LREC 2008 in Marrakech), America (ACL 2011 in Portland and ACL 2017 in Vancouver), Asia (ACL-IJCNLP 2009 in Singapore, ACL-IJCNLP 2015 in Beijing, LREC 2018 in Miyazaki, Japan, COLING 2025 in Abu Dhabi), Europe (LREC 2010 in Malta, ACL 2013 in Sofia, LREC 2014 in Reykjavik, LREC 2016 in Portoroz, RANLP 2019 and RANLP 2023 in Varna, LREC 2022 in Marseille, LREC-COLING 2024 in Torino), and also on the border between Asia and Europe (LREC 2012 in Istanbul). Due to the Corona crisis, the workshop was also held online in conjunction with LREC 2020 and RANLP 2021. The materials of the past workshops and related studies have also been summarised in a recent textbook from Springer:
`https://link.springer.com/book/10.1007/978-3-031-31384-4`.

We want to thank all the people who, in one way or another, helped make this workshop once again a success, especially the LREC workshop chairs.

Our special thanks go to our invited speaker, Els Lefever, and to the members of the program committee, who did a great job in reviewing the submitted papers under strict time constraints. Last but not least, we would like to thank the authors and all workshop participants.

Reinhard Rapp, Ayla Rigouts Terryn, Serge Sharoff, Pierre Zweigenbaum May 2026

Organizing Committee

Workshop Chairs

Reinhard Rapp (University of Mainz, Germany)

Ayla Rigouts Terryn (Université de Montréal & Mila, Canada)

Serge Sharoff (University of Leeds, United Kingdom)

Pierre Zweigenbaum (Université Paris-Saclay, CNRS, France)

Program Committee

- Ebrahim Ansari (Institute for Advanced Studies in Basic Sciences, Iran)
- Eleftherios Avramidis (DFKI, Germany)
- Gabriel Bernier-Colborne (National Research Council, Canada)
- Kenneth Church (VecML.com, USA)
- Patrick Drouin (Université de Montréal, Canada)
- Alex Fraser (Technical University of Munich, Germany)
- Natalia Grabar (CNRS, University of Lille, France)
- Amal Haddad Haddad (Universidad de Granada, Spain)
- Kyo Kageura (University of Tokyo, Japan)
- Natalie Kübler (Université Paris Cité, France)
- Philippe Langlais (Université de Montréal, Canada)
- Yves Lepage (Waseda University, Japan)
- Shervin Malmasi (Amazon, USA)
- Michael Mohler (Language Computer Corporation, USA)
- Emmanuel Morin (Nantes Université, France)
- Dragos Stefan Munteanu (RWS, USA)
- Preslav Nakov (Mohamed bin Zayed University of AI (MBZUAI), United Arab Emirates)
- Ted Pedersen (University of Minnesota, Duluth, USA)
- Reinhard Rapp (University of Mainz, Germany)
- Ayla Rigouts Terryn (Université de Montréal & Mila, Canada)
- Nasredine Semmar (CEA-List, France)
- Serge Sharoff (University of Leeds, United Kingdom)
- Richard Sproat (Sakana.ai, Tokyo, Japan)
- Marko Tadić (University of Zagreb, Croatia)
- François Yvon (CNRS & Sorbonne Université, France)
- Pierre Zweigenbaum (Université Paris-Saclay, CNRS, France)

Table of Contents

Workshop Program

Monday, May 11, 2026

9:00 **Session 1**
Chair: Ayla Rigouts Terryn

9:00 ***Introduction***

9:06 *Keynote: The Cross-Lingual Transfer Myth: Why Modern LLMs Still Fail Without Comparable Corpora and Representations*
Els Lefever

10:06 *A Comparative Study of Parkinsonian Speech Corpora for Deep Learning-Based Detection of Dysarthria*
Clara Ponchard and Pierre Serrano

10:30 **Coffee break**

11:00 **Session 2: Comparable corpora for linguistics research**
Chair: Philippe Langlais

11:00 *Computing Semantic Similarity for Aligning Bilingual Semi-parallel Texts: A Case Study*
Steffen Frenzel, Maximilian Krupop and Manfred Stede

11:24 *A Comparative Study in Corpus Linguistics Applied to Automatic Terminology Extraction*
Mercè Vàzquez, Sergi Alvarez-Vidal and Antoni Oliver

11:48 *Comparable Corpora in Cross-linguistic Research: Nominal Number in English, Czech, and Greek*
Konstantinos Diamantopoulos and Magda Ševčíková

12:12 *Liebe Kolleg:innen, Querid@s Compañer@s: Presenting the GILDEES Corpus*
Marie-Pauline Krielke

12:36 *A Diachronic Comparable Corpus of Spanish Digital News (2017–2026) for the Study of Stylistic Convergence in the GenAI Era*
Hugo Sanjurjo-González

Monday, May 11, 2026 (continued)

13:00 **Lunch break**

14:00 **Session 3: Synthetic corpora**
Chair: Serge Sharoff

14:00 ***Panel discussion: How comparable are synthetic data?***

15:12 *Align and Shine: Building High-quality Sentence-aligned Corpora for Multilingual Text Simplification*
Luis Kenji Hilasaca Sanchez, Nouran Khallaf and Serge Sharoff

15:36 *Bi-Text Mining across German Dialects: On the Role of Synthetic Training Data for Dialect Adaptation*
Jing Wang, Barbara Plank and Robert Litschko

16:00 **Coffee break**

16:30 **Session 4: Building comparable datasets**
Chair: Pierre Zweigenbaum

16:30 *Parallel Corpora of Scholarly Documents for English-French Machine Translation*
Ziqian Peng, Lichao Zhu, Rachel Bawden, Maud Bénard, Éric de la Clergerie, Mathilde Huguin, Natalie Kübler, Paul Lerner, Alexandra Mestivier and François Yvon

16:54 *Validating a Pipeline to Create a Comparable Corpus of Government-Issued Travel Advisories from the Internet Archives*
Laura Braun and Christian Oswald

17:18 *Leveraging Comparable Toxicity Lexicons in Prompt Instructions for Multilingual Text Detoxification*
Yassir El Attar, Esra Dönmez, Nina K. Ohlendorf and Agnieszka Falenska

17:42 ***Closing words***

Invited Talk: The Cross-Lingual Transfer Myth: Why Modern LLMs Still Fail Without Comparable Corpora and Representations

Els Lefever
LT3, Ghent University
Els.Lefever@UGent.be

Abstract

Comparable corpora have long served as a foundation for multilingual NLP, supporting transfer across languages in tasks such as classification, retrieval, translation, and argument mining. Yet in the era of multilingual transformers and generative models, a central question is no longer simply whether texts are comparable, but what kinds of internal representations and downstream behaviors that comparability actually enables. In this keynote, I argue that cross-lingual transfer is best understood as a continuum oscillating between shared semantic structures and language-specific realizations. Drawing on two complementary studies, I demonstrate how this tension manifests both in the data models learn from and in the representations they develop.

The first case study investigates multilingual stance and argument mining using the new Russian LoveHate corpus alongside English debate data. The results indicate that translated or multilingual resources are useful but insufficient proxies for language-specific corpora: local topics, culturally situated argumentation patterns, and stance expression still shape model performance and generalization. The second case study presents a neuron-level analysis of multilingual emotion detection, showing that multilingual encoders such as XLM-R develop both polyglot neurons, which respond consistently across languages, and monolingual neurons, which remain tied to particular linguistic systems. This reveals that even successful cross-lingual emotion transfer depends on only partial internal alignment.

Together, these findings suggest that multilingual NLP needs corpora that preserve culturally specific meaning while supporting robust transfer, as well as interpretability frameworks that can diagnose where multilingual systems genuinely share representations and where they merely approximate them. Comparable corpora are not just training material; they are essential to understand how cross-lingual generalization succeeds, where it breaks down, and how truly multilingual NLP can move beyond English-centric assumptions and conclusions.

Keywords: Comparable Corpora; Representation; Cross-lingual Transfer; Stance and Argument Mining; Emotion Detection; Culturally Specific Meaning

Bio

Els Lefever is associate professor at the LT3 language and translation technology team (Ghent University). She holds a PhD in computer science on ParaSense: Parallel Corpora for Word Sense Disambiguation (2012). Els has a strong expertise in machine learning of natural language and multilingual natural language processing, with a special interest for computational semantics, language modeling of lower-resourced languages and multilingual terminology extraction. She currently supervises research on complex reasoning in large language models, argumentation mining in social media, the automatic detection of irony, stereotypes and bias in online text, multimodal emotion detection and generation, and NLP approaches for low(er)-resourced languages, such as cuneiform, Byzantine Greek, or historical travelogues.

Proceedings of the 19th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC 2026, page 1

A Comparative Study of Parkinsonian Speech Corpora for Deep Learning-Based Detection of Dysarthria

Clara Ponchard, Pierre Serrano
Inria, France
clara.ponchard@inria.fr, pierre.serrano@inria.fr

Abstract

Idiopathic Parkinson's disease is associated with motor speech impairments collectively referred to as hypokinetic dysarthria, which can appear at early disease stages and remain challenging to assess objectively in clinical practice. Most automatic assessment studies rely on individual speech corpora analyzed in isolation, leaving open questions regarding their comparability and their suitability for joint use within unified classification frameworks. This study explicitly investigates the cross-corpus comparability of existing Parkinsonian speech datasets designed for hypokinetic dysarthria assessment. Rather than assuming their compatibility, we evaluate it empirically through the generalization performance of classification systems trained on single or multiple corpora. We examine which datasets can be effectively combined and whether multi-corpus training improves robustness across heterogeneous recording conditions and speech tasks. Four corpora are evaluated under intra-corpus, cross-corpus, and out-of-domain settings. Results demonstrate that multi-corpus training enhances robustness and generalization performance, while also revealing substantial differences in cross-dataset compatibility. These findings provide a clearer understanding of the degree of comparability between existing resources and offer practical guidelines for the design of future corpora and more generalizable tools for the automatic clinical assessment of Parkinsonian speech.

Keywords: Parkinson, dysarthria, deep Learning, self-supervised model

1. Introduction

Idiopathic Parkinson's disease (IPD) is characterized by progressive degeneration of dopaminergic neurons in the substantia nigra, resulting in opaminergic denervation of the striatum (Grabli, 2017). Clinically, it is primarily defined by the motor triad of bradykinesia, rigidity, and resting tremor. Among the additional motor manifestations, speech disorders affect nearly 80% of patients and are perceived as particularly disabling (Hartelius and Svensson, 2009). These impairments, collectively referred to as hypokinetic dysarthria, may involve alterations in respiration, phonation, articulation, resonance, and prosody (Pinto et al., 2010), and can emerge during the early stages of the disease or even in the prodromal (presymptomatic) phase (Logemann et al., 1978; Ho et al., 1998; Sapir, 2014). The most prominent speech abnormalities include monopitch, reduced stress, monoloudness, imprecise consonant production, inappropriate pauses, short rushes of speech, hoarse voice quality, continuous breathiness, altered pitch, and variable speech rate with a tendency toward acceleration (Darley et al., 1969). Perceptual auditory evaluation by clinicians remains the clinical gold standard for the diagnosis and longitudinal monitoring of hypokinetic dysarthria; however,it is widely debated due to its subjective nature (Ghio et al., 2007). Clinicians have therefore emphasized the need for objective and quantifiable tools to complement perceptual assessment and to enable more reliable monitoring of therapeutic interventions (Laaridh, 2017).

Automatic speech processing methods offer a promising alternative by enabling the extraction of discriminative vocal features. However, evaluation on data collected under conditions not represented during training remains essential for real-world applicability. Most existing systems rely on a single corpus or a specific task, which limits their generalization capacity and restricts the broader exploitation of learned representations to analyze data structure, for example through clustering or the identification of similar acoustic profiles. In addition, the majority of studies use embeddings extracted from the last layer of self-supervised audio models such as Wav2Vec 2.0, combined with simple classifiers, which constrains robustness and cross-corpus transferability (Hireš et al., 2023; Klempíř and Krupička, 2024; Ibarra et al., 2023; Postma and Tejedor-Garcia, 2025). Cross-corpus evaluation is therefore crucial to assess the generalization ability of learned representations and to examine the actual comparability of corpora from a classification perspective. However, no standard benchmark currently exists, and commonly used metrics such as accuracy do not always ensure reliable comparisons across studies.

To address these limitations, we propose a deep learning-based system for the automatic detection of hypokinetic dysarthria, inspired by techniques from speaker recognition. The system's robustness and generalization ability are evaluated in a cross-corpus setting using four datasets (Neurovoz, IPVSD, MDVR-KCL, and AHN), covering multiple languages, speech tasks, and recording conditions, and totaling 380 speakers, including 211 patients with Parkinson's disease. This cross-evaluation

Proceedings of the 19th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC 2026, pages 2–8
11 May 2026.

aims not only to assess system robustness but also to determine to what extent these corpora are comparable and can be jointly used within a multi-corpus framework. Among them, the French AHN corpus, previously unused for classification tasks, is the only dataset including recordings in the OFF-DOPA condition, that is, when patients are recorded without the effect of dopaminergic medication. It also includes a range of UPDRS speech scores, offering a unique opportunity to analyze system performance in clinically realistic and more challenging conditions.

The main objectives of this study are: (i) to analyze the impact of single- and multi-corpus training on performance under intra-corpus, cross-corpus, and out-of-domain conditions; (ii) to assess the comparability of existing corpora through system generalization and identify effective corpus combinations for joint training; and (iii) to propose a reproducible and clinically relevant evaluation protocol based on appropriate metrics, such as fixed-point AUC.

This approach enables the quantification of model robustness to Parkinsonian speech variability, provides insight into the compatibility of existing resources, and highlights the limitations of current systems, particularly in realistic screening scenarios. It thus contributes to the development of evaluation strategies that improve the clinical relevance and transferability of automatic hypokinetic dysarthria detection systems.

2. Materials and Methods

2.1. Corpus

In this study, four corpora were used, the first three of which are freely accessible: NeuroVoz (Mendes-Laureano et al., 2024), IPVSD (Dimauro et al., 2017), MDVR-KCL (Jaeger et al., 2019), and AHN (Ghio et al., 2021). A brief description of each corpus is presented below. Table 1 summarizes the number of participants in the Parkinson's disease (PD) and healthy control (HC) groups, along with their distribution by sex and mean age.

2.1.1. NeuroVoz

The NeuroVoz dataset, which is publicly available, is a collection of speech recordings designed for the development and validation of machine learning models for the diagnosis and monitoring of Parkinson's disease (PD) (Mendes-Laureano et al., 2024). It was recorded jointly by the Bioengineering and Optoelectronics Group (ByO) at the Universidad Politécnica de Madrid (UPM) and the Departments of Otorhinolaryngology and Neurology at Hospital General Universitario Gregorio Marañón (HGUGM) and Hospital Universitario de Fuenlabrada (HUF). We analyze a subset comprising voice recordings from 107 native speakers of Castilian Spanish, including 54 healthy control subjects and 53 patients with PD. Patients were recorded in the ON state, after taking their prescribed medication between two and five hours prior to the recording session. The protocol includes four types of speech tasks: (1) sustained phonation of vowels; (2) repetition of 15 predefined sentences; (3) a diadochokinetic (DDK) task based on the rapid repetition of /pa-ta-ka/; and (4) a free monologue based on the description of an illustration. Recordings were conducted under standardized conditions using an AKG C420 head-mounted microphone, with a sampling rate of 44.1 kHz and a 16-bit resolution.

2.1.2. IPVSD

The Italian Parkinson's Voice and Speech Database (IPVSD) corpus was created to assess speech intelligibility in patients with Parkinson's disease using automatic speech recognition systems (Dimauro et al., 2017). It includes recordings from 65 Italian speakers, mainly from the Bari region, divided into three groups: 15 young healthy controls aged 19 to 29 years (13 men, 2 women), 22 older healthy controls aged 60 to 77 years (10 men, 12 women), and 28 patients with Parkinson's disease aged 40 to 80 years (19 men, 9 women). The Parkinsonian patients were recorded in the ON state, after taking their prescribed medication between two and five hours prior to the recording session. The protocol includes several speech tasks: (1) reading of phonetically balanced texts, words, and sentences; (2) sustained vowel phonation; and (3) diadochokinetic (DDK) tasks (/pa/ and /ta/). The acoustic signals were recorded in uncompressed WAV format, with a sampling rate of 44.1 kHz.

2.1.3. MDVR-KCL

The MDVR-KCL corpus, which is publicly available, was recorded at King's College London (KCL) Hospital under conditions designed to replicate a realistic telephone call scenario, with participants holding the phone to their preferred ear and the microphone positioned close to the mouth (Jaeger et al., 2019). It includes voice recordings from 37 native English speakers, comprising 21 healthy control subjects and 16 patients with Parkinson's disease (PD). No information is provided regarding the participants' sex distribution or age, nor about whether patients had taken medication prior to the recording sessions. The protocol includes reading aloud the text "The North Wind and the Sun", optionally followed by the reading of a technical passage, and a spontaneous verbal exchange with the examiner. Recordings were made using a Motorola Moto G4

Corpus	Language	Participants					Age (mean ± SD)			
		Total	Female		Male		Female		Male	
			PD	HC	PD	HC	PD	HC	PD	HC
Neurovoz	Spanish	107	20	26	33	28	67.2 ± 9.1	66.6 ± 12.3	69.8 ± 11.4	61.6 ± 7.5
IPVSD	Italian	65	9	14	19	23	64.3 ± 12.2	58.7 ± 17.0	68.6 ± 6.4	42.0 ± 24.8
MDVR-KCL	English	37	-	-	-	-	-	-	-	-
AHN	French	171	43	40	77	11	64.1 ± 9.9	62.2 ± 8.6	65.8 ± 10.0	66.6 ± 14.1

Table 1: Distribution of Parkinson's disease (PD) and healthy control (HC) participants across corpora, including language, sex, and mean age (± standard deviation, SD).

smartphone via a dedicated application (Toggle Recording App), stored in uncompressed WAV format, with a sampling rate of 44.1 kHz and a 16-bit resolution.

2.1.4. AHN

The AHN (Aix Hôpital Neurologie) corpus consists of acoustic and aerodynamic recordings collected over more than twenty years by the Neurology Department of the Aix-en-Provence Hospital Center (Ghio et al., 2021). It is distinguished by the presence of aerodynamic signals synchronized with the acoustic recordings, as well as by the diversity of recording conditions (with or without L-DOPA, with or without deep brain stimulation). We analyze a subset of the corpus comprising 171 French-speaking participants, including 120 patients with Parkinson's disease (PD) and 51 healthy control subjects. The Parkinsonian patients were treated exclusively with L-DOPA. Most patients were recorded in two pharmacological states: OFF-DOPA, after at least 12 hours of medication withdrawal, and ON-DOPA, after a minimum delay of 1.5 hours following L-DOPA intake. Before each recording session, dysarthria severity was assessed by a neurologist using Item 18 of the UPDRS scale (Fahn and Elton, 1987), with scores ranging from 0 to 3. The analyzed task consisted of repeating the sentence "papa ne m'a pas parlé de beau papa", designed for the evaluation of plosive consonants. Data were collected using the Assisted Voice Evaluation system EVA2 (Teston et al., 1999), which enables synchronized acquisition of acoustic and aerodynamic signals via a handheld device mounted on an adjustable stand and coupled with a sealed silicone mask fitted around the speaker's mouth. The acoustic signal was recorded in WAV format using an AKG C419 directional microphone positioned 4 cm from the mask, with a sampling rate of 25 kHz.

2.2. Method

The task consists of a binary classification aimed at predicting whether a speaker is a healthy control (HC) or a Parkinson's disease (PD) patient from an audio recording. The architecture combines a self-supervised model, Wav2Vec 2.0 XLS-R (Babu et al., 2022), pre-trained on a large multilingual corpus, with a multi-head feature aggregation (MHFA) classification head that aggregates representations from multiple SSL layers with a hidden dimension of 128 (Peng et al., 2022). The parameters of the SSL model are frozen, and only those of the MHFA head are optimized, allowing the model to implicitly select the most informative layers. This approach is inspired by modern solutions for speaker recognition and audio deepfake detection (Wang et al., 2024).

The corpora described above are used in the experiments. For each corpus, the data are split into 80% for training and 20% for testing, with 20% of the training set dedicated to development. Five independent random splits are generated to assess robustness (Postma and Tejedor-Garcia, 2025; Gimeno-Gómez et al., 2025), ensuring strict speaker separation. Performance is reported as mean and standard deviation. Audio recordings are resampled to 16 kHz and segmented into randomly selected one-second chunks for training, with balanced batches and optimization using Adam with a cross-entropy loss function. Training runs for at least 100 epochs, and the last checkpoint is retained. During inference, the full recording is provided to the model, and the MHFA head aggregates temporal representations into a single prediction.

Evaluation is based on accuracy and AUC. Intra-corpus evaluation uses the test set from the same corpus as training, whereas inter-corpus evaluation tests on a different corpus, representing a more challenging out-of-domain scenario. Models can be trained in either mono-corpus or multi-corpus settings. All predictions are made at the recording level, without speaker-level aggregation. The MDVR-KCL corpus, whose size does not allow a split into training and test sets, is used exclusively as an out-of-domain evaluation corpus.

Test →	Neurovoz		IPVSD		AHN		MDVR-KCL	
Train ↓	Acc (%)	AUC (%)	Acc (%)	AUC (%)	Acc (%)	AUC (%)	Acc (%)	AUC (%)
Neurovoz (N)	76.8 ± 0.7	84.9 ± 2.3	69.9 ± 7.4	43.4 ± 7.3	79.6 ± 1.4	51.2 ± 3.3	83.6 ± 4.3	90.2 ± 3.2
IPVSD (I)	65.9 ± 3.0	68.2 ± 5.0	94.4 ± 2.7	95.8 ± 4.1	80.8 ± 4.5	66.7 ± 8.0	70.4 ± 2.5	72.7 ± 4.2
AHN (A)	59.5 ± 3.6	54.7 ± 3.8	69.4 ± 6.6	63.9 ± 2.1	83.4 ± 4.6	79.2 ± 7.4	69.9 ± 6.0	67.1 ± 9.7

Table 2: Cross-corpus performance of systems trained on a single corpus. Accuracy (Acc) and area under the ROC curve (AUC) are reported with their standard deviations (±) computed over five independent splits. Grey cells indicate that evaluation is performed on the test set of the corpus used for training.

Test →	Neurovoz		IPVSD		AHN		MDVR-KCL	
Train ↓	Acc (%)	AUC (%)	Acc (%)	AUC (%)	Acc (%)	AUC (%)	Acc (%)	AUC (%)
I + A	61.2 ± 2.9	61.4 ± 6.6	93.8 ± 4.3	92.9 ± 9.2	83.0 ± 4.6	78.4 ± 8.7	74.5 ± 3.9	74.1 ± 4.6
N + A	75.9 ± 1.8	84.1 ± 3.8	66.8 ± 9.6	35.5 ± 10.1	83.8 ± 2.3	81.3 ± 3.6	79.2 ± 1.8	84.0 ± 3.6
N + I	78.6 ± 2.2	85.9 ± 3.7	91.5 ± 3.6	95.0 ± 4.9	77.7 ± 2.6	59.2 ± 6.7	82.2 ± 1.9	87.2 ± 2.6
N + I + A	77.7 ± 3.4	85.4 ± 4.5	91.5 ± 3.1	93.6 ± 7.1	83.7 ± 3.4	84.3 ± 4.2	80.8 ± 1.4	85.3 ± 1.6

Table 3: Cross-corpus performance of systems trained on corpus combinations. Metrics and conventions follow those of Table 2. Grey cells indicate intra-corpus evaluation.

3. Results

The performance of systems trained on a single corpus is presented in Table 2. In the intra-corpus setting, mono-corpus models generally achieve strong performance. The model trained on IPVSD (I) obtains the highest AUC (95.8%), followed by Neurovoz (N) (84.9%) and AHN (A) (79.2%).

Cross-corpus evaluation reveals more heterogeneous generalization capabilities. The N model transfers effectively to MDVR-KCL (AUC of 90.2%), whereas the I and A models exhibit more limited transfer performance. Notably, applying the N model to the A corpus results in a high accuracy (79.6%) but a near-random AUC (51.2%), indicating unstable or poorly calibrated discrimination behavior.

Bi-corpus combinations (Table 3) show contrasting synergies. The N+I pair stands out in its generalization to MDVR-KCL (87.2% AUC), but fails on AHN (59.2%). All pairs systematically fail on at least one unseen corpus. In contrast, multi-corpus training on all three corpora (N+I+A) yields robust performance across all datasets and avoids the severe degradation observed with mono- or bi-corpus models. Results on MDVR, which was never seen during training, further confirm the robustness of this approach (AUC 85.3%).

ROC curves (Figure 1) indicate that global AUC does not always reflect clinically meaningful performance. At operating points with low false positive rates (FPR), which are more relevant in clinical settings, performance gains differ across corpora. The evaluation also reveals strong sensitivity to random splits, particularly for IPVSD and AHN, highlighting limitations related to the small size of these datasets.

A post-hoc analysis of the N+I+A detector was conducted on a representative split (close to the average performance across the five training runs). [1]

4. Discussion

The results show that multi-corpus training on three corpora (N+I+A) provides the best trade-off between specialization and generalization. While mono-corpus models achieve strong intra-corpus performance, their cross-corpus transfer capabilities remain limited. All bi-corpus combinations fail on at least one unseen corpus, confirming that exposure to only two protocols is insufficient to capture the acoustic diversity associated with Parkinson's disease, as observed in (Ibarra et al., 2023).

The analysis of ROC curves highlights the limitations of accuracy (and, by extension, the F1-score) as primary evaluation metrics for hypokinetic dysarthria corpora, given their strong heterogeneity and often limited size. Systems displaying high accuracy may nevertheless exhibit near-random discrimination ability, as illustrated by the inference of the N system on the AHN corpus (Table 3). Evaluation based on AUC or on fixed operating points, particularly at low false positive rates, appears essential for comparing systems and more faithfully assessing their real clinical applicability, beyond the considerations typically reported in the literature (Ozbolt et al., 2022). Furthermore, the absence of a fixed, shared evaluation corpus strictly independent from training constitutes a major obstacle to reliable

[1] It is not possible to merge systems obtained from different training runs, as such an approach would require a common calibration and the use of identical training data.

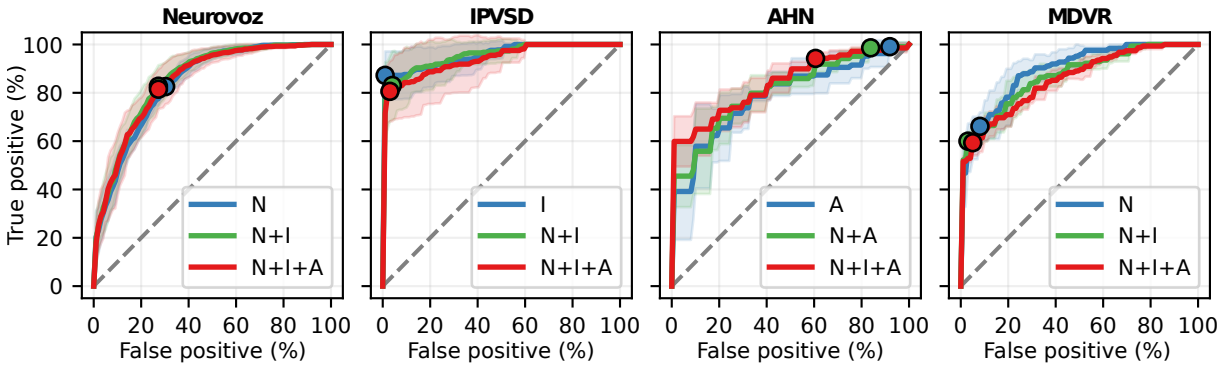


Figure 1: Mean ROC curves (± standard deviation) for models trained on one or several of the Neurovoz (N), IPVSD (I), and AHN (A) corpora. Dots indicate the operating point maximizing classification accuracy for each evaluated system.

comparisons across systems, architectures, and learning parameter configurations.

The use of a frozen self-supervised learning (SSL) model combined with a multi-layer classification head helps mitigate overfitting in a low-data setting. This approach achieves performance comparable to, and in some cases exceeding, the state of the art across all corpora. Although differences in evaluation protocols make direct comparisons challenging, our results are competitive: for the DDK task of the Neurovoz corpus, we achieve 90% accuracy, compared to 85% reported by (Postma and Tejedor-Garcia, 2025) and 88.6% by (Ibarra et al., 2023).

The robustness of our approach is further confirmed by the results on MDVR, which was never seen during training, and by consistently high AUC values across all test sets. A preliminary analysis suggests that intermediate layers, particularly around the 20th layer, play a key role, likely encoding prosodic information relevant for hypokinetic dysarthria detection. Beyond metric selection, evaluation protocols also raise important issues in comparisons of SSL models. While several studies compare such models, these comparisons most often rely on simplified protocols, typically involving a linear classifier or an SVM applied to the final layer representation. However, as recently discussed in (Zaiem et al., 2025), such configurations are not appropriate for establishing reliable benchmarks, particularly in low-resource settings. Our results show that this type of approach may underexploit SSL representations and lead to unstable conclusions, reinforcing the previously highlighted need for stricter and more clinically meaningful evaluation protocols.

5. Conclusion

The use of pre-trained self-supervised audio models, combined with multi-corpus training, enables the development of a hypokinetic dysarthria detection system capable of generalizing across languages, datasets, and recording conditions, including out-of-domain scenarios. Multi-corpus training therefore emerges as a central lever to overcome the limitations of approaches evaluated on single, homogeneous datasets.

Beyond predictive performance, the embeddings extracted from the classification head of the proposed system—explicitly constrained by the hypokinetic dysarthria detection task and validated in a multi-corpus setting—constitute a relevant foundation for future post hoc analyses of clinical databases. They open the way to investigations such as clustering of recordings or patients, with the objective of identifying homogeneous acoustic profiles and examining their associations with clinical or protocol-related factors.

The results highlight the importance of rigorous evaluation protocols and appropriate metrics, such as fixed-operating-point AUC, as well as the need for shared evaluation corpora that are independent from training data and cover diverse clinical conditions.

Finally, post hoc analyses conducted on the AHN corpus reveal the current limitations of the models in cases of mild dysarthria (low UPDRS scores) and under realistic screening conditions (OFF medication state). These findings suggest that targeted data augmentation strategies, fine-grained balancing of clinical tasks, and multimodal approaches represent promising directions to enhance both robustness and clinical relevance.

6. Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation 2026-AD011014982R2 made by GENCI.

7. Bibliographical References

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2022. Xls-r: Self-supervised cross-lingual speech representation learning at scale. In *Proc. Interspeech 2022*, pages 2278–2282.

Frederic L. Darley, Arnold E. Aronson, and J. R. Brown. 1969. Clusters of Deviant Speech Dimensions in the Dysarthrias. *Journal of Speech and Hearing Research*, 12(3):462–496.

Stanley Fahn and Richard L. Elton. 1987. The unified parkinson's disease rating scale. In Stanley Fahn, C. David Marsden, Donald Calne, and Melvin Goldstein, editors, *Recent Developments in Parkinson's Disease*, volume 2, pages 153–163. Macmillan Health Care Information, Florham Park.

Alain Ghio, Gilles Pouchoulin, Antoine Giovanni, Corinne Fredouille, Bernard Teston, Joana Révis, Jean-François Bonastre, Danièle Robert-Rochet, Ping Yu, Maurice Ouaknine, Marie-Dominique Guarella, Christine Spezza, Thierry Legou, and Alain Marchal. 2007. Approches complémentaires pour l'évaluation des dysphonies : bilan méthodologique et perspectives. *Travaux interdisciplinaires du Laboratoire Parole et Langage*, 26:33–74. Autorisation No.3240 : TIPA est la revue du Laboratoire Parole et Langage.

David Gimeno-Gómez, Catarina Botelho, Anna Pompili, Alberto Abad, and Carlos-D Martínez-Hinarejos. 2025. Unveiling interpretability in self-supervised speech representations for parkinson's diagnosis. *IEEE Journal of Selected Topics in Signal Processing*.

David Grabli. 2017. Maladie de parkinson et syndromes parkinsoniens : les signes moteurs. *La Presse Médicale*, 46(2, Part 1):187–194.

L. Hartelius and P. Svensson. 2009. Speech and swallowing symptoms associated with parkinson's disease and multiple sclerosis: A survey. *Folia Phoniatrica et Logopaedica*, 46(1):9–17.

M. Hireš, P. Drotár, N. D. Pah, Q. C. Ngo, and D. K. Kumar. 2023. On the inter-dataset generalization of machine learning approaches to parkinson's disease detection from voice. *International Journal of Medical Informatics*, 179:105237.

A. K. Ho, R. Iansek, C. Marigliani, J. L. Bradshaw, and S. Gates. 1998. Speech impairment in a large sample of patients with parkinson's disease. *Journal of Behavioural Neurology*, 11:131–137.

Emiro J Ibarra, Julián D Arias-Londoño, Matías Zañartu, and Juan I Godino-Llorente. 2023. Towards a corpus (and language)-independent screening of parkinson's disease from voice and speech through domain adaptation. *Bioengineering*, 10(11):1316.

Ondrej Klempíř and Roman Krupička. 2024. Analyzing wav2vec 1.0 embeddings for cross-database parkinson's disease detection and speech features extraction. *Sensors*, 24:5520.

I. Laaridh. 2017. *Évaluation de la parole dysarthrique*. Ph.D. thesis, Université d'Avignon.

J. A. Logemann, H. B. Fisher, B. Boshes, and E. R. Blonsky. 1978. Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of parkinson patients. *Journal of Speech and Hearing Disorders*, 43(1):47–57.

Ana S. Ozbolt, Laura Moro-Velazquez, Irene Lina, Adam A. Butala, and Najim Dehak. 2022. Things to consider when automatically detecting parkinson's disease using the phonation of sustained vowels: Analysis of methodological issues. *Applied Sciences*, 12(3):991.

J. Peng, O. Plchot, T. Stafylakis, L. Mošner, L. Burget, and J. Černocký. 2022. An attention-based backend allowing efficient fine-tuning of transformer models for speaker verification. In *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, pages 555–562.

S. Pinto, A. Ghio, B. Teston, and F. Viallet. 2010. La dysarthrie au cours de la maladie de parkinson. *Revue Neurologique*, 166(10):800–810.

Emmy Postma and Cristian Tejedor-Garcia. 2025. Evaluating the Effectiveness of Pre-Trained Audio Embeddings for Classification of Parkinson's Disease Speech Data. In *Interspeech 2025*, pages 4603–4607.

S. Sapir. 2014. Multiple factors are involved in the dysarthria associated with parkinson's disease. *Journal of Speech, Language, and Hearing Research*, 57(4):1330–1343.

Bernard Teston, Alain Ghio, and Bernard Galindo. 1999. A multisensor data acquisition and processing system for speech production investigation. pages 2251–2254.

X. Wang, H. Delgado, H. Tak, et al. 2024. Asvspoof 5: crowdsourced speech data, deepfakes, and adversarial attacks at scale. In *Proceedings of the ASVspoof 2024 Workshop*, pages 1–8.

Salah Zaiem, Youcef Kemiche, Titouan Parcollet, Slim Essid, and Mirco Ravanelli. 2025. Speech self-supervised representations benchmarking: a case for larger probing heads. *Computer Speech & Language*, 89:101695.

8. Language Resource References

Dimauro, Giovanni and Di Nicola, Vincenzo and Bevilacqua, Vitoantonio and Caivano, Danilo and Girardi, Francesco. 2017. *Assessment of Speech Intelligibility in Parkinson's Disease Using a Speech-To-Text System*.

Ghio, Alain and Pouchoulin, Gilles and Viallet, François and Giovanni, Antoine and Woisard, Virginie and Crevier-Buchman, Lise and Hirsch, Fabrice and Fauth, Camille and Fredouille, Corinne. 2021. *Du recueil à l'exploitation des corpus de parole « pathologique » : comment accéder à la variation physiopathologique ?* Corpus. PID https://doi.org/10.4000/corpus.5677.

Jaeger, Hagen and Trivedi, Dhaval and Stadtschnitzer, Michael. 2019. *Mobile Device Voice Recordings at King's College London (MDVR-KCL) from both early and advanced Parkinson's disease patients and healthy controls*. Zenodo. PID https://doi.org/10.5281/zenodo.2867216.

Mendes-Laureano, J. and Gómez-García, J. A. and Guerrero-López, A. and others. 2024. *NeuroVoz: a Castillian Spanish corpus of parkinsonian speech*. PID https://doi.org/10.1038/s41597-024-04186-z.

Computing Semantic Similarity for Aligning Bilingual Semi-parallel Texts: A Case Study

Steffen Frenzel, Maximilian Krupop, Manfred Stede
University of Potsdam, Applied Computational Linguistics
{steffen.frenzel, maximilian.krupop, stede}@uni-potsdam.de

Abstract

Semi-parallel text refers to versions of the same text that have to some extent been edited by authors, translators, or others. They are of relevance especially in the social sciences and in literary genres. In this paper, we consider the bilingual (English/German) variant of the problem. The philosopher Hannah Arendt, for example, wrote political essays that often exist in multiple versions and in both languages. She repeatedly modified her texts, added or deleted parts, and framed topics differently for target audiences. For researchers to explore the history of such material in detail, and at the same time at scale, automatic alignment (i.e., finding the best match of semantically similar sentences) is a very valuable preprocessing step. In this paper, we compare the performances of a range of methods for this task, based on computing semantic similarity. We present the results and conduct a qualitative error analysis to identify recurring sources of error.
Keywords: Semi-parallel Text, Alignment, Semantic Similarity

1. Introduction

Automatic text alignment has been studied for a long time in the context of machine translation, where the aim is to induce translation models from texts known to be parallel, i.e., close translations of one another. The two widely-used levels in the alignment problem are sentences and words [Jurafsky and Martin, 2026].

In our work we are concerned with pairs of text that are similar to each other but decidedly *not parallel*. These arise in a variety of scenarios. For instance, the translation of literary texts into other languages should intuitively be "close" to the original; however, there are several reasons why such translations may contain different types of changes. Obviously, the language pair itself can cause syntactic or semantic mismatches [Dorr, 1994]; e.g., idiomatic expressions or names often cannot be translated literally and may therefore take a different shape in the target language.

Further, individual translators or publishers (or also government authorities) may invoke changes. An example of this can be found in J.D. Salinger's novel *The Catcher in the Rye*, which we will analyze in more detail below. While the original in American English contains numerous swear words and vulgar language, the German translation from the 1960s shows almost no offensive terms.

Another source of changes can be the authors themselves when they deliberately produce variants of their texts. This is rare in literary works but not uncommon in the social sciences, where adaptations can be done, e.g., in tailoring to a new publication medium or to a different target audience.

For example, the German-American philosopher Hannah Arendt wrote several political essays that were published around and after the Second World War. Depending on the language version, topics such as the persecution of Jews and fascism are framed differently in these texts. In addition, there are revisions that were intended, for example, as radio lectures: Arendt restructured the texts for this purpose, shortening them in places and adding new aspects in others.

Such text adaptations are the subject of debate in literary and translation studies, as well as in the social sciences. Because scholars are interested in studying the nature and potential reasons for adaptions, an interesting problem arises for computational linguistics: determining the "best" alignment of text variants, on the level of sentences. A pre-aligned pair of texts can, after all, be compared and analyzed much more effectively than two raw versions.

Technically, text pairs such as those by Salinger and Arendt provide a good basis for testing the suitability of automatic measurements of *semantic similarity*. For example, the widely-used approach of embedding sentences and comparing them with the cosine metric is assumed to distinguish (mono- or multi-lingual) paraphrases from "unrelated" sentences, but can it capture degrees of shift in meaning that arise in semi-parallel sentences?

In this paper, we test the capabilities of various models for computing semantic similarity as vehicles for automatically aligning semi-parallel text on the sentence level. We create manually-aligned gold data for text material from the two sources mentioned above (Arendt, Salinger) and then compare the similarity models within a window-based alignment algorithm that we designed for handling semi-parallel texts. We report all results using stan-

Proceedings of the 19th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC 2026, pages 9–19
11 May 2026.

dardized evaluation metrics and provide error analyses for both sets of text.

Section 2 discusses relevant related work. Section 3 introduces the data we are using, and Section 4 explains the methods, in particular the semantic similarity measurements. Section 5 reports the results of the experiments, and Section 6 concludes.

2. Related Work

2.1. Semi-parallel Text

When machine translation turned to statistical methods in the 1990s, the term 'parallel corpora' was used to refer to pairs of texts that were understood as direct translations into another language [Wołk and Marasek, 2017]. These corpora formed the foundation for the induction of statistical translation models.

In between 'different' texts and 'parallel' texts, however, there is effectively a continuum of 'similar' texts, often regarded as a range of 'comparable' corpora [Cheung and Fung, 2004]. These materials have sparked quite a bit of research, which often focused on the task of extracting parallel sentences (e.g., Tillmann [2009], Rauf and Schwenk [2011], Smith et al. [2010], Chu et al. [2013]). In this research the term 'quasi-comparable' is sometimes used for texts that address the same or a similar topic [Cheung and Fung, 2004].

In the aforementioned continuum, we see *semi-parallel* text as one step removed from the 'direct translation': Two texts have the same author, the overall intention is the same in both versions, but occasional modifications have been made. This covers both the monolingual and the multilingual case; though in this paper we focus on a bilingual English/German scenario.

For the semi-parallel setting, research on paraphrase detection and paraphrase generation can provide relevant background. The concept of paraphrases describes the possibilities to change sentences on a lexical, morphological or syntactic level without affecting the meaning [Wahle et al., 2023]. The problems of paraphrase detection (e.g., Gold et al. [2019], Liu and Soh [2022]) and paraphrase generation (e.g., Bandel et al. [2022], Yang et al. [2022]) have lea to manifold research efforts, including the conception of shared tasks based on publicly-available data.

Furthermore, paraphrases are also being analyzed as a phenomenon of 'intertextuality' in the context of digital humanities (e.g., Sier and Wöckener-Gade [2019]). Intertextuality refers to connections between texts that were not necessarily intended by the author. It is described as a phenomenon of reception; that is, the relation of two paraphrases is established by the reader and is analyzed from this perspective.

Recently, Frenzel and Stede [2025] tackled the task of sentence alignment in semi-parallel monolingual (German) text, which was taken from news texts and their simplifications, encyclopedia articles on writers, and also an essay by Hannah Arendt. Apart from sentences, they also tried to utilize the concept of Elementary Discourse Units (cf. Frenzel et al. [2026]) as a level for text alignment.

Our notion of semi-parallel text is based on these various strands of research. In our current experiments, we work with bilingual versions of texts written by Hannah Arendt and J.D. Salinger; the former is associated with social science, the latter with fictional writing. In both cases the text versions in our corpus are not literal translations, but include adaptations made by re-writing and by translating, respectively. Detailed information on the corpus is provided in Section 3.

2.2. Text Alignment and Semantic Similarity

In the early days of sentence alignment, the assumption of strong parallelism led to rather successful scoring functions that merely compared the number of words or characters [Brown et al., 1991, Gale and Church, 1993]. Later on, lexical features and heuristics have been utilized to improve the alignment quality while still being temporally efficient (e.g. Moore [2002]). More recently, LERA [Pöckelmann et al., 2022] is an example of a system that models the alignment problem in a graph-theoretic fashion and makes its alignment decisions with a distance function based on the Jaccard index [Jaccard, 1901].

Following the introduction of BERT by Devlin et al. [2019], the use of sentence embeddings has become increasingly popular in this field of research. Notably, Reimers and Gurevych [2019] improved the computation of sentence embeddings with their Sentence-BERT (SBERT) model, which mitigates the computational effort of the classical BERT model.

For comparing sentence embedding vectors to each other, classical similarity calculations such as cosine similarity or Euclidean distance have been employed. One of the first systems that implemented this approach to sentence alignment was VecAlign [Thompson and Koehn, 2019]. Both VecAlign and SentAlign [Steingrimsson et al., 2023] are based on bilingual sentence representations such as LASER [Artetxe and Schwenk, 2019] and LaBSE [Feng et al., 2022]. Building on this work, Molfese et al. [2024] introduced CroCoAlign, an algorithm that incorporates more contextual information for disambiguating possible sentence map-

pings.

In the field of neural information retrieval (IR), recent work commonly follows a retrieve-and-rerank approach: an efficient retriever proposes a small candidate set using independent query/document representations (e.g., dense bi-encoders with cosine similarity), which is then refined by a stronger reranker that jointly models query and candidate text [Nogueira and Cho, 2020]. Beyond single-vector representations, late-interaction models such as ColBERT [Khattab and Zaharia, 2020] compute relevance via token-level matching after independent encoding and provide a middle ground between dense retrieval and full cross-encoders. More recently, listwise rerankers allow richer interactions by processing a query together with multiple candidates in one context window and producing per-candidate relevance scores [Wang et al., 2025]. We adopt this IR perspective to model bilingual alignment as retrieving the best-matching target sentence for each source sentence in semi-parallel text.

3. Data

We are using two datasets for our alignment experiments: a book chapter from *The Human Condition*, written by Hannah Arendt, and a part of J. D. Salingers *The Catcher in the Rye*. These texts differ in terms of style, syntactic complexity and also in the level of 'semi-parallelism' and are therefore well suited to test our alignment approaches.

Hannah Arendt's works are currently studied as part of the online publication of a Critical Edition[1], which will allow users to follow the emergence of her works step by step. This is achieved by publishing all versions, including manuscripts, revisions and translations into other languages. Depending on the date of origin, target audience and publication format, these text versions contain numerous variations and differences. Apart from that, Hannah Arendt is known to use long, syntactically complex sentences to express her thoughts. For these reasons, the alignment of the text versions is expected to be challenging.

J.D. Salingers *Catcher in the Rye* is an American coming-of-age novel that was partially published in 1945-46 before being novelized in 1951. This text is interesting for our work for different reasons as well: the novel contains heavily stylized language, containing slang, vulgar language, and many old-fashioned words and expressions. The differences between the American original and the German translation are striking – vulgar language has been considerably "softened" in the German version and the numerous idioms could not be translated literally.

[1]The Critical Edition is published online and can be found here: https://hannah-arendt-edition.net/home

Corpus Statistics				
Datasets:	**Arendt**		**Salinger**	
Language:	DE	EN	DE	EN
Sentences	71	59	135	148
Words	3055	2486	2089	2310
Avg. Words / Sentence	43.0	42.1	15.5	15.6

Table 1: Statistics for both corpora.

Results of the Manual Annotation		
Datasets:	**Arendt**	**Salinger**
IAA (Cohen's Kappa)	0.75	0.97
Non-aligned Share	20.0%	17.0%

Table 2: Results of the manual annotation.

The basic corpus statistics in Table 1 show that the text versions have different lengths: for Arendt's text, the German version is longer than the English one. Average sentence length, on the other hand, is very similar across languages - with more than 40 words per sentence on average, the sentences are very long. For Salinger's text, the English version is longer. Again the sentence length is similar across languages, but with around 15 words per sentence the complexity is much lower compared to *The Human Condition*.

3.1. Manual Annotation and IAA

We conducted a manual annotation study to create gold data for the automatic alignment experiments. Two linguistically-informed annotators worked on both texts, and we can therefore compute Inter-Annotator-Agreement (IAA) for both datasets. Our annotation guidelines were derived from a study that was conducted for monolingual text versions by Frenzel and Stede [2025]. Their guidelines specify that the basis for alignment must always be semantic similarity rather than surface form. It is specified that multiple alignments of the same element should only be made in justified exceptions and that, in contrast, there is no obligation to align all elements. According to these guidelines, the following alignment patterns are allowed: `[1:0, 0:1, 1:n, n:1]`. However, `[n:m]` alignments are not possible.

We use Cohen's Kappa to measure chance-corrected agreement. The results in Table 2 show that the IAA is considerably higher for Salinger's text. This finding is not surprising, since the text is more parallel; this is underscored also by the amount of text that remains non-aligned: In Arendt's essay, 20% out of the 70 sentences from the source text are not aligned (14 sentences), but for Salinger only 17% remain non-aligned (23 sentences).

After both annotators had processed the texts independently and the IAA had been measured, all disagreements were discussed, and a gold standard was derived from both annotations.

4. Methods

4.1. VecAlign

`VecAlign` [Thompson and Koehn, 2019] is a widely-used alignment model based on contextualized embeddings. It uses a Dynamic Programming approach on the cosine distance of `LASER` sentence embeddings [Artetxe and Schwenk, 2019] to align bilingual text versions. However, this approach was originally designed to align parallel texts and cannot produce mappings that violate parallel sentence ordering. For our purposes, we use `VecAlign` as a baseline to test whether a more flexible approach leads to improvements in alignment quality.

4.2. Alignment Window

As our framework for testing different approaches to automatic alignment, we cast bilingual alignment as a retrieval problem: Given a source sentence s_i as a query, the goal is to retrieve the most relevant target sentence(s) t_j from the target text. In contrast to parallel sentence alignment, our semi-parallel setting contains insertions, deletions, and local rewrites. Therefore, we operationalize relevance in terms of semantic similarity instead of surface overlap between segments.

To reflect the largely (though not entirely) monotonic progression of both versions between source and target texts, we restrict the retrieval search space to a sliding window of size w around the current source index. For each query sentence s_i we construct a candidate set $\mathcal{C}_i = \{t_{i-w}, \ldots, t_{i+w}\}$ and compute a relevance score $f(s_i, t_j)$ for each candidate $t_j \in \mathcal{C}_i$. We then select the best-matching target sentence (`1:1`). Candidates can be reused across different queries, i.e. the same target element may be selected by multiple queries (n:1).

In our implementation, a strict `1:1` matching can be enforced but it may put a ceiling on model performance compared to the human annotation we use as gold data. Spans of consecutive target elements may be aligned together, resulting in `1:m` matching; however, span alignment does increase the number of candidates per source element linearly.

Our main interest is to compare several different similarity metrics inside the sliding window to find the best match in the respective candidate set. We test cosine similarity of sentence embeddings in conjunction with three different embedding models; BERTScore and NLI Entailment probability; a neural reranker model; and finally we also prompt GPT-5-Mini for this task. In the following we describe these alternative scorers in more detail.

Cosine Similarity: Calculating the cosine similarity of sentence embeddings is a very popular approach for text alignment, as it offers several practical advantages: The calculation is relatively inexpensive, as embeddings can be precomputed once, i.e., for the similarity computation no additional forward passes through the encoder are required. The approach can also be combined with different embedding models, making it flexible in terms of language coverage and computational complexity. We test cosine similarity within the window approach described above with three different multilingual embedding models: we use `LASER` [Artetxe and Schwenk, 2019] to be able to compare our window algorithm to `VecAlign` directly. In addition to that, we use the two multilingual, BERT-based models `LaBSE` [Feng et al., 2022] and `paraphrase-multilingual-MiniLM-L12-v2` [Reimers and Gurevych, 2019].

BERTScore: The BERTScore [Zhang et al., 2020] is very similar in principle to the cosine approach mentioned above. The algorithm was developed to evaluate the quality of translations or summaries by first calculating the pairwise cosine similarity for all tokens of a candidate text and a reference text. The final BERTScore is obtained by selecting the maximum similarity for all tokens using greedy matching and calculating an average per sentence from this. Optionally, IDF importance weighting can be activated to give rare words greater weight in the BERTScore.

NLI Entailment Probability: Natural Language Inference (NLI) refers to the modeling of inference relationships, which are expressed in the form of binary relations between two textual units (e.g., sentences). An entailment relation holds whenever the truth of one text fragment follows from another text. Therefore, the alignment is in this case modeled as a text classification task: the relation between source sentence and the target sentences is to be labeled as 'entailment', 'neutral'

or 'contradiction'. Our scorer selects the target sentence that is assigned the 'entailment' label with the highest probability. To perform the classification task, we used `joeddav/xlm-roberta-large-xnli`, a RoBERTa model that was explicitly finetuned on NLI-labeled datasets in 15 languages, including German and English.

Neural Reranker: We use `jina-reranker-v3` [Wang et al., 2025] as a multilingual neural reranker. The model scores each query segment against the corresponding set of candidate segments from our context window. Unlike the bi-encoder retrieval from the previous approaches, the reranker models query-candidate interactions during encoding. The model returns a relevance score for each candidate, where we select the highest-scoring candidate as the alignment for our query segment, thus giving us a `1:1` alignment.

LLM Prompting: We prompt OpenAI's GPT-5-Mini[2] via the Responses API. We set the reasoning effort to `low` and use a zero-shot prompt with minimal instructions, where we ask the LLM to align the sentences from German to English based on their indices. We include the query segment s_i and the list of candidate sentences from our context window. To better compare the model to the previous approaches, we test two different alignment settings: in one approach we prompt the model to follow a strict `1:1` matching, while in the other we also specify rules that allow these different forms of alignment: `[1:1, 1:0, 1:n]`. In both cases we then prompt the model to output only valid `JSON`, which we can easily parse for further processing. The full system prompts for both alignment settings can be found in Appendix A.

5. Experimental Results

In the following section, we present the quantitative results of our experiments with the range of models described above. In Section 5.2, we provide a qualitative error analysis of the predictions of selected models.

5.1. Quantitative Evaluation

Model Ranking: In order to compare all models as fairly as possible, for the first set of experiments we reset our window algorithm to its default settings: For each sentence in the source text, exactly one sentence in the target text is required – `1:0` or `1:n` alignments are not permitted in this setup. There are several practical reasons for this: The amount of gold data is not large enough to form a dedicated validation and test set. However, testing `1:0` alignments, which are based on the definition of thresholds using the confidence scores of the models (see below), would require such a validation set. The same applies to span alignment, which in turn must be limited to varying degrees depending on the model. Within the scope of this paper, we therefore provide the performance on the default settings and explore potential improvements by fine-tuning the aforementioned hyperparameters directly on the test set.

Other parameters, such as window size, are not affected by this, as we specified them per dataset and they are therefore identical for all models. While a small window (10 elements in each direction) should suffice for the Arendt text, a larger window was chosen for the Salinger text (10 elements backward and 20 elements forward). This is due to the fact that this text contains more sentences overall and that the target text in this case is longer than the source text – the model should therefore be given a larger context forward than backward.

Recall that `VecAlign` is a standalone system that is not integrated into the window algorithm. It receives the entire German and English text as input and then aligns the sentences using its inherent logic.

The results of all approaches on both datasets are listed in Table 3. While GPT-5-Mini shows the best performance on the Arendt data, the Cosine/LaBSE approach achieves the best f1-score on the Salinger data. Although the best values for both datasets are close to each other, several important differences can be observed: In the Salinger data, precision is higher than recall for all models, while this trend is reversed in the Arendt data. However, since alignment is enforced for each source element in this baseline run, these figures primarily allow conclusions to be drawn about the gold data: There are 8 source elements (11%) in the Arendt text that have not been aligned. The models cannot predict these cases correctly because they are forced to align – and so the recall increases (i.e., the models 'over-align'). One possible way to mitigate this evaluation problem would be to filter the outputs before calculating the scores: if all items that are not aligned `1:1` in the gold data are deleted, possible problems with false negatives and false positives in the evaluation could be avoided. However, the `1:0` and `1:n` alignments in the gold data are indicating a high alignment difficulty - and therefore the evaluation scores would be too generous if these items were excluded from the calculation.

This problem does not exist with Salinger, as all source elements in the gold data have been aligned. The f1 scores therefore confirm the pecu-

[2] `gpt-5-mini-2025-08-07`

Scorer	P	R	f1
Salinger			
GPT-5-Mini	0.89	**0.88**	0.88
jina-reranker-v3	0.84	0.84	0.84
BERTScore	0.87	0.79	0.83
NLI Entailment Prob	0.90	0.81	0.86
Cosine Sim / MiniLM	0.90	0.81	0.86
Cosine Sim / LaBSE	**0.95**	0.86	**0.90**
Cosine Sim / LASER	0.83	0.79	0.81
VecAlign	0.78	0.81	0.79
Arendt			
GPT-5-Mini	**0.86**	**0.95**	**0.90**
jina-reranker-v3	0.85	0.92	0.88
BERTScore	0.84	0.92	0.88
NLI Entailment Prob	0.74	0.81	0.78
Cosine Sim / MiniLM	0.80	0.87	0.83
Cosine Sim / LaBSE	0.85	0.93	0.89
Cosine Sim / LASER	0.77	0.80	0.78
VecAlign	0.74	0.79	0.76

Table 3: Results for different datasets and alignment algorithms using the basic settings - an exact 1:1 alignment is enforced for all models.

liarities of the two datasets, but they tend to be too poor for the predictions on the Arendt text, as the models are forced into some incorrect alignments here.

Apart from that, it is striking that VecAlign achieves the worst results in both cases. Since approaches using cosine similarity generally produce good results, this is probably due to the logic of the aligner, which was developed for parallel data and does not allow alignments that contradict the sentence ordering in the two texts. Cases of such "crossing alignments" are not frequent, but do occur in both datasets.

The NLI Entailment Probability delivers good results on the Salinger data, but falls behind on the Arendt data. One possible explanation for this is the high sentence complexity in this text, which can blur entailment relations and thus lead to incorrect alignments.

Aligning Spans: Our window algorithm allows the alignment of multiple target elements per source element, provided that the target elements follow each other directly. The maximum length of such spans can be controlled by the user; to test the potential of this span annotation, we ran several trials with different span lengths. However, the results were ambivalent: All models based on cosine similarity almost always chose the longest allowed span in this scenario, even though there were only a few cases in the gold data where spans were aligned at all.

This points to a general weakness of cosine similarity: as long as the sentences are semantically related, they seem to have almost exclusively positive effects on cosine similarity when combined. Individual semantically distant words therefore have a much smaller negative effect than related words have a positive effect. This observation is also supported by the fact that even the cosine similarity between widely differing sentences within our datasets never reaches a negative value, even though the cosine scale ranges from -1 to 1. Other approaches, such as NLI Entailment Probability, were affected by this problem to a much smaller extent.

In order to use span alignment in a meaningful way, it must therefore be penalized to an appropriate extent. We set different levels of penalties for all metrics in our experiments in order to achieve positive results in the end. However, since very few spans were aligned in the gold data and the models occasionally predicted incorrect span alignments, this option did not result in any significant improvements.

Thresholds for 1:0 Alignments: Another option for improving the quality of automatic alignments is to allow 1:0 alignments, i.e., cases where no matching element is found in the target text for a source element. The only way to implement this option is to define thresholds: based on the confidence scores, a threshold value can be set for each metric that must be exceeded in order for the alignment to be allowed.

However, since the confidence scores of the various metrics are not comparable with each other, they must be set individually for each model. In the course of our experiments, we also found that, ideally, the scores need to be adjusted in relation to the datasets. Table 4 shows the thresholds with which we achieved the best results for each model and dataset and it also shows the change in the f1 score compared to the values in Table 3. Additionally, we also tested GPT-5-Mini with the 1:0 alignment option. In this case, we adjusted the prompt instead of using a threshold.

It is striking that, for the Cosine/miniLM ap-

Model	Threshold	f1
Salinger		
Cosine / miniLM	0.5	0.85 (-0.01)
Cosine / LaBSE	0.565	0.94 (+0.04)
NLI	0.87	0.88 (+0.02)
GPT-5-Mini	-	0.96 (+0.06)
Arendt		
Cosine / miniLM	0.45	0.80 (-0.03)
Cosine / LaBSE	0.562	0.92 (+0.03)
NLI	0.85	0.81 (+ 0.03)
GPT-5-Mini	-	0.93 (+0.03)

Table 4: Results with thresholds for seleceted models.

proach, the application of thresholds did not show any improvement for either dataset. Looking at the alignments in detail, it can be seen that, although some correct 1:0 alignments are predicted, false negatives also occur. This suggests that the boundary between correct and incorrect alignments is very blurred in this model, i.e., very similar confidence scores can underlie both cases.

It is also noticeable that the thresholds for the Arendt text have to be set slightly lower for all models than for Salinger. This could be due to the fact that the Salinger text (presumably due to its lower sentence complexity) is easier to align overall and the confidence scores are correspondingly higher.

Efficiency: Finally, we will briefly address the issue of efficiency on the theoretical level. Unfortunately, we cannot report comparable data on computing time for all models, since we cannot control hardware use for the LLM approaches, and other factors like internet speed would further blur the results. However, even from a theoretical standpoint, significant differences between the individual metrics become obvious. While approaches that use cosine similarity can generally be implemented very efficiently, the NLI Entailment Probability and the BERTScore in particular are very expensive. The reason for this lies in the underlying logic of these models.

Cosine-based approaches (e.g., Sentence-BERT embeddings, LaBSE, LASER) follow a two-stage computational paradigm: First, each sentence is mapped independently into a fixed-dimensional vector space using a pre-trained encoder. This embedding step is performed once per segment and the resulting vectors can be cached. For multilingual sentence encoders such as LaBSE or LASER, the embedding model processes each segment independently, i.e., no interaction between candidate pairs occurs at this stage. Alignment scoring is then reduced to computing cosine similarity between vector representations, which is computationally inexpensive.

In contrast, the BERTScore does not operate on precomputed sentence embeddings. Instead, it takes raw text as input and performs contextual token-level comparisons using a full transformer model. Thus, for every candidate span, a complete forward pass through a large transformer model is required. When multiple candidate spans are evaluated per source segment, this cost multiplies accordingly. Even when batched, the system must process all candidate pairs jointly through the model, which remains computationally expensive relative to the cosine similarity computations.

NLI models introduce an even stronger computational coupling between candidate pairs, because for each alignment candidate the candidate pair is concatenated as a premise–hypothesis input. The combined sequence is processed jointly by a transformer, and a classification head then predicts probabilities for entailment, contradiction, or neutrality. Unlike independent sentence embeddings, NLI models rely on cross-attention between the two texts, and even when batched, the cost remains proportional to the number of candidate spans, since cross-text attention must be computed for each pair.

5.2. Error Analysis

As already mentioned in Section 3, the Salinger text is syntactically less complex than the Arendt essay. Furthermore, there are no `1:0` alignments in the gold data here, which means that very few errors are enforced by the basic settings described in Section 5.1. Almost all errors that do occur are caused by span alignments in the gold data. Since the English text contains 13 more sentences than the German version, several English target sentences sometimes have to be aligned with one German source sentence. In the basic settings, all models are limited to `1:1` alignments, but even when span alignment is allowed, almost all errors occur there. A good example of such a case is shown in Example 1. In this case, all models align only the longer, first target sentence to the source sentence and miss the second English sentence. However, this second sentence is very short and does not contain semantically strong words. The fact that it is not a literal translation makes the alignment even more difficult in this case.

In contrast, the Arendt essay contains more errors that are also more diverse. As mentioned

- Dafür hat Pencey einen guten Ruf als Schule, das muss man sagen.
  (Pencey does have a good reputation as a school, that has to be said.)
- It has a very good academic rating, Pencey.
- It really does.

Examples 1: Alignment error in Salingers *Catcher in the Rye*. Literal translations of the German sentences are provided in round brackets.

- Es handelt nur von den allerelementarsten Gliederungen, in die das Tätigsein überhaupt zerfällt, also von denjenigen, die der Überlieferung wie unserer eigenen Meinung zufolge offenbar innerhalb des Erfahrungshorizonts jedes Menschen liegen sollten.
  (It deals only with the most basic categories into which human activity can be divided, that is, those which, according to tradition and our own judgment, should clearly lie within the scope of every person's experience.)
- It deals only with the most elementary articulations of the human condition, with those activities that traditionally, as well as according to current opinion, are within the range of every human being.
- This, obviously, is a matter of thought, and thoughtlessness – the heedless recklessness or hopeless confusion or complacent repetition of "truths" which have become trivial and empty–seems to me among the outstanding characteristics of our time.

Examples 2: Alignment error in Arendts *The Human Condition*. Literal translations of the German sentences are provided in brackets.

above, there are 8 cases (11%) in which no match for a source element was found in the gold annotations. In two other cases (3.4%), spans of two elements per source element are annotated, even though in this case the source text is longer than the target text. These statistics again indicate a lower degree of parallelism compared to the Salinger text.

However, errors in the model predictions arise not only from these special cases, but also from some incorrect 1:1 alignments. Such an incorrect alignment is shown in Example 2.

In this case the German source sentence was aligned with the third sentence by several models, although it should actually be aligned with the second sentence. The two English sentences are only three index positions apart and are semantically quite closely related: words such as *thoughts*, *opinions*, *articulations*, *range*, etc. come from similar semantic fields, making alignment not unreasonable. The general length of the sentences and their structure are also quite similar. Nevertheless, other words without a counterpart in the German sentence (e.g., *heedless recklessness*, *hopeless confusion*) and formal peculiarities (e.g., the dash or the quotation marks around the word *truth*) should actually interfere with the alignment. It is possible that errors like this one are happening because of sentence length and complexity (the German source sentence contains 32 words), but since most sentences in the Arendt text are equally long and get aligned correctly, this alone may not be the problem.

6. Conclusion

This paper reports a case study to systematically test various existing metrics for the task of aligning semi-parallel bilingual texts. The results are promising: both GPT-5-Mini as a generative language model, as well as cosine similarity approaches and the NLI entailment probability achieve f1 scores well above 0.8. Our window logic helps to make alignment more efficient without imposing severe restrictions on the alignment options.

Nevertheless, these results can only be the starting point for further work on aligning semi-parallel texts. First of all, going beyond the present case study by running the experiments on larger datasets is a central next step.

From an empirical perspective, an interesting experiment would be to swap the source and target texts to see if this results in changes to the alignment decisions, even though our annotation guidelines call for consideration of semantic similarities in both directions.

Each similarity metric currently has its own problems, which have become particularly apparent in the area of 1:0 and 1:n alignment. Another important future task will be the systematic testing of ensemble scores to overcome the weaknesses of individual approaches. Further work is also needed in the area of alternative alignment levels: for example, preliminary alignment of paragraphs could help to further narrow down the search areas. Segmenting long sentences into clauses could also simplify difficult cases where parts of sentences overlap while other parts do not match.

Acknowledgments

We thank our student assistant Dietmar Benndorf for annotating training data, and we are grateful to the anonymous reviewers for their helpful feedback. Our work is supported by the Deutsche Forschungsgemeinschaft (DFG), project (524057241) "Semi-automatische Kollationierung verschiedensprachiger Fassungen eines Textes“.

A. LLM System Prompts

```
You align German sentences to English
    sentences.
Align ONLY the current German sentence
    labeled "DE".
If DE is a fragment or clause, still
    choose the best matching English
    candidate.
Choose ONLY from the candidate IDs below
    .

German sentence:
{de_instance}

English candidates:
{candidates}

Return JSON only, no extra text. en_span
     must reference English candidate
    IDs:
{{
  "type": "1-1",
  "en_span": [j, k],
  "confidence": 0.0-1.0
}}

Rules:
- Choose exactly ONE candidate ID from
    the list.
- Output must be type "1-1".
- en_span must be [j, j] with j from
    candidate IDs.
- Do not output "none" or "1-many".

Do not include explanations. Output must
     be a single JSON object.
```

Listing 1: System prompt used for strict alignment setting

```
You align German sentences to English
    sentences.
Align ONLY the current German sentence
    labeled "DE".
If DE is a fragment or clause, still
    choose the best matching English
    candidate(s).
Return "none" if no clear match exists
    in the candidates.
Choose ONLY from the candidate IDs below
    .

German sentence:
{de_instance}

English candidates:
{candidates}

Return JSON only, no extra text. en_span
     must reference English candidate
    IDs:
{{
  "type": "1-1" | "1-many" | "none",
  "en_span": [j, k],
  "confidence": 0.0-1.0
}}

Rules:
- If 1-1: en_span must be [j, j].
- If 1-many: en_span must be contiguous
    [j, k] with (k-j+1) <= {max_span}.
- If none: en_span must be [-1, -1].

Do not include explanations. Output must
     be a single JSON object.
```

Listing 2: System prompt used for non-strict alignment setting

B. References

Mikel Artetxe and Holger Schwenk. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019.

Elron Bandel, Ranit Aharonov, Michal Shmueli-Scheuer, Ilya Shnayderman, Noam Slonim, and Liat Ein-Dor. Quality Controlled Paraphrase Generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 596–609, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.45.

Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics*, ACL '91, page 169–176, USA, 1991. Association for Computational Linguistics. doi: 10.3115/981344.981366.

Percy Cheung and Pascale Fung. Sentence Alignment in Parallel, Comparable and Quasi-comparable Corpora. In *LREC Workshop on the amazing utility of parallel corpora*,

2004. URL https://api.semanticscholar.org/CorpusID:29228186.

Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. Chinese–Japanese Parallel Sentence Extraction from Quasi–Comparable Corpora. In Serge Sharoff, Pierre Zweigenbaum, and Reinhard Rapp, editors, *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 34–42, Sofia, Bulgaria, 2013. Association for Computational Linguistics. URL https://aclanthology.org/W13-2505/.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019. URL https://arxiv.org/abs/1810.04805.

Bonnie J. Dorr. Machine translation divergences: A formal description and proposed solution. *Computational Linguistics*, 20(4):597–633, 1994. URL https://aclanthology.org/J94-4004/.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT Sentence Embedding. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.62.

Steffen Frenzel and Manfred Stede. Sentence-Alignment in Semi-parallel Datasets. In Anna Kazantseva, Stan Szpakowicz, Stefania Degaetano-Ortlieb, Yuri Bizzoni, and Janis Pagel, editors, *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 87–96. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.latechclfl-1.9.

Steffen Frenzel, Maximilian Krupop, and Manfred Stede. Discourse segmentation of german text with pretrained language models. *Journal for Language Technology and Computational Linguistics*, 2026.

William A. Gale and Kenneth W. Church. A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, 19(1):75–102, 1993. URL https://aclanthology.org/J93-1004.

Darina Gold, Venelin Kovatchev, and Torsten Zesch. Annotating and analyzing the interactions between meaning relations. In Annemarie Friedrich, Deniz Zeyrek, and Jet Hoek, editors, *Proceedings of the 13th Linguistic Annotation Workshop*, pages 26–36, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4004.

Paul Jaccard. Etude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, page 547–579, 1901.

Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*. Pearson, 3rd edition, 2026. URL https://web.stanford.edu/~jurafsky/slp3/. Online manuscript released January 6, 2026.

Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert, 2020. URL https://arxiv.org/abs/2004.12832.

Timothy Liu and De Wen Soh. Towards Better Characterization of Paraphrases. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8592–8601, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.588.

Francesco Molfese, Andrei Bejgu, Simone Tedeschi, Simone Conia, and Roberto Navigli. CroCoAlign: A Cross-Lingual, Context-Aware and Fully-Neural Sentence Alignment System for Long Texts. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2209–2220, St. Julian's, Malta, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-long.135.

Robert C. Moore. Fast and accurate sentence alignment of bilingual corpora. In Stephen D. Richardson, editor, *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 135–144, Tiburon, USA, 2002. Springer. URL https://link.springer.com/chapter/10.1007/3-540-45820-4_14.

Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with bert, 2020. URL https://arxiv.org/abs/1901.04085.

Marcus Pöckelmann, André Medek, Jörg Ritter, and Paul Molitor. LERA—an interactive platform for synoptical representations of multiple text witnesses. *Digital Scholarship in the Humanities*, 38(1):330–346, 2022.

Sadaf Abdul Rauf and Holger Schwenk. Parallel sentence generation from comparable corpora for improved SMT. *Machine Translation*, 25(4): 341–375, 2011. ISSN 09226567, 15730573. URL http://www.jstor.org/stable/41487466.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, 2019. URL https://arxiv.org/abs/1908.10084.

Kurt Sier and Eva Wöckener-Gade. Paraphrase als Ähnlichkeitsbeziehung. Ein digitaler Zugang zu einem intertextuellen Phänomen. In *Platon Digital. Tradition und Rezeption*. Propylaeum, 2019.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment. In Ron Kaplan, Jill Burstein, Mary Harper, and Gerald Penn, editors, *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 403–411, Los Angeles, California, 2010. Association for Computational Linguistics. URL https://aclanthology.org/N10-1063/.

Steinthor Steingrimsson, Hrafn Loftsson, and Andy Way. SentAlign: Accurate and Scalable Sentence Alignment. In Yansong Feng and Els Lefever, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 256–263, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-demo.22.

Brian Thompson and Philipp Koehn. Vecalign: Improved Sentence Alignment in Linear Time and Space. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1136.

Christoph Tillmann. A Beam-Search Extraction Algorithm for Comparable Data. In *Annual Meeting of the Association for Computational Linguistics*, 2009. URL https://api.semanticscholar.org/CorpusID:7798552.

Jan Wahle, Bela Gipp, and Terry Ruas. Paraphrase Types for Generation and Detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, page 12148–12164. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.746. URL http://dx.doi.org/10.18653/v1/2023.emnlp-main.746.

Feng Wang, Yuqing Li, and Han Xiao. jina-reranker-v3: Last but not late interaction for listwise document reranking, 2025. URL https://arxiv.org/abs/2509.25085.

Krzysztof Wołk and Krzysztof Marasek. Unsupervised Construction of Quasi-comparable Corpora and Probing for Parallel Textual Data. In Aleksander Zgrzywa, Kazimierz Choroś, and Andrzej Siemiński, editors, *Multimedia and Network Information Systems*, pages 307–320, Cham, 2017. Springer International Publishing.

Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Haibo Zhang, Xue Zhao, Wenqing Yao, and Boxing Chen. GCPG: A General Framework for Controllable Paraphrase Generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4035–4047, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.318.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020. URL https://arxiv.org/abs/1904.09675.

A Comparative Study in Corpus Linguistics applied to Automatic Terminology Extraction

Mercè Vàzquez, Sergi Alvarez-Vidal, Antoni Oliver

Universitat Oberta de Catalunya, Universitat Autònoma de Barcelona, Universitat Oberta de Catalunya
Barcelona (Spain)
mvazquezga@uoc.edu, sergi.alvarez@uab.cat, aoliverg@uoc.edu

Abstract

Parallel and comparable corpora are the main linguistic resources to identify multilingual terminology using automatic term extraction tools. However, parallel corpora are available only for certain languages, domains and genres, and comparable corpora have some limitations when identifying corresponding terms. To implement a more effective selection of multilingual terminology, we compared the performance of using specialised parallel and comparable corpora applied to languages with various forms of capital in linguistic resources. This paper presents a comparative study in corpus linguistics in which we automatically identify terms in Catalan, Spanish and English in Legislation and Administrative Law using parallel corpora, comparable corpora and a combined methodology based on both types of corpora together with word embeddings. We observe that the combined methodology implemented obtains a higher number of terms than when working exclusively with parallel or comparable corpora. The evaluation of the results is performed using a terminological thesaurus as a gold standard. The new methodology presented in our study permits us to identify multilingual terminology in an effective way, especially in Catalan-Spanish languages.

Keywords: comparable corpora, parallel corpora, automatic terminology extraction, computational terminology

1. Introduction

Terms are elements within specialised documents that are used for the creation and enrichment of ontologies and dictionaries (Maynard & Ananiadou 2001; El-Sappagh et al. 2018; Durán-Muñoz, 2019), and terms are relevant for multitude of applications such as information retrieval (Lingpeng et al., 2005), machine translation (Haque et al., 2019; Michon et al., 2020; Moslem et al., 2023), and sentiment analysis (Mayorov et al., 2015). Since the nineties computational terminology has been widely developed through the availability of corpus linguistics and the contribution of different natural language processing (NLP) domains, such as terminology extraction, information retrieval, ontology building, machine translation or computer-aided translation (L'Homme et al., 1998).

In order to implement different ATE strategies to identify terms from specialised information, monolingual and parallel corpora - a corpus that contains source texts and their translations (McEnery & Xiao, 2007) - have become the main linguistic resources to automatically extract candidates to compile terminology from a specific domain to be manually supervised by linguists and terminologists (Kupiec, 1993; Gaussier, 1998; Ha et al., 2008; Macken et al., 2013; Haque et al., 2014; Baisa et al., 2015). However, parallel corpora are usually available only for certain languages, domains and genres, and term extraction tools have been developed exclusively for the needs of major European and non-European languages. Furthermore, compiling parallel corpora from authoritative sources of information for terminology extraction is a resource-intensive task, particularly for less-resourced languages in terms of data scarcity in specialised domains. Indeed, access to authoritative sources sometimes may be restricted, or may need the permission from authors, companies or institutions (Daille & Morin, 2005; Gornostay et al. 2012; Gurrutxaga et al. 2013; Rigouts Terryn et al., 2018). Consequently, comparable corpora – documents that are comparable in content and form in various degrees and dimensions across several languages or language varieties (Zweigenbaum, Rapp and Sharoff, 2024) – are an alternative for extracting domain-specific terms because it is much easier to collect data (Fung, &Yee, 1998; Rapp, 1999; Chiao & Zweigenbaum, 2002; Daille & Morin, 2005; Aker et al., 2013; Bouamor et al., 2013; Morin & Hazem, 2014; Hazem & Morin, 2016; Hazem & Morin, 2017). However, certain limitations have been observed when implementing automatic term extraction from comparable corpora related to the terms identification, as the information compiled in both languages is similar but not equivalent, due to comparable corpora are not aligned; the construction of a gold standard dataset to automate the task, and also the evaluation of candidates term extracted using comparable corpora (Rigouts Terryn et al., 2020; Rigouts Terryn, 2023).

In order to make a more effective selection of terminology from corpus linguistics, we analysed the performance of specialised parallel and comparable corpora applied to languages with various forms of capital in linguistic resources**.** We present a comparative analysis on corpus linguistics in which we automatically identify terms in Catalan, Spanish and English in Legislation and

Proceedings of the 19th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC 2026, pages 20–29
11 May 2026.

Administrative Law using parallel corpora, comparable corpora and a combined methodology based on both types of corpora together with word embeddings. This comparison aims to determine whether a particular corpora type is more suitable for processing using ATE tools. To conduct this analysis, we utilise TBXTools, an open-access term extraction tool capable of employing both statistical and linguistic term extraction methods and automatic search of translation equivalents of terms in corpora (Oliver and Vàzquez, 2015), to extract candidate terms from comparable and parallel corpora in the Legislation and Administrative Law in order to assess the reliability of the results obtained.

The primary objective of our study is to assess the performance and reliability of comparable corpora in comparison with parallel corpora to automatically identify multilingual terminology using a term extraction tool, particularly in languages with limited linguistic resources. This primary objective is based on two hypotheses. The first is that comparable corpora allows us to easily compile a higher volume of specialised information compared to parallel corpora, due to the fact that collecting original texts in more than one language for one domain is easier to compile than a collection of translated and aligned texts. The second is that comparable corpora can be an effective and reliable mechanism to compile specialised information for all domains, genres and languages, which is especially important for less-resourced languages.

To achieve this aim, this paper conducts a comparative analysis of results obtained using parallel corpora from the same domain to ascertain the effectiveness and reliability of Legislation and Administrative Law comparable corpora. The findings will enable us to determine whether employing comparable corpora (a) yields a larger corpus volume for terminology extraction, (b) improves terminology identification in languages with restricted linguistic resources such as Catalan, and (c) achieves a satisfactory level of terminological reliability.

The remainder of the present paper is structured as follows: in Section 2 we describe the background of parallel and comparable corpora applied to terminology extraction. In Section 3 we present the materials and tools used and the method implemented to compare the performance of comparable corpora with parallel corpora to identify multilingual terminology. The results and discussion are described in detail in Section 4. The paper is concluded with some final remarks and ideas for future research.

2. Materials, tools, and methods

With the aim of making a more effective selection of term equivalents from corpora using ATE tools, we analyse and compare the performance of specialised parallel and comparable corpora applied to automatic terminology extraction for two language pairs (English-Spanish and Catalan-Spanish) in one domain, Legislation and Administrative Law.

2.1 Materials

We have processed two parallel corpora to obtain parallel and comparable corpora of different sizes. For the English-Spanish language pair, we have used the DGT Corpus (Steinberger, 2013), and for the Catalan-Spanish, the DOGC corpus, a Catalan Spanish parallel corpus created from laws of the Catalan government (Oliver, 2022). The *Diari Oficial de la Generalitat de Catalunya*[1] (DOGC) is an official media outlet in which the laws and regulations of the Government of Catalonia are published. In Table 1 we can observe the size of these corpora once the segments have been deduplicated and shuffled. These corpora contain unique sentences.

Corpus	Segments	L1 tokens	L2 tokens
DGT unique eng-spa	3,640,761	73,883,784	84,702,886
DOGC unique cat-spa	8,472,786	188,929,206	197,986,300

Table 1: Tokens and segments included in the used corpora

From these corpora we have created one subset of parallel corpora: 1M segments. To create the comparable corpora from the parallel corpora we have selected 1M segments from the top of the corpus for the source language and from the bottom of the corpus for the target language. We have then a comparable corpus of 1M segments for each language pair.

For our study, we also need a set of source terms to find their translation, and the valid translation equivalents to perform the evaluation of the methodologies. We have used a subset of the Catalan IATE terminology glossary[2] consisting of 1,722 terms in English, Spanish and Catalan extracted from the Europarl Corpus (Koehn, 2005). More precisely, the glossary has 1,621 terms with equivalents in English and Spanish; and 1,232 with equivalents in Catalan and Spanish.

We have created subsets of this glossary with the terms present in all the created parallel and comparable corpora to evaluate the performance of terminology extraction using parallel corpora,

[1] https://dogc.gencat.cat/ca/inici/

[2] https://www.termcat.cat/en/diccionaris-en-linia/264

comparable corpora and also a combined methodology. From each source and target terms we also know the frequency of apparition of each source and target term. This data will help us to provide a more detailed analysis of the evaluation figures. In Table 2 we can see the number of source-target terms present in each subcorpus.

Language	Corpus	Size	Terms
eng-spa	parallel	1M	617
eng-spa	comparable	1M	852
cat-spa	parallel	1M	814
cat-spa	comparable	1M	845

Table 2: Number of terms present in the parallel and comparable corpora used in the experiments

2.2 Tools

To undertake the comparative analysis, we have used TBXTools, a Python class performing a series of methods for automatic term extraction and automatic search of translation equivalents of terms in parallel and comparable corpora. For the experimental part we have used the capabilities of TBXTools for automatic detection of translation equivalents of known terms in parallel and comparable corpora.

2.2.1 Detection of translation equivalents in parallel corpora

In TBXTools the automatic detection of translation equivalents in parallel corpora is performed in the following way: we have a set of terms in the source language (L1) and we want to know the translation equivalents of these terms in the target language (L2). We have a parallel corpus L1-L2 for the given subject. The algorithm takes one term in L1 and creates a L2 subcorpus with the target segments whose source segments contain the given term. Then the algorithm performs an ATE task (it can be either statistical or linguistic) on this L2 subcorpus. The most frequent L2 term candidate has big chances of being the translation equivalent of the given L1 term. We repeat this procedure for each term in the set of terms in L1.

In this strategy, the ATE process on the target corpus can be either statistical or linguistic. In our experiments, the statistical methodology has been used. One important parameter is the relation of n-gram order (*n*) between the source term and the target term candidate. For example, one uni-gram source term can have a uni-gram translation equivalent (as in *agreement - acuerdo*), a bi-gram (as in *bailiff - agente judicial*) and even a tri-gram (as in *affidavit - acta de notoriedad*). As these relations are not known in advance, several relations should be explored. Therefore, two parameters: maximum *n* increment (max_inc) and maximum *n* decrement (max_dec) should be set to better identify the translation equivalents.

This simple strategy works quite well when the L1 terms appear several times in the L1 part of the parallel corpus and most of the times the same translation equivalent is used in the L2 part of the parallel corpus. In the experimental part we present figures of precision, recall and F_1 for this strategy in several scenarios.

2.2.2 Detection of translation equivalents in comparable corpora

In TBXTools a method for automatic detection of translation equivalents in comparable corpora based on word embeddings, is implemented. In this methodology the embeddings for all terms in the list of terms to search is calculated using the source language part of the comparable corpus. As the source language terms are known, we can convert the complex terms, that is, terms formed by more than one word, into single tokens joining the components of the terms by some symbol, for example "_". In this way, a complex term as for example *interest rate* is converted into a single token *interest_rate*. We call this process *compoundifying*.

Hence, source language terms can be compoundified because they are known, as they are in the list of terms we want to find the translation equivalent from. Then, we need to calculate the embedding for the target terms using the comparable corpus for the target language. But now these target language terms are not known in advance, as we are precisely looking for these terms. This is not a problem for the simple terms, that is, for the terms formed by a single word, as we can calculate the embeddings for all the words in the target comparable corpus. But we don't know *a priori* the complex terms in the target language, so the algorithm performs an ATE process in the target comparable corpora to detect target term candidates in order to compoundify them. This ATE process, again, can be either statistical or linguistic. We implemented the statistical process.

Once we have the embeddings for the source terms and the target extracted terms, we have two different vector spaces that should be mapped. To do so, we use the vecmap[3] algorithm (Artetxe et al., 2018). Once the two vector spaces are mapped, the translation equivalent of a source language term should be the nearest target language term in the mapped vector space. But taking the nearest target language term is dangerous, as this target term can be closer to a different source term. For this reason, a margin score is calculated, as defined in Artetxe and Schwenk (2019) (changing the sentence embeddings by word embeddings): the margin score between two candidate term equivalents *x* and *y* is defined as the ratio between the cosine distance between the two word embeddings, and the average cosine similarity of its nearest neighbors in both directions. This strategy, however, fails in the cases where the translation

[3] https://github.com/artetxem/vecmap

equivalent is not present in the target comparable corpus. For this reason, a minimum margin score should be defined to reject the equivalents detected with a lower margin score.

2.3 Methods

The methods we implemented to identify a more effective procedure to select term equivalents from corpora are based on parallel corpora, comparable corpora and a combined methodology which combines both types of corpora.

2.3.1 Parallel corpora

In this strategy, as stated above, two important parameters must be set: max_dec and max_inc. When searching for the translation equivalent of a term, the translation equivalent might have the same number of tokens, or a different number. For example, the translation of a bigram term can be a unigram (max_dec=1) or a trigram (max_inc=1). Depending on the language pair in the experiments, these parameters can be set with different values. To set these parameters in our experimental setting we have performed a statistical analysis using the complete Catalan IATE e-dictionary (Vàzquez, Oliver, 2019) from Termcat's Terminologia Oberta. This e-dictionary contains 15,997 terms in Catalan, Spanish, English and French. From this e-dictionary we have analyzed all the English-Spanish and Catalan-Spanish pairs of terms to count for the increments and decrements in the n-gram relation between source and target term. In Table 3 we can observe the results of this analysis, and we can set for English-Spanish max_dec=1 and max_inc=1 and for Catalan-Spanish we can set max_dec=0 and max_inc=0.

Increment	% English-Spanish	% Catalan-Spanish
-3	2,16	2.95
-2	4.75	6.21
-1	10.46	4.32
0	60.04	69.55
1	12.88	5.21
2	4.04	3.2
3	1.1	1.1

Table 3: Results of the statistical analysis to set the max_dec and max_inc parameters

For each source term the algorithm provides a set of translation equivalents sorted by a confidence score, being the first candidate the one with more chances to be the correct one.

2.3.2 Comparable corpora

To find the translation equivalents of terms in comparable corpora, we need to perform the following processes:

1. Compoundify the source language terms in the source language comparable corpus. This process can be performed as the source language terms are known in advance to evaluate the performance of terminology extraction using comparable corpora.
2. Compoundify the target language terms in the target language comparable corpus. As the target language terms are not known in advance, we should make an unsupervised ATE process in the target corpus to get a set of term candidates to compoundify.
3. Calculate the source language word embeddings using the compoundified source language comparable corpus.
4. Calculate the target language word embeddings using the compoundified target language comparable corpus
5. Map the source and target language embeddings.
6. Extract the target term candidate for each source language term, using the margin score.

2.3.3 Combined methodology

In the combined methodology the translation equivalent candidates are obtained using the parallel corpora method, but the translation equivalents will be sorted using mapped word embeddings calculated by concatenating the source part of the parallel corpus with the comparable corpus for the source language, and the target part of the parallel corpus with the comparable corpus for the target language. The steps performed in this methodology are the following:

1. Perform the search of translation equivalents using the parallel corpus method.
2. Concatenate the source part of the parallel corpus and the comparable corpus for the source language.
3. Concatenate the target part of the parallel corpus and the comparable corpus for the target language.
4. Compoundify the concatenated corpus for the source language using the source terms we want to search for.
5. Compoundify the concatenated corpus for the target language using all the translation equivalents candidates obtained in the first step.
6. Calculate the source language embeddings using the compoundified source language concatenated corpus.
7. Calculate the target language embeddings using the compoundified target language concatenated corpus.
8. Map the source and target language embeddings.
9. Resort the translation equivalents calculated in step one calculating the margin score.

The combined methodology has the additional advantage that the target language corpora can be compoundified without the need of performing an unsupervised ATE process. Instead, we can compoundify the target language corpora using all the translation equivalents candidates obtained using the parallel corpus methods, as these candidates are precisely the ones we want to resort with the margin score.

3. Results

In this section we present the evaluation results for three tasks: automatic translation equivalent detection in parallel corpora, in comparable corpora and a combined methodology using parallel corpora and embeddings calculated from the parallel corpus and a comparable corpus to resort the candidates. These methodologies are tested for 1M segments of the corpora and two language pairs: English-Spanish and Catalan-Spanish. We show the results obtained with 1M corpus segments.

3.1 Parallel corpora

3.1.1 English-Spanish

The evaluation figure for the automatic detection of translation equivalents in parallel corpora for English-Spanish and corpus size of 1M segments are shown in Table 4. For this corpus size we experimented with two sets of max_dec=0 and max_inc=0, and max_dec=1 and max_inc= 1, which offers higher results. In the Table we can observe the precision (P), recall (R) and F_1 for the first translation candidate, and for the cases where the correct candidate is among the first 5 candidates (P 5, R 5 and F_1 5), and among the top-ten candidates (P 10, R 10 and F_1 10). We also present figures for those source terms appearing at least 1, 2, 5, and 10 times in the corpus. As a general behavior, the most frequent the source term is, the higher the precision. But as the recall has been calculated regardless the frequency of apparition, the recall and F_1 score drop drastically.

Another general and obvious behavior is that the P5 and P10 (the precision taking into account the top 5 or 10 candidates) is higher than P1 (the precision taking into account only the first candidate). But it is worth knowing P5 and P10, because it simulates the practical case where a terminologist is presented with the list of candidates to choose the correct one.

An interesting conclusion from Table 4 is that enlarging the size of the parallel corpora does not improve the results. Enlarging the size of the parallel corpora causes some source terms to appear with a higher frequency, but some new source terms with lower frequency are also included in the experiment, yielding no improvement.

Freq.	P	R	F_1	P 5	R 5	F_1 5	P 10	R 10	F_1 10
1	21.88	21.88	21.88	42.46	42.46	42.46	50.57	50.57	50.57
2	26.08	21.56	23.6	50.39	41.65	45.61	58.43	48.3	52.88
5	32.44	19.61	24.44	57.64	34.85	43.43	64.61	39.06	48.69
10	36.77	17.34	23.57	62.89	29.66	40.31	70.1	33.06	44.93

Table 4. Evaluation figures for parallel corpus 1M segments English-Spanish with max_dec=1 and max_inc=1

3.1.2 Catalan-Spanish

The results for the automatic detection of translation equivalents in parallel corpora for Catalan-Spanish in 1M segments corpora are presented in Table 5. For this language pair we have only considered max_dec and max_inc of 0. The first thing we notice is that the results for Catalan-Spanish are much better than for English-Spanish. This can be explained by different causes. The 0 value of max_inc and max_enc covers a larger percentage of cases for Catalan-Spanish than for English-Spanish. The fact that Catalan-Spanish are more similar than English-Spanish should have no direct influence on the results, as no linguistic information is used. But the similarity between languages may cause a more consistent use of translation equivalents in the corpus, making them easier to detect.

Freq.	P	R	F_1	P 5	R 5	F_1 5	P 10	R 10	F_1 10
1	44.72	44.72	44.72	78.26	78.26	78.26	83.78	83.78	83.78
2	48.25	44.1	46.08	83.74	76.54	79.97	88.04	80.47	84.08
5	51.89	40.54	45.52	88.05	68.8	77.24	91.19	71.25	80.0
10	53.61	36.49	43.42	88.81	60.44	71.93	91.88	62.53	74.42

Table 5. Evaluation figures for parallel corpus 1M segments Catalan-Spanish with max_dec=0 and max_inc=0

3.2 Comparable corpora

3.2.1 English-Spanish

The evaluation results for comparable corpora for the English-Spanish pair are now presented. As explained in section 2.3.2, one important step in this methodology is the compoundifying of complex terms, that is, converting the terms formed by more than one word, into a single token, replacing the blank spaces by a "_". As commented, this process can be done for the source terms, as they are already known. But target terms are still not known, so we cannot directly compoundify them.

We present two cases:

1. False compoundifying, where we cheat and use the list of known target terms used for evaluation. The results obtained are higher than in a real situation, but allow us to assess the capability of mapped word embeddings to find translation equivalents.
2. Statistical compoundifying, where an unsupervised statistical term extraction has been performed on the target comparable corpus. We then take all the target term candidates and we use them to compoundify. In Table 6 we can observe the overall number of term candidates and the number of terms with a frequency of 5 or higher, used in the compoundifying process.

	Term candidates	
Corpus size	freq. >=1	freq. >=5
1 M	943,544	238,005

Table 6. Number of term candidates of the automatic term extraction for compoundifying the Spanish comparable corpora in the English-Spanish experiments

False compoundifying

In Table 7 we can observe the evaluation results for the English-Spanish experiments using comparable corpora, but doing the cheating of compoundifying the target Spanish terms, for the size of 1M segments. The results for 1M segment corpus are low, but with about 10 points of increment in precision for the first candidate and for the first 5 candidates; and up to 20 points for the first 10 candidates. Still, however, the precision results are not enough for fully automatic tasks, but they can be reliable enough for manual tasks performed by terminologists to provide a set of suggestions.

But we must remember that the results from Table 7 are obtained using the known Spanish translation candidates to compoundify the target corpus. In most real situations this is not doable, as these Spanish translation equivalents are still unknown. Only the case when a terminologist performs an ATE task and a manual revision of the term candidates, and then tries to search the relation with the set of source English terms would be similar to this experimental setting.

Freq.	P	R	F_1	P 5	R 5	F_1 5	F 10	R 10	F_1 10
1	13.02	12.91	12.96	18.46	18.31	18.39	40.24	39.91	40.07
2	13.02	12.91	12.96	18.46	18.31	18.39	40.24	39.91	40.07
5	15.44	15.33	15.39	21.95	21.8	21.88	47.73	47.4	45.57
10	17.38	17.27	13.33	24.24	24.09	24.17	52.79	52.46	52.62

Table 7. Evaluation results for the comparable 1M corpora using false compoundifying

Compoundifying with statistical ATE

In Table 8 the results of the search of translation equivalents using comparable corpora, and compoundifying using an unsupervised statistical ATE task on the target Spanish comparable corpus are presented. This experimental setting is more similar to a real practical situation. These results cannot be used in a full automatic setting, but can be presented as an aid to a terminologist.

Freq.	P	R	F_1	P 5	R 5	F_1 5	P 10	R 10	F_1 10
1	4.97	4.93	4.95	9.59	9.51	9.55	24.38	24.18	24.28
2	4.97	4.93	4.95	9.59	9.51	9.55	24.38	24.18	24.28
5	5.95	5.91	5.93	11.47	11.39	11.43	29.18	28.97	29.08
10	6.7	6.66	6.68	12.92	12.84	12.88	32.54	32.33	32.43

Table 8. Evaluation results for the comparable 1M comparable corpora using unsupervised statistical ATE for compoundifying

3.2.2 Catalan-Spanish

In this section we present the evaluation of the task on comparable corpora for Catalan-Spanish. As we have done for English-Spanish, we present the results with two compoundifying processes: firstly, doing the cheating of using the already known Spanish terms; and secondly, performing an unsupervised statistical ATE process on the Spanish comparable corpora, and using the term candidates with a frequency equal or higher than 5 to compoundify the target corpus. In Table 9 we can see the number of extracted terms and those used for compoundifying.

	Term candidates	
Corpus size	freq. >=1	freq. >=5
1 M	759,116	209,872

Table 9. Number of term candidates of the automatic term extraction for compoundifying the Spanish comparable corpora in the English-Spanish experiments

False compoundifying

In this setting, the results shown in Table 10 are much better than the results for English-Spanish (see Table 7). For the precision of the first candidate we obtain an improvement of 21.23 for the 1M segments corpora. For Catalan-Spanish we get precisions of around 90% with very good F_1 scores if we take into account the top-ten candidates for source terms appearing at least 10 times. But we must keep in mind that the compoundifying step has been performed with the cheating of using the Spanish terms in the evaluations set, and these good results will not be obtained in a real situation.

Compoundifying with statistical ATE

Now, in a more realistic situation where the compoundifying step has been performed through unsupervised statistical ATE (Table 11), the precision values drop drastically, but they are much better than the obtained for the English-Spanish corpora (see Table 8), with an improvement of 16.92 precision points for the 1M segments corpora.

3.3 Combined method using word embeddings

3.3.1 English-Spanish

From the results presented so far, we can see that the precision values obtained using the methodology based on parallel corpora is much higher than those based on comparable corpora. The main problems for the methodology based on parallel corpora are, on one hand, to determine the translation equivalent for terms with very low frequency; and on the other hand, to know the n-gram relation between the source term and the target term. In this section we explore the use of mapped word embeddings to resort the list of translation equivalents. In this combined methodology we use both the parallel and the comparable corpora to calculate the word embeddings. Then, for each source term, we get the list of translation equivalent candidates using the parallel corpus. Once we get the list of the *n* best candidates, we resort them using the margin score calculated with the source and target word embeddings.

In Table 12 we can observe the results using the parallel and comparable corpora with 1M segments. These results should be compared with the results presented in Table 4. Note that we are using the worst values of max_dec and max_inc parameters, as the embeddings will do the job of selecting the correct translation equivalent regardless of the *n* order relation. Comparing these two tables we can see that this reordering methodology using word embeddings is very productive for the first candidates. If we analyze the results taking into account the precision of the first candidate (P), we get an improvement of 19.1 points for a frequency of 1 (terms appearing at least one time), an improvement of 16.14 points for frequency of 5, and 15.13 for a frequency of 10. These figures drop when considering the first 5 candidates (P5) to 9.83, 4.71 and 3.78 respectively. But when we observe the results for the top-ten candidates (10), we improve a little (3.32 points), but get worse results for frequency 5 (-1.86 points) and frequency 10 (-3.87 points). This is explainable because resorting a large set of candidates has no effect if we take all of them to calculate the precision. So the results seem to indicate that the methodology can be very productive to select the best candidate into the first position when considering a limited number of candidates.

Freq.	P	R	F_1	P 5	R 5	F_1 5	P 10	R 10	F_1 10
1	34.25	32.43	33.31	37.38	35.38	36.35	75.75	71.72	73.68
2	34.24	32.43	33.31	37.38	35.38	36.35	75.75	71.72	73.68
5	38.32	36.53	37.41	41.68	39.73	40.68	84.48	80.83	82.46
10	40.94	38.99	39.94	43.84	71.74	42.76	88.58	84.35	86.41

Table 10. Evaluation results for the comparable 1M corpora using false compoundifying for Catalan-Spanish

Freq.	P	R	F_1	P 5	R 5	F_1 5	P 10	R 10	F_1 10
1	21.89	19.76	20.77	27.92	25.21	26.49	59.24	53.49	56.22
2	21.89	19.76	20.77	27.92	25.21	26.49	59.24	53.49	56.22
5	24.13	22.13	23.09	30.81	28.27	29.49	65.12	59.73	62.31
10	26.06	24.06	25.02	33.12	30.58	31.8	69.54	64.2	66.77

Table 11. Evaluation results for the comparable 1M comparable corpora using unsupervised statistical ATE for compoundifying

Freq.	P	R	F_1	P 5	R 5	F_1 5	P10	R 10	F_1 10
1	40.98	21.72	28.39	52.29	27.71	36.23	53.82	28.53	37.29
2	44.0	21.39	28.79	56.0	27.23	36.64	57.67	28.04	37.73
5	48.58	19.45	27.78	62.35	24.96	35.65	62.75	25.12	35.88
10	51.9	17.67	23.36	66.67	22.69	33.86	67.14	22.85	34.1

Table 12. Evaluation figures for parallel corpus 1M segments English-Spanish with max_dec=1 and max_inc=1 combined methodology

3.3.2 Catalan-Spanish

This combined methodology is also very productive for the Catalan-Spanish pair. If we observe the results in Table 13 for the 1M segments corpora and compare the results in Table 5 for the parallel corpus methodology for the same corpora size, we see an improvement of 31.85 precision points for the first candidate of

frequency 1 or higher. As in this combined methodology the results of the parallel corpus methodology are resorted with the mapped word embeddings calculated with both the parallel and comparable corpora, the improvements are much higher for the first position, dropping to 4,06 points for the first five candidates and yielding to no improvement for the top-ten candidates.

Freq.	P	R	F_1	P 5	R 5	F_1 5	P 10	R 10	F_1 10
1	76.57	71.87	74.14	82.33	77.27	79.72	83.77	78.62	81.12
2	81.43	70.02	75.3	87.14	79.94	80.58	88.43	76.04	81.77
5	85.26	63.27	72.64	90.23	66.95	76.87	91.06	67.57	77.57
10	86.34	55.9	67.86	91.08	58.97	71.59	91.65	59.34	72.04

Table 13. Evaluation figures for parallel corpus 1M segments Catalan-Spanish with max_dec=0 and max_inc=0 combined methodology

If we now compare the improvements of the combined methodology for Catalan-Spanish and English-Spanish (comparing Table 12 with Table 13), we can observe higher improvements for Catalan-Spanish (31.85 precision points vs. 19.1 points for the first position and frequency 1).

4. Conclusions and perspectives

In this article, we have presented a novel usage of parallel and comparable corpora to effectively identify multilingual terminology from specialised domains. The methodology combines the information contained in parallel corpora and comparable corpora related to a specific domain and introduces a mapped word embeddings procedure to effectively identify term equivalents from specialised corpora. In order to determine the reliability of the method, especially addressed to less-resourced languages that suffer from a lack of available linguistic resources to build parallel corpora, we have conducted a comparative analysis on corpus linguistics in which we automatically identify terminology in Catalan, Spanish and English in Legislation and Administrative Law using parallel corpora, comparable corpora and a combined methodology based on both types of corpora together with word embeddings. The evaluation results applied in two language pairs (English-Spanish and Catalan-Spanish) in the domain of Legislation and Administrative Law shows that combining parallel and comparable corpora to identify terminology from specialised domains outperforms those using parallel corpora or comparable corpora separately. We have used a terminological glossary manually compiled as a gold standard from the Catalan IATE e-dictionary to evaluate the reliability of the method applied. The promising results obtained contribute to expanding the methodology applied in corpus linguistics to maximise the terminology compilation, which has a relevant impact in the context of less-resourced languages with a lack of corpus linguistics availability.

The novel usage of corpus linguistics has been implemented in TBXTools, an open-access term extraction tool created to automatically identify multilingual terminology from specialised domains. The present methodology can be used in any other specialised area that has similar resources to identify terminology.

The present research provide a promising perspective in terminology identification with a novel usage of corpus linguistics with the aim to provide a larger volume of corpora for terminology extraction, especially relevant in the context of languages with limited linguistic resources; enhance terminology detection, and achieve a satisfactory level of reliability in the extracted terminology.

As a future work, we plan to introduce general-domain data to improve translation terms identification, due to general content completing the information given for each term candidates in source and target corpora. And also we plan to evaluate the performance of the methodology applied in other domains together with other evaluation methods.

5. Acknowledgments

This work is supported by the project TamTAS PCI2025-167063-2, funded by MICIU/AEI/10.13039/501100011033 and European Union in the Chist-era call 2025 Science in your own language

6. Bibliographical References

Aker, A., Paramita, M. and Gaizauskas. R. (2013). Extracting bilingual terminologies from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (pp 402–411). Sofia, Bulgaria. Association for Computational Linguistics.

Artetxe, M. Labaka, G. and Agirre. E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (pp 789–798).

Artetxe, M., and Schwenk. H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the association for computational linguistics*, *7*, 597–610.

Baisa, V., Ulipová, B. and Cukr. M. (2015). Bilingual terminology extraction in Sketch Engine. *9th Workshop on Recent Advances in Slavonic Natural Language Processing*, 61–67.

Bouamor, D., Semmar, N. and Zweigenbaum. P. (2013). Context vector disambiguation for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (pp 759–764).

Castor, A. and Pollux, L. E. (1992). The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.

Chiao, Y. C., and Zweigenbaum. P. (2002). Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics* (pp 1–5). Association for Computational Linguistics.

Daille, B., and Morin. E. (2005). French-English terminology extraction from comparable corpora" *International Conference on Natural Language Processing*, 707–718. Berlin, Heidelberg: Springer Berlin Heidelberg.

Durán-Muñoz, I. (2019). Methodological Proposal to Build a Corpus-Based Ontology in Terminology" *Lingue e Linguaggi*, *29*, 581–597.

El-Sappagh, S., Franda, F., Ali, F., and Kwak K. S. (2018). SNOMED CT standard ontology based on the ontology for general medical science. *BMC medical informatics and decision making*, *18*, 1–19.

Fung, P., and Yee, L. Y. (1998). An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th International Conference on Computational Linguistics* (pp 414–420).

Gaussier, E. (1998). Flow network models for word alignment and terminology extraction from bilingual corpora. In *Proceedings of the 17th International Conference on Computational Linguistics* (pp 444–450).

Gornostay, T., Ramm, A., Heid, U., Morin, E., Harastani R., and Planas E. (2012). Terminology Extraction from Comparable Corpora for Latvian. *HLT 2012: 5th International Conference Human Language Technologies*, 66–73. Estonia.

Gurrutxaga, A., Leturia, I., Saralegi, X., and Vicente, I. S. (2013). Automatic comparable web corpora collection and bilingual terminology extraction for specialized dictionary making. *Building and using comparable corpora*, 51–75.

Ha, L. A., Fern, G., Mitkov, R., and Corpas, G. (2008). Mutual bilingual terminology extraction" In *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (LREC'08) (pp 1818–1824).

Haque, R., Penkale, S., and Way, A. (2014). Bilingual termbank creation via log-likelihood comparison and phrase-based statistical machine translation. In *Proceedings of the 4th International Workshop on Computational Terminology* (Computerm) (pp 42–51).

Haque, R, Hasanuzzaman, Md., and Way, A. (2019). Investigating Terminology Translation in Statistical and Neural Machine Translation: A Case Study on English-to-Hindi and Hindi-to-English. *Proceedings of the International Conference on Recent Advances in Natural Language Processing* (RANLP 2019) (pp 437–446). Varna, Bulgaria.

Hazem, A., and Morin, E. (2016). Efficient Data Selection for Bilingual Terminology Extraction from Comparable Corpora. *26th International Conference on Computational Linguistics* (COLING), 3401–3411. Osaka, Japan.

Hazem, A., and Morin, E. (2017). Bilingual word embeddings for bilingual terminology extraction from specialized comparable corpora. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing* (pp 685–693).

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. *The 10th Machine Translation Summit Proceedings of Conference*. 79–86. International Association for Machine Translation.

Kupiec, J. (1993). An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics* (pp 17–22). Association for Computational Linguistics.

L'Homme, M.-C., Bourigault, D., and Jacquemin, C. (1998). *First Workshop on Computational Terminology* (COMPUTERM'98), Montréal, Canada.

Lingpeng, Y., Donghong, J., Guodong, Z., and Yu, N. (2005). Improving Retrieval Effectiveness by Using Key Terms in Top Retrieved Documents. In Losada, D.E., Fernández-Luna, J.M. (Eds.) Advances in Information Retrieval. ECIR 2005. Lecture Notes in Computer Science, *3408* (pp 169–184). Springer, Berlin, Heidelberg.

Macken, L., Lefever, E., and Hoste, V. (2013). ExSIS: Bilingual Terminology Extraction from Parallel Corpora Using Chunk-Based Alignment. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *19*(1), 1–30.

Maynard, D., and Ananiadou, S. (2001). TRUCKS: A model for automatic multi-word term recognition. *Journal of Natural Language Processing*, *8*(1), 101–125.

Mayorov, V., Andrianov, I., Astrakhantsev, N., Avanesov, V., Kozlov, I., and Turdakov. D. (2015). A High Precision Method for Aspect Extraction in Russian. *Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference "Dialogue."*, *2*, 34–43. Moscow, Russia.

McEnery, A., and Xiao, R. Z. (2007). Parallel and comparable corpora: What are they up to. *Incorporating corpora: Translation and the linguist. Translating Europe. Multilingual matters*, 1–13.

Michon, E., Crego, J., and Senellart, J. (2020). Integrating Domain Terminology into Neural Machine Translation. In P*roceedings of the 28th International Conference on Computational Linguistics* (pp 3925–3937). Barcelona, Spain. International Committee on Computational Linguistics.

Morin, E., and Hazem, A. (2014). Looking at unbalanced specialized comparable corpora for bilingual lexicon extraction. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp 1284–1293).

Moslem, Y., Romani, G., Molaei, M., Kelleher, J. D., Haque, R., and Way, A. (2023). Domain Terminology Integration into Machine Translation: Leveraging Large Language Models. In *Proceedings of the 8th Conference on Machine Translation* (pp 902–911). Singapore. Association for Computational Linguistics.

Oliver, A. 2022. El corpus paral·lel del Diari Oficial de la Generalitat de Catalunya. *Linguamática*, 2023, 14 (2).

Oliver, A., and Vàzquez. M. (2015). TBXTools: A Free, Fast and Flexible Tool for Automatic Terminology Extraction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing* (pp 473–479).

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics* (pp 519–526).

Rigouts Terryn, A., Hoste, V., and Lefever, E. (2020). In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora. *Language Resources and Evaluation*, *54*(2), 385–418.

Rigouts Terryn, A. (2023). Supervised Feature-based Classification Approach to Bilingual Lexicon Induction from Specialised Comparable Corpora. In *Proceedings of the Workshop on Computational Terminology in NLP and Translation Studies (ConTeNTS) Incorporating the 16th Workshop on Building and Using Comparable Corpora (BUCC)* (pp. 59–68).

Steinberger, R., Eisele A., Klocek, S., Pilos, S. and Schlüter, P. (2013). *DGT-TM: A freely available translation memory in 22 languages. European Commission*.

Vázquez, M., Oliver, A., and Casademont, E. (2019). Using open data to create the Catalan IATE e-dictionary. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication, 25*(2) (pp. 175–197).

Zweigenbaum, P., Sharoff, S., & Rapp, R. (2024). Preface. In *Proceedings of the 17th Workshop on Building and Using Comparable Corpora*. LREC-COLING-2024.

Comparable corpora in cross-linguistic research: Nominal number in English, Czech, and Greek

Konstantinos Diamantopoulos, Magda Ševčíková
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, Prague, Czech Republic
{diamantopoulos, sevcikova}@ufal.mff.cuni.cz

Abstract

The paper examines the use of comparable corpora for contrastive research on the category of nominal number across three languages—English, Czech, and Greek. Two objectives are pursued: a cross-linguistic analysis of number and an assessment of the impact of automatic annotation on linguistic findings. For this study, corpora of comparable size and composition were compiled for the three languages from the Leipzig Corpora Collection. The data were automatically annotated using two open-access tools, Stanza and UDPipe, producing six datasets (two per language), each containing about 5 million sentences and 100 million tokens. Although derived from the same source, the paired datasets for each language differ in sentence and word segmentation, in the number of nouns identified, and in the number values assigned. These differences, nevertheless, do not appear to substantially affect the overall picture of number in the languages examined. The distribution of lemmas by the ratio of singular and plural forms challenges the view commonly presented in grammars that most nouns occur in both numbers and that singular-only and plural-only nouns are rare. However, a closer analysis of nouns assumed to have defective number indicates that answers to more nuanced questions vary depending on the annotation tool used.

Keywords: number, comparable corpora, morphosyntactic annotation, English, Czech, Greek

1. Introduction

The comparison of morphological categories across languages has long been central to contrastive and typological research. Information on morphological features has been collected for many languages in typological databases. However, recent advances in cross-linguistic research on word order have shown that the verification of distinctions recorded in such databases against corpus data leads to a more realistic picture (e.g. Choi et al. 2021; Levshina et al. 2023; Jing et al. 2023). In this respect, cross-linguistic research on morphology has lagged behind syntactic research in using corpus data, possibly due to the limited availability of suitable resources.

The present paper examines the morphological category of number in nouns in three languages with different morphological profiles, namely English, Czech, and (Modern) Greek. The languages were selected based on the availability of the required resources and tools, and on the availability of native speakers' judgments, ensuring full control over both the data and their interpretation. In order to avoid the problems that monolingual corpora (varied size, time period covered, etc.) and parallel corpora (translationese, etc.) pose for cross-linguistic comparisons, we compiled corpora of comparable size and composition for the three languages from the Leipzig Corpora Collection. However, the aim is not only to compare the category of number, but also to assess the potential of automatically annotated data for such research. To this end, the corpora for each language are annotated using two open-access tools Stanza and UDPipe.

The paper is structured as follows. Section 2 outlines basic facts about the category of nominal number, first as a cross-linguistically attested category and then with specific reference to the three languages examined. Section 3 describes the construction and morphosyntactic annotation of the comparable corpora and discusses in particular the differences between the paired datasets for each language. Section 4 presents the quantitative and qualitative analysis of the data aimed at identifying both language-internal and cross-linguistic patterns in the use of number. The results of the study are summarized in Section 5.

2. Nominal Category of Number

2.1. Cross-linguistic attestation

Grammatical number is a fundamental morphological category across the world's languages. However, as Corbett (2000, p. 2) notes, although it is often regarded as a simply structured category with singular and plural values and overt marking on nouns, one or more of these assumptions may not hold universally. Typological databases provide a broad-coverage view of some aspects: Of the 291 languages for which the World Atlas of Language

Proceedings of the 19th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC 2026, pages 30–40
11 May 2026.

Structures (WALS; Dryer and Haspelmath 2013) reports on plural marking, 28 lack nominal plural, and several dozen mark plurality only to a limited extent, for example only on human nouns (Feature 34A). The Grambank database (Skirgård et al., 2023) identifies plural marking on nouns in 1,282 of 2,389 languages (Feature GB044); the other languages may express plurality, for example, by means of a free-standing marker, noun reduplication, or not at all (cf. Feature 33A in WALS).

2.2. Number in grammars of English, Czech, and Greek

English, Czech, and Greek are Indo-European languages in which, despite their differing morphological profiles, the category of number is structured in a similar way. In the three languages, number in nouns is primarily realized as a binary opposition between singular and plural. English marks number most commonly by the suffix *-s* (as in *leg – legs*), with limited inflectional variation and a modest set of irregular forms (e.g. *foot – feet*; cf. Quirk et al. 1985; Huddleston and Pullum 2002; Bauer et al. 2013, among others).

Czech exhibits a morphologically rich inflectional system with numerous noun inflectional classes employing different formal markers to express singular and plural, while also retaining residual traces of the historical dual, preserved in a small group of nouns, especially those referring to parts of the body (e.g. Komárek et al. 1986; Havránek and Jedlička 2002). Number is realized jointly with morphological case in a single inflectional (portmanteau) ending; cf. selected forms of the noun *noha* 'foot': *noh-a* foot-NOM.SG, *noh-ám* foot-DAT.PL, *noh-ama* foot-INSTR.DUAL.

Greek is likewise fusional, with nominal number and case expressed cumulatively in portmanteau endings. Although it distinguishes singular and plural across several declensional classes, the inventory of formal number markers is comparatively smaller and more regular than in Czech (cf. Τριανταφυλλίδης 1979; Τζεβελέκου et al. 2007; Holton et al. 2012 (Triantaphillídis 1979, Tzevelékou 2007), or Χατζησαββίδης and Χατζησαββίδου 2014 (Khatzisavvídis and Khatzisavvídou 2014), or Κλαίρης and Μπαμπινιώτης 2010 (Klaíris and Bambiniótis 2010). Cf. selected forms of the noun *πόδι* (pód-i) 'foot', with portmanteau markers delimited: *πόδ-ι* (pód-i) foot-NOM/ACC.SG, *ποδ-ιού* (pod-ioú) foot-GEN.SG, *πόδ-ια* (pód-ia) foot-NOM/ACC.PL, *ποδ-ιών* (pod-ión) foot-GEN.PL.

2.3. Singularia and pluralia tantum

Although the grammars of individual languages follow different traditions, the assumption that "[m]ost nouns have both singular and plural" (Huddleston and Pullum 2002, p. 340), recurs, to varying degrees of explicitness, across them (cf. Komárek et al. 1986, p. 45 or Holton et al. 2012, Chapter 2). Nouns occurring exclusively in the singular or in the plural are typically treated as peripheral cases. These nouns are called singularia tantum, singular-only nouns, or singular invariable nouns for the first group, and pluralia tantum, plural-only nouns, or plural invariable nouns for the second group, and their precise delimitation may vary not only between languages, but also between individual works on a single language (Corbett, 2019; Acquaviva and Gardelle, 2023).

For English, Quirk et al. (1985, pp. 297–318), for example, assesses the number of nouns primarily on the basis of syntactic behavior, meaning that nouns ending in *-s* are also classified as singularia tantum, such as the names of disciplines (*acoustics, linguistics*) or diseases (*measles, rickets*) that are used with a verb in the singular form. In contrast, Bauer et al. (2013, p. 124) favor morphological criteria and classify the names of diseases as pluralia tantum due to the presence of the ending *-s* and the lack of forms without this ending.

Grammars of Czech, as well as those of Greek, proceed relatively consistently within each linguistic tradition, starting from semantic criteria but checking for the presence of plural markers in the noun forms and the absence of singular forms, thus arriving at similar, though relatively modest, lists of singularia and pluralia tantum. Singularia tantum cover several semantic categories: abstract nouns (cf. Cz. *spravedlnost*, Gr. *δικαιοσύνη* (dikaiosíni), both 'justice'), mass nouns (Cz. *popel*, Gr. *στάχτη* (stákhti), both 'ash'), collective nouns (Cz. *nábytek* 'furniture', Gr. *κλήρος* (klíros) 'clergy'). Pluralia tantum include inherently paired objects (Cz. *brýle*, Gr. *γυαλιά* (yialiá), both 'glasses'), plural mass concepts (Cz. *splašky*, Gr. *λύματα* (límata), both 'dregs'), or names of events (Cz. *narozeniny*, Gr. *γενέθλια* (yenéthlia), both 'birthday').

2.4. A view from Universal Dependencies

A comment is due on the representation of the category of number in the Universal Dependencies collection, as the treebanks from this collection, in the respective versions specified below, were used to train the Stanza and UDPipe models, which we employ to annotate the comparable corpora for the present study.

Within Universal Dependencies treebanks, which are constructed on the basis of a unified annotation scheme (de Marneffe et al., 2021), number is encoded as a morphosyntactic feature

	Raw comparable corpora		Processed datasets		
Language	**Sentences**	**Tokens**	**Tool**	**Sentences**	**Tokens**
English	5,000,000	104,430,900	Stanza	5,095,753	115,948,006
			UDPipe	5,038,278	117,091,745
Czech	5,000,000	81,762,710	Stanza	5,043,844	87,456,006
			UDPipe	5,067,301	87,546,410
Greek	5,000,000	106,908,786	Stanza	5,046,190	112,260,066
			UDPipe	6,051,461	109,330,298

Table 1: Size of the raw comparable corpora and of the datasets processed by the two annotation tools.

of individual noun forms. This `Number` feature takes the values `Sing` (singular) and `Plur` (plural) in the treebanks for English, Czech, and Greek. While Greek is limited to these two values, the English data additionally include the value `Ptan` (plurale tantum) with nouns that appear only in the plural. In the Universal Dependencies treebanks of Czech, besides singular and plural, the value `Dual` is attested for forms referring to two entities, as well as the value `Coll` (collective), used for nouns that employ grammatical singular to denote sets of objects.

Linguistically, these values are clearly heterogeneous: At the level of the grammatical opposition between singular and plural, we find the class of pluralia tantum, defined precisely by the absence of one of the two values, alongside the lexical category of collective nouns. The inclusion of these values—together with additional ones attested in Universal Dependencies treebanks of languages other than the three analyzed here—likely reflects annotation decisions inherited from the original datasets prior to their harmonization within the Universal Dependencies framework.[1]

2.5. Expectations arising from research on paradigmatic defectiveness

A final line of research relevant to our study concerns morphological defectiveness. While not focusing specifically on number, it examines inflectional paradigms more broadly, challenging the assumption of paradigmatic completeness and regularity. Baerman et al. (2010) show that large portions of the lexicon exhibit systematic restrictions, with grammatically predictable forms often unattested. Based on corpus evidence, Janda and Tyers (2021) demonstrate that many Russian nouns systematically avoid certain case–number combinations, with defectiveness varying across inflectional classes. Nikolaev and Bermel (2022) report similar patterns in Czech, arguing that paradigmatic gaps extend across entire semantic domains rather than being isolated lexical anomalies.

[1] Cf. the Universal Dependencies documentation on English, Czech, and Greek, and on the feature itself.

3. Data and Methods

3.1. Construction of the comparable corpora

For the purposes of the present study, we constructed three comparable monolingual corpora of 5 million sentences each for Czech, English, and Greek, drawn from the Leipzig Corpora Collection (Goldhahn et al., 2012). The Leipzig Corpora Collection was chosen for its consistent preprocessing methodology, comparable text types, and sentence-level organization across languages. Each corpus combines news and Wikipedia texts in a 4:1 ratio (80% news, 20% Wikipedia), with news components spanning 2019–2024 and Wikipedia snapshots from 2016–2021. Despite identical sentence counts, the corpora differ in total tokens (cf. the left-hand side of Table 1), corresponding to average sentence lengths of 20.9 tokens for English, 16.4 tokens for Czech, and 21.4 tokens for Greek.

3.2. Morphosyntactic annotation

We annotated each corpus using two tools trained on treebanks from the Universal Dependencies collection: the latest version of Stanza, Stanza 1.11.0 (Qi et al., 2020) with models trained on Universal Dependencies 2.15, and the latest version of UDPipe 2 (Straka, 2018) with models trained on Universal Dependencies 2.17 (`english-gum`, `czech-pdtc`, `greek-gud`).[2] Both tools were ap-

[2] For the UDPipe models, high accuracy is reported across all tasks relevant to our study (`https://ufal.mff.cuni.cz/udpipe/2/models`); cf. the results for sentence segmentation, word-level tokenization, part-of-speech tagging, morphological feature prediction, and lemmatization: `english-gum`: 95.77, 99.74, 98.12, 98.04, 98.83; `czech-pdtc`: 94.82, 99.96, 99.24, 98.88, 99.54; `greek-gud`: 95.24, 99.94, 98.08, 94.36, 95.93.

Lang	Tool	Noun tokens	Sing	Plur	Dual	Ptan	No value
English	Stanza	22,079,865	16,098,396 (72.9%)	5,918,052 (26.8%)	—	63,417 (0.3%)	— —
	UDPipe	21,604,055	15,651,183 (72.4%)	5,867,912 (27.2%)	—	71,894 (0.3%)	13,066 (0.1%)
Czech	Stanza	20,730,449	14,595,098 (70.4%)	5,537,569 (26.7%)	4,603 (0.02%)	—	593,179 (2.9%)
	UDPipe	20,629,395	14,358,900 (69.6%)	5,636,277 (27.3%)	5,315 (0.03%)	—	628,903 (3.0%)
Greek	Stanza	21,671,723	14,642,436 (67.6%)	6,054,245 (27.9%)	—	—	975,042 (4.5%)
	UDPipe	21,526,746	14,656,751 (68.1%)	6,120,727 (28.4%)	—	—	749,268 (3.5%)

Table 2: Distribution of the values of the `Number` feature across noun tokens. Percentages calculated relative to total noun token counts in the individual datasets (rows). A dash (—) indicates the absence of the specific value in the dataset.

plied with default settings including sentence segmentation and word-level tokenization, ensuring easy replicability of the data compilation process and direct comparability with other Universal Dependencies research. By using two tools, we can validate quality through their agreement.

All processing was conducted on a high-performance computing cluster at the Institute of Formal and Applied Linguistics of Charles University using parallel processing strategies: input sentences were divided into 5,000-sentence bundles distributed across multiple cluster partitions. Output files follow CoNLL-U format (Nivre et al., 2016). The right-hand side of Table 1 lists sentence and token counts for the six resulting datasets.

The datasets have been made available in the LINDAT/CLARIAH-CZ repository at `http://hdl.handle.net/11234/1-6120`. Additional materials, including analysis scripts, manually annotated files for the evaluation of part-of-speech tagging and lemmatization quality (see Section 3.3), as well as grammar-derived lists of singularia and pluralia tantum (used in Section 4.3), are available on GitHub.

3.3. Annotation quality validation

Automated annotation tools may alter pre-existing segmentation—even when processing pre-segmented input (Demrozi et al., 2023; Bindi, 2025). This effect was observed in our processing, prompting an evaluation of consistency between Stanza and UDPipe. We therefore assessed the preservation of sentence and token segmentation relative to the original corpora and inter-tool differences in part-of-speech distributions.

Sentences and tokens in the CoNLL-U outputs were matched against the original plain-text corpora. Preservation refers to sentences or tokens remaining unchanged relative to the raw corpus; inter-tool agreement corresponds to the proportion preserved intact by both tools.

At the **sentence** level, Czech and English show high preservation rates and strong agreement between the tools. In Czech, Stanza preserved 97.62% and UDPipe 96.80% of sentences, with 96.07% jointly preserved. In English, Stanza preserved 96.70% and UDPipe 95.94%, with 94.20% jointly preserved. Greek shows lower agreement: Stanza preserved 93.35% of sentences, whereas UDPipe preserved 79.35% (77.35% jointly).

At the **token** level, preservation rates are comparable across tools. In Czech, Stanza preserved 85.33% and UDPipe 85.13% of tokens (85.08% jointly). In English, Stanza preserved 86.97% and UDPipe 88.26% (86.55% jointly). In Greek, Stanza preserved 89.22% and UDPipe 91.09%, with 88.20% jointly preserved.

Since the raw corpora are not tagged for grammatical categories, **part-of-speech** distributions were compared only between Stanza- and UDPipe-annotated datasets, focusing on nouns as the category of interest. As shown in Table 2, Stanza identified more noun forms than UDPipe in all three languages. The difference is largest in English (nearly 476 thousand tokens), smaller in Greek (146 thousand), and smallest in Czech (101 thousand tokens).

To assess the quality of part-of-speech tagging and lemmatization, 500 tokens per tool (1,000 per language, and 3,000 instances in total) were manually inspected. For Czech, POS tagging accuracy reached 95.6% for Stanza (22 incorrect assignments) and 97.6% for UDPipe (12 errors), while lemmatization accuracy was 95.0% (25 errors) and 98.4% (8 errors), respectively. For English, POS tagging accuracy was 94.6% (27 errors) for Stanza and 94.8% (25 errors) for UDPipe, with lemmatization accuracy at 92.6% (35 errors) and

93.6% (30 errors), respectively. For Greek, POS tagging accuracy was 88.4% for Stanza (58 errors) and 84.2% for UDPipe (79 errors), while lemmatization accuracy reached 84.4% (78 errors) and 77.2% (114 errors), respectively.

3.4. Extraction of number values and calculation of the plural ratio

From each annotated corpus, we extracted all tokens tagged as nouns[3] and, for each such token, retrieved and stored the value of the `Number` feature from the CoNLL-U `FEATS` column. As shown in Table 2, singular and plural forms were consistently attested in all six datasets. For English, forms annotated with the `Number` value `Ptan` (plurale tantum) were identified by both tools. For Czech, `Dual` forms were identified in both datasets, but no forms were assigned the value `Coll` (unlike in the Universal Dependencies treebanks of Czech; cf. Section 2.4).

To analyze number distribution within and across languages, we aggregated token counts by lemma and calculated a plural ratio for each distinct lemma. Forms with the values `Ptan` and `Dual` were treated as non-singular and added to plural forms. The plural ratio, representing the proportion of non-singular forms among all number-marked tokens of a lemma, was calculated for each language as follows:

– **English:**

$$\text{plural ratio} = \frac{\text{count}_{\texttt{Plur}} + \text{count}_{\texttt{Ptan}}}{\text{count}_{\texttt{Sing}} + \text{count}_{\texttt{Plur}} + \text{count}_{\texttt{Ptan}}}$$

– **Czech:**

$$\text{plural ratio} = \frac{\text{count}_{\texttt{Plur}} + \text{count}_{\texttt{Dual}}}{\text{count}_{\texttt{Sing}} + \text{count}_{\texttt{Plur}} + \text{count}_{\texttt{Dual}}}$$

– **Greek:**

$$\text{plural ratio} = \frac{\text{count}_{\texttt{Plur}}}{\text{count}_{\texttt{Sing}} + \text{count}_{\texttt{Plur}}}$$

A plural ratio of 0 indicates the lemma occurs only in singular, while a ratio of 1 indicates the absence of singular forms among the attested tokens of the lemma. Ratios between 0 and 1 show the lemma occurs in both singular and plural in varying proportions (see Section 4.2 for details).

Apart from the values included in the plural ratio, Table 2 also reports counts of forms with no assigned value. For English, this affects only one dataset (<0.1% of forms), while for Czech it is about 3% and for Greek up to 4.5%. These forms merit further investigation, as they reflect the tools' performance and data used for their training. For Section 4, however, they are disregarded.

[3]Thus, tokens tagged `PROPN` (reserved for proper nouns in the Universal Dependencies scheme) are not included in the analysis, as proper nouns are expected to be biased toward singular-only or plural-only usage and could skew the distributional patterns.

3.5. Compilation of defective-number noun lists for validation in the data

In order to assess how the automatically annotated comparable corpora reflect phenomena described in grammatical accounts within and across languages, we used the reference grammars cited in Section 2.2 to compile lists of nouns considered as singularia tantum and pluralia tantum. For English, in cases of conflicting classifications of particular lexemes—which, as noted above, can be explained by prioritizing morphological over syntactic criteria, or vice versa, in delimiting these categories—we followed Quirk's classification.

The lists compiled in this way contain 54 singularia tantum and 87 pluralia tantum candidates for English, 27 and 54 respectively for Czech, and 22 and 59 for Greek. For each candidate, the corresponding lemma was identified in the annotated datasets, and its plural ratio was determined; see Section 4.3 for a discussion of the attestation and distribution of these items.

4. Results and Linguistic Analysis

4.1. Singular vs. plural at the token level within and across languages

All six datasets exhibit a strong singular bias (67–73% of noun tokens) when considering the proportion of singular and plural forms—without yet linking forms to lemmas. The ratios of singular and plural forms are consistent for each language across both datasets and are also similar across languages: 2.7:1 for English, 2.6:1 for Czech, and 2.4:1 for Greek. Table 2 shows the distribution of `Number` values across all noun tokens in each dataset.

Before merging tokens marked with the values `Ptan` and `Dual` with plural forms for the calculation of the plural ratio in the next section, we briefly examine these categories, as they may reveal potential inconsistencies in the automatic annotation.

The `Ptan`, which occurs only in the English datasets, is assigned by Stanza to 63 thousand forms belonging to 363 lemmas, whereas UDPipe assigns it to 71 thousand forms associated with 800 lemmas. In both datasets, most cases involve expressions containing digits (e.g. *1970s*) or canonical examples of pluralia tantum (e.g. *thanks, clothes*). However, while in the Stanza dataset the lemma *clothes* is instantiated exclusively by `Ptan` forms, the UDPipe dataset also con-

tains some Sing and Plur forms under the same lemma. A similar inconsistency appears with *remains* and *remain* (the latter of which should not exist): cf. *remains* by Stanza: 117 Plur, 1,682 Ptan, and by UDPipe: 57 Plur, 764 Ptan; *remain* by Stanza: 76 Sing, 44 Plur, 7 Ptan, and by UDPipe: 38 Sing, 762 Plur, 321 Ptan.

The Dual value, provided only for Czech, was assigned by Stanza to 4.6 thousand forms belonging to three body-part lemmas (*oko* 'eye', *ruka* 'arm', *noha* 'leg'), which is consistent with Universal Dependencies guidelines. In contrast, the UDPipe dataset contains 5.3 thousand such tokens distributed across 260 lemmas, including nouns referring to persons or objects (e.g. *holka* 'girl', *droga* 'druh'), where the forms likely represent colloquial instrumental plurals rather than genuine dual forms. Among these lemmas, we also find instrumental forms that were not recognized as inflected forms and are incorrectly considered to be lemmas (e.g. *kanálama*, which should be lemmatized as *kanál* 'canal'), as well as non-existent strings (e.g. *pracha* instead of *prachy* 'money').

4.2. Proportion of singular and plural forms per lemma

A simple way to examine noun behavior with respect to number is to rank noun lemmas by their plural ratio, i.e., the percentage of non-singular forms among all forms of a lemma, as introduced above. Figures 1–3 depict this ranking for each dataset. The x-axis shows the plural ratio, with the value 0 on the left (i.e. exclusively singular forms and no plurals) and the value 1 on the right (i.e. exclusively plural forms and no singulars). The y-axis, on a logarithmic scale, represents the number of lemmas corresponding to a given plural ratio. The absolute frequencies of individual lemmas are not taken into account; thus, both high- and low-frequency lemmas fall into the same bar if they have the same proportion of plural forms. However, only lemmas with at least ten occurrences were included in the plots.

All figures show the same trend: A high number of singular-only noun lemmas creates a peak on the left in all bar plots, while plural-only nouns form a smaller peak on the right. While this pattern deviates from grammatical expectations, it is consistent with our previous results based on different datasets annotated with different tools (namely, on the British National Corpus for English and the SYN2020 corpus for Czech; Ševčíková and Diamantopoulos 2025).

Absolute counts and percentages for the extremes, for (arbitrarily chosen) adjacent ranges (plural forms up to 10 or above 90%) and for the intermediate range are reported on the right-hand side of Table 3, which, like Figures 1–3, includes only lemmas with at least ten occurrences. Lemmas with ten or more forms but no plural among them account for roughly 30% of nouns in each dataset—far exceeding their peripheral status in grammars. Relaxing the ten-token threshold (left-hand side of Table 3) increases the proportion of singular-only nouns to 60–70%. Plural-only nouns remain lower, between 2–12% with the threshold and 12–25% without it.

The difference between the two sides of the table shows that low-frequency lemmas with fewer than 10 occurrences are primarily found in the extreme groups with a plural ratio of 0 and 1. Applying a minimum threshold of 10 forms reduces these extreme groups dramatically; at the same time, however, it leads to a relative increase—given the newly established totals—in the groups adjacent to these extremes.[4]

4.3. Singular-only and plural-only nouns: data vs. grammars

The correspondence between the analyzed datasets and the picture of paradigmatic defectiveness depicted in the grammars is illustrated by colored dots in Figures 1–3. Nouns listed in grammars as singularia tantum (red dots) and pluralia tantum (blue dots) are plotted on bar plots showing the distribution of words according to the plural ratio; the candidate dots are placed at arbitrary vertical positions with randomly varied sizes for visual distinction.

For English, candidate words in both datasets are only partially located in the expected extremes, with a substantial proportion scattered across the middle range among nouns attested in both singular and plural. Examples of words whose annotation matches grammatical expectations include singularia tantum such as *honesty* (plural ratio 0.00 in both datasets) and *physics* (Stanza 0.000, UDPipe 0.001; the same order applies throughout this section), as well as pluralia tantum like *people* (0.98, 1.00). However, the tools differ in annotating other predicted singularia tantum (*economics*: 0.35, 0.01; *politics*: 0.94, 0.16) and pluralia tantum (*police*: 0.60, 0.53; *vermin*: 0.69, 0.00).

For Czech and Greek, the data show greater agreement with the grammars: In the UDPipe-annotated dataset for Czech and the Stanza dataset for Greek, most candidates cluster in the respective extremes or close to them. In the Stanza dataset for Czech and the UDPipe dataset for Greek, singularia tantum remain on

[4] The absolute counts of lemmas in the (0, 0.1] and [0.9, 1) intervals remain unchanged, as only lemmas with at least ten occurrences fall into these intervals, even without applying a minimum frequency threshold.

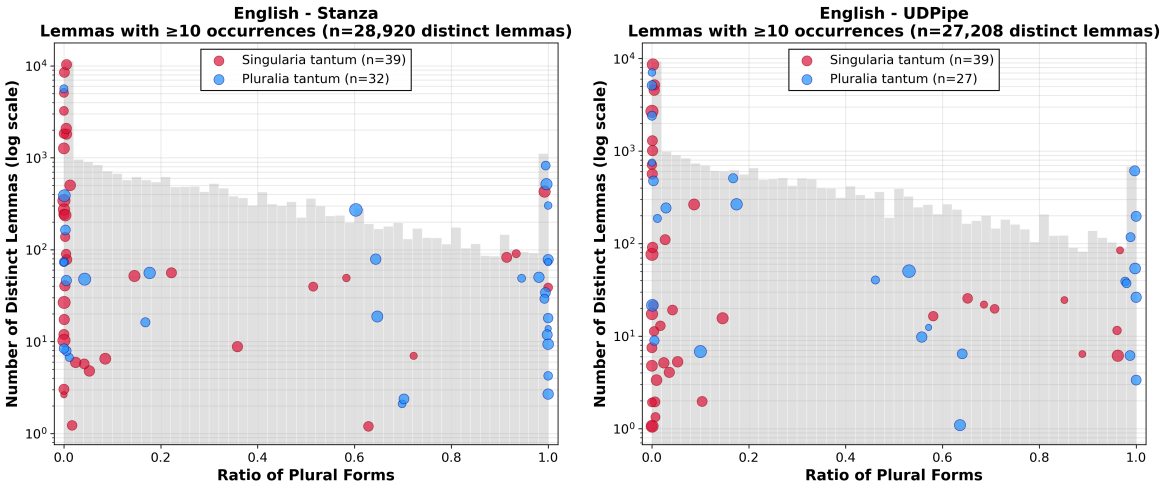


Figure 1: Distribution of English noun lemmas (with ≥10 occurrences) by plural ratio. Singularia tantum (red, 54 candidates) and pluralia tantum (blue, 87 candidates) overlaid. Left: Stanza; right: UDPipe.

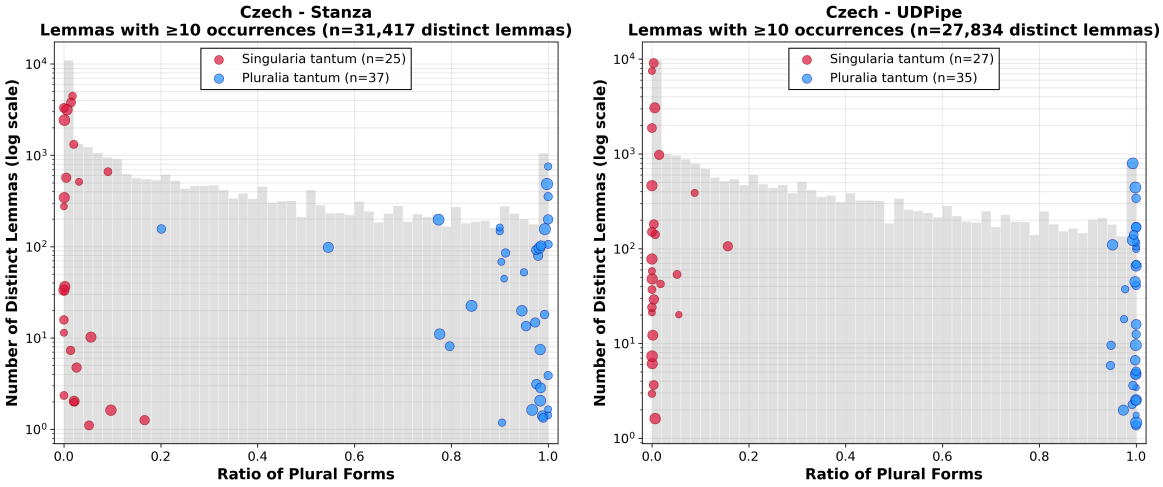


Figure 2: Distribution of Czech noun lemmas (with ≥10 occurrences) by plural ratio. Singularia tantum (red, 27 candidates) and pluralia tantum (blue, 54 candidates) overlaid. Left: Stanza; right: UDPipe.

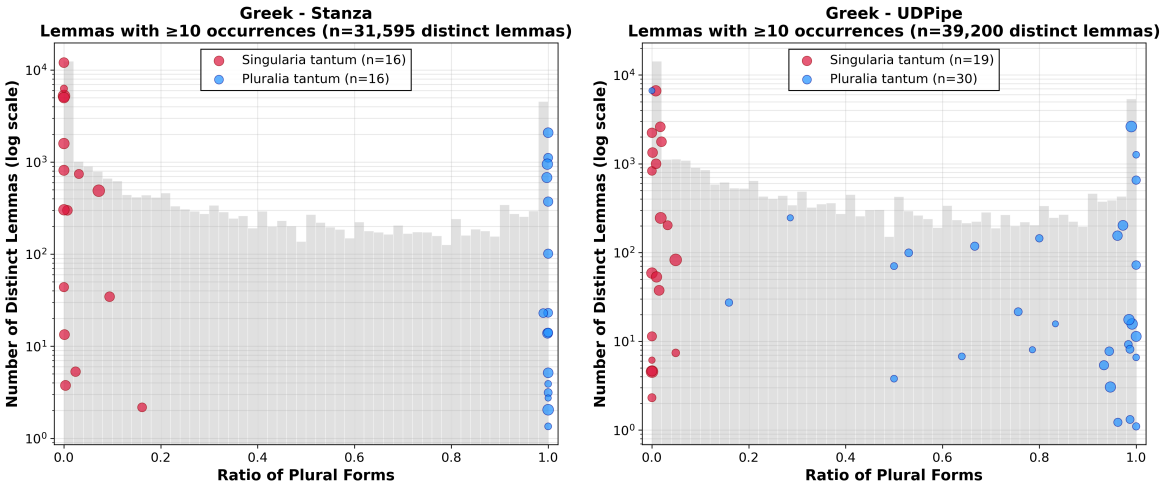


Figure 3: Distribution of Greek noun lemmas (with ≥10 occurrences) by plural ratio. Singularia tantum (red, 22 candidates) and pluralia tantum (blue, 59 candidates) overlaid. Left: Stanza; right: UDPipe.

All noun lemmas						Noun lemmas with ≥10 forms					
Total nouns	Plural ratio = 0	(0,0.1]	(0.1,0.9)	[0.9,1)	1	Total nouns	Plural ratio = 0	(0,0.1]	(0.1,0.9)	[0.9,1)	1
English Stanza											
146,892	102,554	4,897	20,827	522	18,092	28,920	9,930	4,897	12,549	522	1,022
	(69.8%)	(3.3%)	(14.2%)	(0.4%)	(12.3%)		(34.3%)	(16.9%)	(43.4%)	(1.8%)	(3.5%)
English UDPipe											
147,966	100,204	4,990	22,451	528	19,793	27,208	8,024	4,990	13,064	528	602
	(67.7%)	(3.4%)	(15.2%)	(0.4%)	(13.4%)		(29.5%)	(18.3%)	(48.0%)	(1.9%)	(2.2%)
Czech Stanza											
145,635	91,321	6,921	25,504	993	20,896	31,417	8,922	6,921	13,633	993	948
	(62.7%)	(4.8%)	(17.5%)	(0.7%)	(14.3%)		(28.4%)	(22.0%)	(43.4%)	(3.2%)	(3.0%)
Czech UDPipe											
124,745	75,155	5,597	22,303	851	20,839	27,834	7,947	5,597	12,727	851	712
	(60.2%)	(4.5%)	(17.9%)	(0.7%)	(16.7%)		(28.6%)	(20.1%)	(45.7%)	(3.1%)	(2.6%)
Greek Stanza											
194,559	118,945	5,076	21,219	1,622	47,697	31,595	10,943	5,076	9,877	1,622	4,077
	(61.1%)	(2.6%)	(10.9%)	(0.8%)	(24.5%)		(34.6%)	(16.1%)	(31.3%)	(5.1%)	(12.9%)
Greek UDPipe											
272,126	163,821	6,195	30,235	2,200	69,675	39,200	12,603	6,195	13,393	2,200	4,809
	(60.2%)	(2.3%)	(11.1%)	(0.8%)	(25.6%)		(32.2%)	(15.8%)	(34.2%)	(5.6%)	(12.3%)

Table 3: Lemma-level distribution of `Number` values for all nouns (left) vs. nouns with ≥10 forms (right). Lemma counts by plural ratio: 0 (only singular); (0,0.1] (mostly singular); (0.1,0.9) (both singular and plural); [0.9,1) (mostly plural); 1 (only plural). Percentages sum to 100% per total count of lemmas.

		English		Czech		Greek	
		Sg tantum	Pl tantum	Sg tantum	Pl tantum	Sg tantum	Pl tantum
Tool	**Lemmas**	Candidates / Attested / Confirmed					
Stanza	All	54 / 40 / 10	87 / 37 / 12	27 / 25 / 7	54 / 39 / 8	22 / 20 / 10	59 / 29 / 21
	≥10	54 / 39 / 9	87 / 32 / 8	27 / 25 / 7	54 / 37 / 7	22 / 16 / 7	59 / 16 / 12
UDPipe	All	54 / 39 / 5	87 / 33 / 9	27 / 27 / 10	54 / 38 / 20	22 / 20 / 8	59 / 45 / 14
	≥10	54 / 39 / 5	87 / 27 / 3	27 / 27 / 10	54 / 35 / 18	22 / 19 / 7	59 / 30 / 6

Table 4: Attestation of grammar-derived singularia tantum and pluralia tantum candidates in the datasets. For each language, the table shows the number of candidates, how many of them are attested (in the full dataset vs. among lemmas with ≥10 occurrences), and how many of those attested are found exclusively in singular or plural in the datasets.

the left side of the plot, whereas pluralia tantum are dispersed across the scale, reaching the plural ratios of singularia tantum. In Czech, singularia tantum like *lidstvo* 'humankind' (0.00 in both datasets) and *kvítí* 'flowers' (0.00, 0.09), and pluralia tantum such as *kleště* 'pliers' (1.00) and *dveře* 'door' (0.99, 1.00) match grammatical expectations, though *vrátka* 'gate' shows large discrepancies (0.20, 0.99). In Greek, both tools largely agree on singularia tantum like *οξυγόνο* (oxigóno) 'oxygen' (0.00 in both datasets) and *ζάχαρη* (zákhari) 'sugar' (0.00, 0.01), and pluralia tantum such as *περίχωρα* (períkhora) 'suburbs' (1.00, 0.96) and *γενέθλια* (yenéthlia) 'birthday' (0.99, 0.95). However, expected pluralia tantum like *πρόθυρα* (próthira) 'threshold' are analyzed differently (1.00, 0.50). Other predicted pluralia tantum, such as *έξοδα* (éxoda) 'expenses', are absent in the Stanza dataset, while UDPipe treats them mostly as singular forms (plural ratio 0.29).

Table 4 reports the exact counts of candidate nouns, how many were found in the data, and how many were confirmed as restricted exclusively to singular or plural forms. For all languages, more singularia tantum than pluralia tantum candidates appeared on the lists. Many pluralia tantum were not found even when searching the entire dataset, with further reductions when

limiting to lemmas with ≥10 occurrences; this effect is smaller for singularia tantum. The absence of the Greek plurale tantum *άρματα* (ármata) 'chariots' likely reflects its archaic character, while the failure to find singularia tantum like *γυμναστική* (yimnastikí) 'gymnastics' or pluralia tantum like *μαθηματικά* (mathimatiká) 'mathematics', as with English pluralia tantum *outskirts* or *shorts*, points to lemmatization issues. Problematic lemmatization is also seen in the Czech pluralia tantum *Velikonoce* 'Easter', which was not found in the UDPipe dataset, as it appears there only in lowercase (*velikonoce*), along with two incorrect lemmas (*velikonoc* and *velikonce*).

Last but not least, the data also reveal that, particularly among nouns attested in the datasets exclusively in the singular (plural ratio 0.00), there are nouns that grammatical descriptions do not consider to be singularia tantum. In the languages analyzed, these include spatial or temporal expressions; cf. English *north* and *southwest*; Czech *jih* 'south' and *sever* 'north', or *minulost* 'past', *budoucno* 'future', *dnešek* 'this day', *leden* 'January', *úterý* 'Tuesday'; and Greek *μέλλον* (méllon) 'future', μεσημέρι (mesiméri) 'noon', *νύχτα* (níkhta) 'night', *Κυριακή* (Kiriakí) 'Sunday'.

5. Conclusions

In this study, we examined the morphological category of number in English, Czech, and Greek using comparable corpora annotated with Stanza and UDPipe, tools reported to achieve excellent performance. By systematically comparing paired datasets for each language, we aimed to assess whether such annotated corpora constitute a reliable basis for drawing linguistic conclusions.

Despite differences between the paired datasets in sentence and word segmentation, noun counts, and number values, these discrepancies did not appear to substantially affect the overall patterns. In all datasets for each language, singular forms were two to three times more frequent than plurals, and lemma-level analysis revealed a picture differing from grammar descriptions, showing a strong inclination of a substantial portion of nouns toward singular and, to a lesser extent, toward plural forms.

A more detailed inspection of a small set of nouns expected to exhibit defective number revealed inconsistencies, indicating that such annotated comparable corpora may yield divergent results on linguistically nuanced questions. The observed differences seem to stem from variations in lemmatization—Stanza producing more stable and linguistically conformant lemmas, while UDPipe sometimes generates nonexistent lemmas—and likely point to additional issues that should be addressed in dedicated follow-up experiments.

6. Acknowledgements

The research reported in the present paper has been supported by the Czech Science Foundation (Project No. GA26-21822S). It has been using data and tools provided by the LINDAT/CLARIAH-CZ Research Infrastructure (Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2023062).

7. Bibliographical References

Paolo Acquaviva and Laure Gardelle. 2023. Pluralia tantum and singularia tantum. *The Wiley Blackwell Companion to Morphology*, pages 1–28.

Matthew Baerman, Greville G Corbett, and Dunstan Brown. 2010. *Defective paradigms: Missing forms and what they tell us*. Liverpool University Press.

Laurie Bauer, Rochelle Lieber, and Ingo Plag. 2013. *The Oxford reference guide to English morphology*. Oxford University Press, Oxford.

Beatrice Bindi. 2025. Evaluating Stanza and UDPipe for Morphosyntactic Annotation of Old Russian: A Case Study on Maximus the Greek. *Scripta & e-Scripta*, pages 39–60.

Hee-Soo Choi, Bruno Guillaume, and Karën Fort. 2021. Corpus-based language universals analysis using Universal Dependencies. In *Proceedings of the Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021)*, pages 33–44, Sofia, Bulgaria. Association for Computational Linguistics.

Greville G. Corbett. 2000. *Number*. Cambridge University Press.

Greville G Corbett. 2019. Pluralia tantum nouns and the theory of features: A typology of nouns with non-canonical number properties. *Morphology*, 29(1):51–108.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Florenc Demrozi, Cristian Turetta, Fadi Al Machot, Graziano Pravadelli, and Philipp H. Kindt. 2023. A comprehensive review of automated data annotation techniques in human activity recognition.

Matthew S. Dryer and Martin Haspelmath. 2013. The World Atlas of Language Structures Online.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Bohuslav Havránek and Alois Jedlička. 2002. *Stručná mluvnice česká*. Fortuna, Praha.

David Holton, Peter Mackridge, Irene Philippaki-Warburton, and Vassilios Spyropoulos. 2012. *Greek: A comprehensive grammar of the modern language*. Routledge.

Rodney Huddleston and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge.

A. Laura Janda and M. Francis Tyers. 2021. Less is more: why all paradigms are defective, and why that is a good thing. *Corpus Linguistics and Linguistic Theory*, 17(1):109–141.

Yingqi Jing, Paul Widmer, and Balthasar Bickel. 2023. Word order evolves at similar rates in main and subordinate clauses. *Diachronica*, 40(4):532–556.

Miroslav Komárek, Jan Kořenský, Jan Petr, and Jarmila Veselková. 1986. *Mluvnice češtiny 2. Tvarosloví*. Academia, Praha.

Natalia Levshina, Savithry Namboodiripad, and Marc Allassonnière-Tang et al. 2023. Why we need a gradient approach to word order. *Linguistics*, 61(4):825–883.

Alexandre Nikolaev and Neil Bermel. 2022. Explaining uncertainty and defectivity of inflectional paradigms. *Cognitive Linguistics*, 33(3):585–621.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal Dependencies v1: A multilingual treebank collection. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

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.

Randolph Quirk, Sidney Greenbaum, Geoffrey N. Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.

Magda Ševčíková and Konstantinos Diamantopoulos. 2025. Word-formation of singular-only nouns: A pilot study in four languages. Presented at the Word-Formation Theories VII & Typology and Universals in Word-Formation VI conference, Košice, Slovakia.

Hedvig Skirgård, Hannah J. Haynie, and Damián E. Blasi et al. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9(16):eadg6175.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Χρήστος Κλαίρης and Γεώργιος Μπαμπινιώτης. 2010. *Γραμματική της νέας ελληνικής: δομολειτουργική-επικοινωνιακή. Το όνομα της νέας Ελληνικής.* Ελληνικά Γράμματα.

Μαρία Τζεβελέκου, Βασιλική Κάντζου, and Σπυριδούλα Σταμούλη. 2007. *Βασική γραμματική της Ελληνικής*. Ινστιτούτο Επεξεργασίας του Λόγου - Ε.Κ. "Αθηνά".

Μανόλης Α. Τριανταφυλλίδης. 1979. *Νεοελληνική γραμματική: αναπροσαρμογή της Μικρής Νεοελληνικής Γραμματικής του Μανόλη Τριανταφυλλίδη*. Οργανισμός Εκδόσεων Διδακτικών Βιβλίων.

Σοφρώνης Χατζησαββίδης and Αθανασία Χατζησαββίδου. 2014. *Γραμματική της Νέας Ελληνικής Γλώσσας Α', Β', Γ' Γυμνασίου*. Οργανισμός Εκδοσης Διδακτικών Βιβλίων - Υπουργείο Παιδείας και Θρησκευμάτων.

8. Language Resource References

Leipzig Corpora Collection. 2019. *English news corpus based on material from 2019*. University of Leipzig. Accessed: 2026-03-04.

Leipzig Corpora Collection. 2020a. *Czech news corpus based on material from 2020*. University of Leipzig. Accessed: 2026-03-04.

Leipzig Corpora Collection. 2020b. *English news corpus based on material from 2020*. University of Leipzig. Accessed: 2026-03-04.

Leipzig Corpora Collection. 2021a. *Czech Wikipedia corpus based on material from 2021*. University of Leipzig. Accessed: 2026-03-04.

Leipzig Corpora Collection. 2021b. *English Wikipedia corpus based on material from 2021*. University of Leipzig. Accessed: 2026-03-04.

Leipzig Corpora Collection. 2021c. *Modern Greek news corpus based on material from 2021*. University of Leipzig. Accessed: 2026-03-04.

Leipzig Corpora Collection. 2021d. *Modern Greek Wikipedia corpus based on material from 2021*. University of Leipzig. Accessed: 2026-03-04.

Leipzig Corpora Collection. 2022a. *Czech news corpus based on material from 2022*. University of Leipzig. Accessed: 2026-03-04.

Leipzig Corpora Collection. 2022b. *Modern Greek news corpus based on material from 2022*. University of Leipzig. Accessed: 2026-03-04.

Leipzig Corpora Collection. 2023a. *Czech news corpus based on material from 2023*. University of Leipzig. Accessed: 2026-03-04.

Leipzig Corpora Collection. 2023b. *English news corpus based on material from 2023*. University of Leipzig. Accessed: 2026-03-04.

Leipzig Corpora Collection. 2023c. *Modern Greek news corpus based on material from 2023*. University of Leipzig. Accessed: 2026-03-04.

Leipzig Corpora Collection. 2024a. *Czech news corpus based on material from 2024*. University of Leipzig. Accessed: 2026-03-04.

Leipzig Corpora Collection. 2024b. *English news corpus based on material from 2024*. University of Leipzig. Accessed: 2026-03-04.

Leipzig Corpora Collection. 2024c. *Modern Greek news corpus based on material from 2024*. University of Leipzig. Accessed: 2026-03-04.

Diamantopoulos, Konstantinos. 2026. *Automatically Annotated Corpora with Stanza and UDPipe for Czech, English, and Greek*. LINDAT/CLARIAH-CZ digital library.

Zeman, Daniel and Nivre, Joakim and Abrams, Mitchell et al. 2024. *Universal Dependencies 2.15*. LINDAT/CLARIAH-CZ. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Zeman, Daniel and Nivre, Joakim and Abrams, Mitchell et al. 2025. *Universal Dependencies 2.17*. LINDAT/CLARIAH-CZ. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Liebe Kolleg:innen, querid@s compañer@s: presenting the GILDEES corpus

Marie-Pauline Krielke
Saarland University
Department of Language Science and Technology
Campus A2.2, 66123 Saarbrücken, Germany
mariepauline.krieke@uni-saarland.de

Abstract

We present a multi-register (WEB, NEWS, and GOVernment texts), diachronic (2015-2024), comparable corpus annotated for lexical gender-inclusive language (GIL) features in German and Spanish. Apart from rule-based annotations, we train a transformer-based classifier to resolve semantically ambiguous neutral expressions like epicenes to reliably annotate true human referents. In a sample study, we analyze register variation in the three registers in terms of GIL features both contrastively and diachronically. We show that GIL usage increases and varies diachronically in terms of register in both languages. German texts show a higher overall frequency and diversity of GIL features than Spanish texts. However, across languages, registers behave similarly, with government text showing the strongest usage of GIL followed by news and web texts, and web texts showing the strongest innovation in terms of features. The results of our study are valuable to linguistic areas such as human and machine translation, SLA, and contribute to register-conform gender inclusive NLP downstream tasks such as machine translation, summarization, or text generation. From a diachronic point of view, our corpus and analyses are a valuable contribution to observing language change in the making.

Keywords: Gender inclusive language, German-Spanish, contrastive corpus linguistics, diachronic change

1. Introduction

In recent years, gender-inclusive language (GIL) has become increasingly common in everyday usage, particularly in grammatically gendered languages such as German and Spanish. The equal representation of female and male referents has been debated in feminist linguistics since the 1970s in both German (e.g. Trömel-Plötz, 1978; Guentherodt et al., 1980; Pusch, 1979) and Spanish scholarship (Suardiaz, 1973; Hampares, 1976), and psycholinguistic studies on German masculine generics (e.g. Braun et al., 1998; Stahlberg et al., 2001; Gygax et al., 2008) show that they predominantly evoke male representations. However, the practical implementation of these findings in actual language use has been gradual and has achieved broader institutional uptake only in the past two decades, for instance through official guidelines (Günthner, 2019).

Since the early 2000s, the gender binary has increasingly been questioned, leading to proposals for including non-binary identities (Günthner, 2019), particularly through new orthographic forms (Diewald and Steinhauer, 2022) such as *Amigxs* or *Freund*innen*. As this development remains ongoing and socially contested, there is neither broad agreement on the use of GIL nor consensus on its lexical, grammatical, or orthographic realization.

It is therefore interesting to trace recent developments of GIL features, paying attention to the communicative contexts they are used in. While for German, a monolingual diachronic database of GIL in language use exists (Dick et al., 2024), it does not cover all lexical features of GIL, lacking, for instance, epicenes and collective terms, two strategies to refer to people without explicit gender marking. For Spanish, there is no comparable resource.

We present the GILDEES corpus, the first multi-register diachronic Spanish – German comparable corpus annotated for a comprehensive set of GIL features, i.e., explicit forms of gender inclusiveness such as spelling variants (*amig@*) and double mention (*amiga o amigo*), and implicit forms like epicenes (*persona*), collective terms (*equipo*), and nominalizations (*Studierende*).

A central part of this paper is dedicated to a detailed description of corpus building (Section 3) and annotation. Special focus is put on the annotation of implicit forms of GIL (Section 3.4), often ambiguous in their reference to personal or non-personal referents (e.g., *Haushaltshilfe*, domestic help, or the activity of helping in the household). To handle this kind of ambiguity, we manually annotate a gold standard set of epicenes and train transformer-based binary classifiers to disambiguate between personal and non-personal reference.

We present sample analyses (Section 5) to illustrate applications of the resource. Diachronically, Spanish lags behind German in both the frequency and variability of GIL features. Register patterns are similar across languages: government texts (GOV) show the highest GIL usage, web texts (WEB)

Proceedings of the 19th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC 2026, pages 41–52
11 May 2026.

the lowest overall but the highest proportion of non-binary features, and news texts (NEWS) occupy an intermediate position with little feature variation. The GILDEES annotations thus enable tracking overall gender inclusiveness across registers as well as processes of diversification (increasing variability) and conventionalization (decreasing variability) over time, reflecting ongoing social negotiation and consensus formation around inclusive language.

2. Previous work

As awareness has increased that language can reflect and reinforce social inequalities and shape perceptions of gender and identity (e.g., Kaufmann and Bohner, 2014, for Spanish) and (Braun et al., 1998; Stahlberg et al., 2001; Gygax et al., 2008, 2019, for German), interest in research on the active usage of gender-inclusive language has grown substantially.

Practical implementations of GIL have been pursued through actively shaping language use via guidelines for a gender-inclusive usage in public discourse (e.g., Aguilar Gavira et al., 2019, for Spanish) and (e.g., Diewald and Steinhauer, 2022, for German). Especially in the area of gender-fair (machine) translation, research has been growing in the past few years (Daems, 2023; Piergentili et al., 2023; Lardelli et al., 2024; Savoldi et al., 2025).

Extant corpus-based research on GIL in German has shown that its usage underlies both temporal and register-dependent variation. In newspaper texts of the early 2000s, the generic masculine was still more prevalent than forms overtly including female referents (Bühlmann, 2002), while an integration of forms making both genders visible could be observed in job announcements (Demey, 2002). Well into the 2010s, an analysis of Swiss authorities texts shows a decrease in generic masculine compared to alternative forms such as double mentions (Elmiger et al., 2017). At the beginning of the 2020s, Sökefeld (2021) using a diachronic corpus (2000-2019) of newspaper and blog articles, reports the integration of overt non-binary GIL forms (*Freund*in*) from the early 2010s onwards, as well as a persisting preference for generic masculine forms in newspaper texts over alternative forms, while blog texts show a preference for neutral forms. Exemplifying the impact of political orientation in news texts, Rauth (2025) shows a temporal increase of GIL features in 2023 compared to 2021 in the German leftist newspaper *die Tagezeitung*. Political orientation also impacts GIL usage in the spoken public domain: Stecker et al. (2021) identify a general GIL increase in plenary protocols of the German Bundestag spanning 1949 - 2021 from the 1980s onward. Representatives of left-wing and green parties make stronger use of GIL than those of conservative parties. Regarding cross-register variation Dick et al. (2024) compile a multi-genre, diachronic corpus spanning 1993-2023 and find clear register-dependent differences (Twitter texts > NEWS texts > EU parliamentary text > academic texts).

Research on GIL in the Spanish-speaking community began later than in the German-speaking community (Zapf, 2024). Similarly, there are fewer corpus-based studies tracing the temporal development of GIL usage. Qualitative studies (e.g Papadopoulos, 2022; Linares, 2022) reiterate the strong influence of the Spanish Royal Academy (RAE) hindering its adoption, especially regarding the recommendation to avoid overt non-binary GIL forms. Pino (2022), however, reports on a gradual increase in mentions of non-binary GIL features in press texts. In a pilot study using the Spanish reference corpus (CREA), Medina Guerra (2016) found an increase in neutral forms compared to generic masculine forms. Comprehensive resources and/or accounts on the diachronic development of GIL usage in Spanish, taking register variation into account, are still lacking.

To our knowledge, no comparable corpora currently exist for the German–Spanish language pair. For multilingual applications, the only related resource is a parallel dataset specifically designed to improve translations involving gender-neutral language (mGeNTE; Savoldi et al., 2025). However, even where corpus resources are available, the extraction and detection of GIL present substantial challenges. A major bottleneck concerns semantically ambiguous neutral forms, which may refer to both human and non-human entities, as well as grammatically ambiguous constructions such as nominalizations (Dick et al., 2024).

On the sociolinguistic level, the use of GIL is closely tied to speaker attitudes and ideological positioning (Greene and Rubin, 1991; Matheson and Kristiansen, 1987; Cremades and Fernández-Portero, 2022), and varies across demographic factors such as age (Parks and Roberton, 2008). Since such attitudinal and social factors are reflected in register-specific usage patterns, the creation of a multi-register resource is crucial for capturing the full variability of gender-inclusive language in practice.

3. Corpus Building

3.1. Corpus compilation

The corpus consists of a balanced number of texts covering the time span 2015-2024, which were crawled from the German and Spanish web, and covers three registers: blogs and forums (WEB),

Language	Register	Texts	Tokens
German	GOV	4,803	3,683,367
	NEWS	200,000	6,387,219
	WEB	5,270	5,990,239
Total		210,073	16,060,825
Spanish	GOV	5,150	5,436,287
	NEWS	200,000	8,591,051
	WEB	5,122	4,954,390
Total		210,272	18,981,728
Total		**420,345**	**35,042,553**

Table 1: Total corpus size by language and register.

governmental press releases (GOV), and news texts (NEWS). The WEB texts consist of blog articles from 10 different domains (psychology, cooking, gastronomy, travel, sports, tech, video games, lifestyle, finance, literature, education). For temporal attribution, they were selected according to their time stamps.

The NEWS texts consist of sentences taken from the downloadable Leipzig Corpora Collection (Goldhahn et al., 2012) and are derived from news feed texts. To control for regional differences, we restricted the dataset to sources from Spain and Germany, using DNS resolution of domain names to IP addresses. Geographic attribution of domains was obtained via the requests library (Reitz, 2016), which queried the public ipinfo.io API (IPinfo) to retrieve country-level metadata associated with resolved IP addresses. To make the sub-corpora's sizes roughly comparable, we took a random sample of 20,000 texts per language and year.

The government texts were semi-hand-crawled using BootCaT (Baroni et al., 2006). They were derived from nine Spanish and German governmental institutions, including the Ministries of Public Health, Ministries of Education, Offices for Migration, Ministries of Labor, the Police, Ministries for Family and Women, Ministries of Economy, Ministries of Finance, and the President's Offices.

The exact numbers of text and tokens per register are presented in Table 1.

3.2. Metadata

Each text in the corpus was annotated with metadata such as text ID, year as derived from timestamps, register (WEB, NEWS, GOV), URL, and author (i.e., newspaper, ministry, or blog title). WEB texts are additionally annotated with the topic of the blog, e.g., psychology, cooking, books, etc. Figure 1 shows the number of authors per register and per year, indicating that NEWS texts show the highest variability in authors with a decreasing trend. The WEB texts are relatively stable in the number of different authors per year and per language. GOV texts are completely stable with nine authors per language and year since their authors correspond to the different ministries.

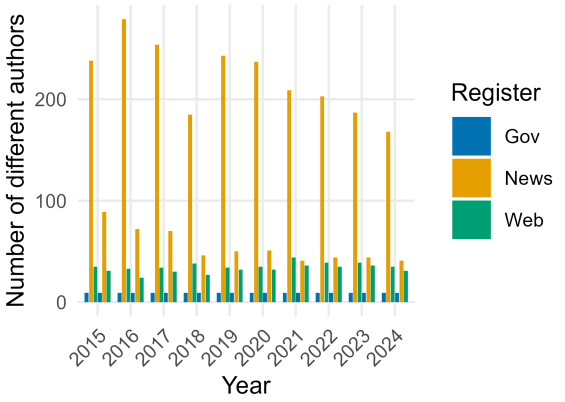


Figure 1: Authorial variation per register and year.

3.3. Morpho-syntactic annotation

Morpho-syntactic information was annotated using Stanza (Qi et al., 2020). The Spanish texts were parsed using the Stanza AnCora models to obtain language-specific XPOS tags. For German, we used Stanza with off-the-shelf settings. Morphological information is especially useful to detect gender-specific features. For instance, the double mention (amigos y amigas) can be detected via the morphological information.

3.4. Annotation of GIL features

The main focus of this work lies in the annotation of GIL features, which have been annotated using two different approaches: automatically with manual correction and manually with automatic generalization using machine learning approaches. Specifically, we divide GIL forms into explicit and implicit ones. Explicit GIL features are forms that explicitly mark the inclusion of other genders than merely the male. Implicit forms refer to those forms that do not encode natural gender and abstract away from any binary gender reference (Zapf, 2024, p.162). Explicit GIL is relatively easy to detect automatically, while implicit forms require manual approaches. These approaches will be described in the following. For an overview of GIL feature types, sub-types, and examples in Spanish and German, see Table 2.

3.4.1. Explicit GIL

Explicit GIL features are divided into two subcategories (Sökefeld, 2021): visibility and diversification. Visibility features refer to those visibly referring to both the male and the female gender (Zapf, 2024),

Type	Subtype	German	Spanish	Annotation
Explicit strategies				
Visibility	double mention	Bürger und Bürgerinnen	Los y las alumnos y alumnas	ex|vis|double
	infix-I	BürgerInnen	–	ex|vis|I
	grapho-stylistic disturbances	Bürger/innen, Bürger-innen, Bürger(innen)	alumnos/as, alumnos,-as, alumno(a)s	ex|vis|orth
	at-sign	–	alumn@s	ex|vis|at
Diversification	asterisk	Bürger*innen	alumn*s	ex|div|star
	underscore	Bürger_innen	alumn_s	ex|div|uscore
	colon	Bürger:innen	–	ex|div|colon
	-x	Bürgx	alumnxs	ex|div|x
	-e	–	alumnes	ex|div|e
Implicit strategies				
Neutralization	nominalization	Studierende, Beschäftigte, Arbeitslose	–	imp|neut|nom| (adj/part/num)
	derivation	-schaft, -hilfe, -kraft	–	imp|neut|der
	comunes	–	estudiantes, alumnado	imp|neut|com
Neutral/abstract forms	epicene	Person	persona	imp|epi
	collective terms	Team	equipo	imp|col

Table 2: Gender-inclusive linguistic strategies in German and Spanish (explicit vs. implicit).

most explicitly by mentioning both forms (double mention, e.g. *amigos y amigas*, coordinated by conjunctions like *o*/ *oder*, *y*/*und*, *bzw.* etc.), or using an infix-I (e.g., *FreundInnen*) or several forms of "grapho-stylistic disturbances" (cf. Gautherot, 2017, p.43) word-internal punctuation, e.g., the slash or the @-sign (*Schüler/innen*, *alumno/as*, *alumn@s*). Their explicit encoding facilitates automatic, rule-based annotation. We include nominal forms, but also pronominal forms (*Keine/r*, *niguno/a*) and determiners (*der/die*, *ein/-e*, *l@s*, *el/la*, etc.).

3.4.2. Implicit GIL

Implicit GIL forms are further divided into the subcategories neutralization and neutral/abstract forms (Bühlmann, 2002). **Neutralizations** refer to gender-neutral lexemes actively created (Gautherot, 2017; Sökefeld, 2021), such as derivations containing -kraft, -schaft, -hilfe, e.g. *Bürger-schaft* (citizens), *Haushalts-hilfe* (domestic help), *Führungs-kraft* (manager), and nominalizations derived from participles, e.g. Studierende (students), Geflüchtete (refugees), adjectives, e.g. Alte (elderly), or numerals, e.g. Hunderte (hundreds). These forms are highly productive in German; however, only inclusive in the plural form. In Spanish, this type of word formation is not inclusive since nominalizations are always gender-marked. Comparable forms in Spanish are the so-called comunes, i.e., lexemes lacking an explicit gender ending like *-a* or *-o*, e.g., *estudiantes* when intentionally used without a gender marked article (e.g. *Estudiantes de matemática se deben presentar a clase*) and were annotated following a manually curated list. Detection of nominalizations in German can be facilitated using UD-morphological annotation (de Marneffe et al., 2021). Plural nouns annotated without gender specification are possible candidates for nominalizations. These were extracted and manually cleaned for noise. Since not all lemmas encountered in this way were uniformly annotated without gender, in a second step, the cleaned set of lemmas was annotated with the word formation base form (adjective/participle/numerals) and used as a lookup list to automatically annotate all matching lemmas in the corpus.

Neutral/abstract forms include epicenes and collective terms (Bühlmann, 2002). Epicenes are inherently neutral expressions referring to persons, e.g., Person. However, many of them are semantically ambiguous, e.g. *Besuch* (visit or visitor). Collective terms in their most narrow definition refer to groups of people, such as *Team* or *equipo*. However, in the literature, abstract nouns referring to professions or positions (e.g., *consejo municipal*) or ministries (*Innenministerium*) are often also regarded as collective terms, referring metonymically to the group of people working within these institutions (Zapf, 2024). They represent the most problematic group, since most of them, apart from refer-

ring to groups of persons, can refer to institutions in their legal form, or buildings. Also, derivations with *-kraft* and *-hilfe* are semantically ambiguous, i.e. do not always refer to a person (e.g. *-hilfe* can refer to a person as in *Haushaltshilfe* but also to a technical device *Gehhilfe* (walking aid). The annotation guidelines are available on GitHub[1].

3.5. Personal reference resolution

To facilitate reliable identification and annotation of "true" derivations, epicenes, and collective nouns, linguistic experts compiled manually curated lemma lists. The German list contains both standalone lemmas and lemmas that may appear as compound constituents (e.g., -besuch), with compound status explicitly marked. Each lemma was further annotated for semantic ambiguity. The curated lists were subsequently used to automatically retrieve lemma occurrences from the corpus (including German compounds), and all unambiguous instances were automatically labeled accordingly (see Table 2). Ambiguous cases were annotated with the label "check" and held back for manual disambiguation in their context (see Section 3.5.1).

In a pilot trial, we used a machine learning approach to disambiguate automatically between epicenes referring to a person or not. For this, sentences with unique epicene lemmas previously identified as ambiguous, were extracted from the corpus. For each ambiguous lemma, we took a sample of at least one and up to 10 random sentences containing it, resulting in 2140 sentences with 716 unique lemmas for German. For Spanish, we allowed a sample size of 50 for each unique lemma since there are no compounds in Spanish. The sample resulted in 1145 sentences with 31 unique lemmas. Due to the great diversity of lemmas in German, the list of unique lemmas was divided into groups of ambiguity level (mostly person, 50/50 person/non-person, mostly non-person). To create the gold-standard training data, we down-sampled the 2140 sentences to 50% for manual correction, preserving the representativeness of the ambiguity groups. The Spanish set was left unchanged. Both sets were manually disambiguated and annotated with a binary label, e.g., imp|epi|person=yes or imp|epi|person=no, resulting in a total number of 1117 gold-annotated sentences with a distribution of 578 person=yes and 539 person=no for German and 1145 sentences for Spanish with 474 person=yes and 800 person=no showing a strong bias towards non-personal references.

3.5.1. Binary classifier

We implement a span-level transformer-based classifier that encodes full sentential context while restricting the classification decision to the contextualized representation of a manually annotated target expression, i.e., ambiguous epicenes. The model was implemented in Python 3 using PyTorch (Paszke et al., 2019). Pretrained transformer models and tokenizers were accessed via the Hugging Face Transformers library (Wolf et al., 2020), specifically the base version of XLM-RoBERTa (Conneau et al., 2020), since it is a multilingual model and can thus be used for both German and Spanish. The model was trained for six epochs with batch size 16, using cross-entropy loss and the AdamW optimizer (learning rate 2e^{-5}). Evaluation metrics (macro-averaged F1 and confusion matrix) were computed with scikit-learn (Pedregosa et al., 2011). Experiments were conducted in Google Colab with GPU acceleration via CUDA. The training data and code are available on GitHub[2].

Each instance consisted of the full sentence with explicit $< TARGET >$ epicene$< /TARGET >$ markers, enabling the model to attend explicitly to the referentially ambiguous token. Inputs were truncated or padded to a maximum sequence length of 192 tokens. Evaluation was conducted with predefined train (80%), development (10%), and test (10%) splits. For German, due to the high lemma diversity (many different compounds with huge overlap in head nouns), we constructed a strict lemma-held-out test set. On the held-out test set, the model achieved a macro-averaged F1 score of ≈ 0.90. The confusion matrix ($TN = 51$, $FP = 10$, $FN = 1$, $TP = 47$) indicates a slight asymmetry in errors, with more not_person instances misclassified as person (10 cases) than the reverse (1 case). At the same time, the model demonstrates particularly high recall for person-referential instances. Manual inspection of the false positives showed that, especially in cases where humans would also have struggled, led to wrong model predictions, e.g., compounds containing the lemma *Figur*, e.g., Spielfigur (play figure), which are ambiguous even in context. The Spanish model was trained with the same hyperparameters but without a lemma-held-out restriction, as the number of unique lemmas in the gold dataset was comparatively small. In Spanish, nouns are frequently modified by prepositional phrases rather than forming compounds (e.g., *personalidades del ámbito cultural*, i.e., personalities in the cultural sphere), which provide informative contextual cues for classification. The Spanish model performed slightly better than the German model, achieving a macro-F1 of ≈ 0.92 and demonstrating robust performance with minimal false negatives

[1] https://github.com/MariPeKa/GILDEES

[2] https://github.com/MariPeKa/GILDEES

for person-referential instances.

To assess practical usability, we simulated a selective prediction strategy in which thresholds of prediction probability were used to estimate the recall in automatic annotation. With restriction to $p \geq 0.8$ for $person = true$ and $p \leq 0.2$ for $person = false$, for Spanish, this approach achieved 93.7% automatic coverage, with a manual review rate of 6.3%. Within the automatically labeled subset, performance was highly reliable (precision ≈ 0.91), suggesting that model uncertainty effectively identifies borderline cases. Applying the same confidence-based filtering strategy to the German model resulted in 91.7% automatic coverage, with 8.3% of instances deferred for manual review. Among auto-decided cases, precision for *person* predictions was 0.849. Notably, recall for person-referential instances among automatically classified cases reached 1.00, indicating that no high-confidence person instances were missed.

Final automatic annotation in the corpus was therefore restricted to high-confidence predictions ($p \geq 0.8$ or $p \leq 0.2$), while ambiguous cases were labeled `check` and scheduled for manual review. Given the effectiveness of this approach for epicene annotation, we plan to extend it to collective terms in future work.

4. Access and Usage

The corpus is currently available in a derived format[3] to circumvent copyright restrictions on the complete web and government texts. In the latter case, we removed all sentences without GIL features and retained only those containing GIL features. Of the remaining sentences, we masked the word forms and lemmas of all but the GIL forms and their 4 preceding and 5 following words (see example 7). The full corpus may only be made available to specific individuals for the purpose of reviewing my research work.

5. Sample analysis

To examine the temporal development of GIL features in both languages, we conduct quantitative analyses based on all previously annotated, unambiguous features according to the following hypotheses. We first analyze overall feature frequencies. We then compare feature variability across languages and compute yearly feature diversity using entropy.

Hypotheses

1. Contrastive dimension (H1): German texts exhibit earlier and stronger adoption of GIL features.

2. Register dimension (H2): GIL usage frequencies are highest in GOV, followed by NEWS and WEB.

3. Diachronic dimension (H3): GIL usage increases across languages and registers.

4. Variability (H4): (a) German shows greater feature variability than Spanish. (b) WEB texts show the highest feature variability among all registers.

5.1. GIL over time per register in German and Spanish

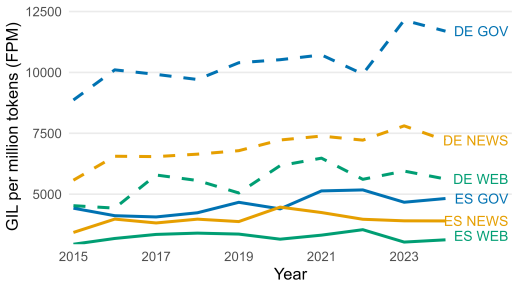


Figure 2: GIL frequencies in ES and DE per year and register.

Normalized frequencies of GIL features per language year and register, reveal both language- and register-specific differences (Figure 2).

For German, we observe consistently higher frequencies than in Spanish (H1). GOV and NEWS texts show a clear upward trend, both peaking in 2023. As expected, WEB texts show the lowest usage of GIL features overall, with stronger oscillations and an earlier peak in 2021.

In the Spanish corpus, the order of frequencies is the same, with GOV texts showing the strongest GIL usage, and an increasing trend. They are followed by NEWS and WEB texts.

The similar proportions per register are plausible: Government press releases rely most heavily on personal references when addressing topics of public interest. NEWS texts show an intermediate picture reporting about recent events and mostly specific persons of public interest whose gender identity is mostly known, and abstraction/inclusion is not needed. WEB texts show the lowest amount of (GIL) as blogs and forums often represent reports written from a first-person perspective, their experiences and opinions, rather than referring to others. The results are in line with H2.

[3] https://zenodo.org/records/19236744

Diachronically, we only see a consistent increase in GIL in German, while in Spanish, GIL usage only increases in GOV texts, partially confirming H3.

5.2. GIL feature diversity

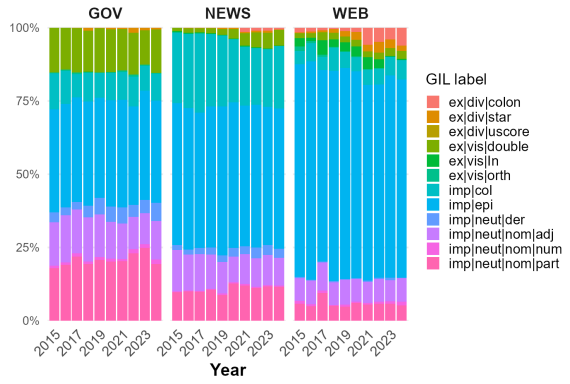


Figure 3: GIL types in German across years and registers.

German The distributions of GIL features differ across registers (Figure 3): Overall, implicit GIL features show the highest proportional representation in all registers, with epicenes being the most represented feature. This is not surprising since epicenes (e.g. *persona*) are highly frequent lexemes of personal reference in any register. Their dominance is especially high in the WEB register. A look into the WEB subcorpus reveals that especially the lemma *Kind* is highly frequent, describing a frequent topic of interest in the blogs. In NEWS texts, collective terms take the second-highest proportion, plausibly, since NEWS texts frequently report on governmental activities and their impact on the citizens. Frequent lemmas are *Regierung* (government) and *Bevölkerung* (population). Neutralizations, especially nominalizations derived from participles and adjectives, are the second most prevailing group in GOV and WEB texts. Explicit features take a lower proportion in all registers.

GOV texts show a high proportion of double mentions (e.g., *Bürgerinnen und Bürger*), a long-established and comparatively uncontested GIL strategy in German that explicitly addresses women and men alike. Their distribution remains stable over time, suggesting institutionalized use in public discourse. The only other explicit feature is the gender star, attested from 2018 onward (*Migrant*nnen*). GOV exhibits the most even and diverse distribution of GIL features across registers, with minimal temporal variation, indicating register-specific consolidation.

NEWS texts exhibit a relatively stable distribution dominated by implicit features. Among explicit forms, only double mentions increase in frequency, along with a small share of colon forms (*Bürger:innen*), which peak in 2021.

WEB texts show the highest and constantly increasing proportion of features of diversification, most prominently the colon (*Musiker:innen*) and the gender star (*Trainer*innen*). Of all registers, WEB is the only register with a notable proportion of using the infix-I (*AutorInnen*). The corpus data shows that the infix-I is also often used in hybrid form, e.g., in combination with a slash (*Follower/Innen*). Compared to the other registers, the proportion of double mentions is comparatively low. The diversity of the explicit GIL features increases over time and is biggest compared to the other registers. This is in line with our assumptions (H4b), as authors of WEB texts have the biggest freedom to linguistic innovation and diversity due to lower editorial regulation.

From a probabilistic point of view, the three registers differ in terms of overall entropy (Figure 4), which is highest for GOV texts, in line with the most even distribution of features in this register. The high entropy also indicates that German institutional texts make a balanced use of available GIL features. The fairly stable trend in entropy points to conventionalization of an already established set of GIL choices with little temporal variation or innovation. For NEWS and WEB texts, entropy across features both increases due to a diversification of feature usage over time. This trend is in line with our expectations since the increasing social demand for gender inclusion in language seems to have pushed towards a stronger integration of explicit forms apart from already existing neutral features. The lowest overall entropy in WEB texts can be attributed to the strong dominance of epicenes in the register. It disguises, however, the great variability in explicit features.

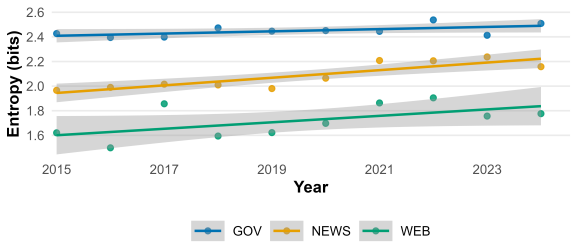


Figure 4: Entropy in German across years and registers.

Spanish For Spanish (Figure 5), we observe a more similar distribution of features across registers than in German, with epicenes being by far the most frequent ones.

GOV texts show the most visible changes in their GIL feature proportions over time. Starting out with

a high dominance of epicenes, their proportion decreases over time while both comunes (e.g. *habitantes*) and double mentions increase proportionally. Overall, the usage is restricted to rather conservative forms of GIL: neutral and visibility forms are represented, forms including gender diversity (beyond binary gender) are not used in official GOV texts in Spain.

NEWS texts show the lowest and a decreasing trend of variability in the set of GIL options used. The features used are very stable in temporal distribution with only a slight shift towards stronger epicene proportions. This stability is especially remarkable considering the high authorial variability in the NEWS register (cf. Figure 1).

WEB texts are fairly stable in their proportions of epicenes and comunes. As in the German WEB texts, epicenes show the highest proportions compared to the other registers (+50%). Also, similar to the German texts, WEB shows low but comparatively the largest proportions of explicit GIL features across registers, but with strong oscillations, reflecting more heterogeneous authorship, with personal choices of GIL rather than a concerted strategy of gender-inclusiveness. Counter to our intuition, the diversification feature x (*amigxs*) is practically not used. Also, the use of @ (*amig@* seems to decrease over time. The visibility features double mention and orthographical features like the slash (*amigos/as*) show opposite trends: double mentions proportion increases while orthographical features decrease.

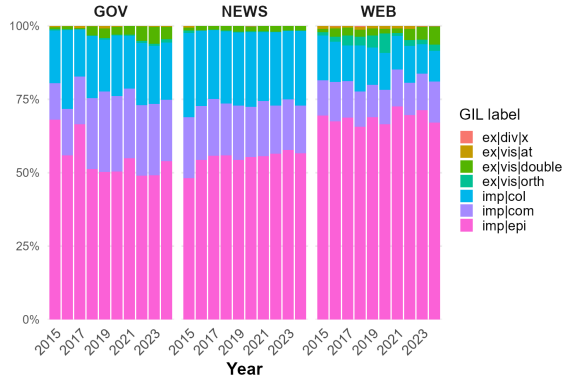


Figure 5: GIL types in Spanish across years and registers.

In probabilistic terms, the Spanish GOV texts show a notable increase in entropy, while in NEWS texts, entropy decreases (Figure 6). The increase in the GOV texts reflects the decreasing bias for epicenes, indicating an ongoing diversification with different ways of using GIL in the public discourse, possibly influenced by the change in administrations from the conservative People's Party (PP) to the left-wing Spanish Socialist Workers' Party (PSOE) in

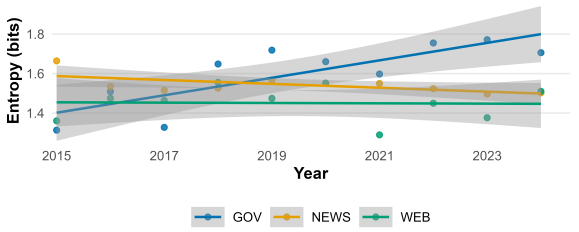


Figure 6: Entropy in Spanish across years and registers.

2018. The decrease in variability in NEWS texts is driven by a shift towards a stronger preference for epicenes and indicates conventionalization. The strong scattering of entropy values across the years in the WEB texts is an indicator of an ongoing experimentation with GIL options and temporal oscillations in preference for a specific option.

Comparing the developments in German and Spanish, we can conclude that the GOV and WEB texts in both languages behave similarly. In both German and Spanish, GIL usage is highest overall and shows the highest variability. The German GOV texts, however, start at higher frequencies and higher entropy rates, while the Spanish texts show a much milder increase in both over time. This reflects our assumptions about the time-shifted development of Spanish GIL usage compared to German. Both German and Spanish WEB texts show the highest usage of epicenes and explicit GIL features, including diversification strategies representing non-binary gender identities.

6. Conclusion

We have presented a diachronic, comparable German-Spanish corpus annotated with explicit and implicit GIL features. We have described corpus compilation and annotation, especially that of GIL features. We discussed the problem of ambiguity with collective and epicene terms and proposed a binary classification technique using a transformer model to solve ambiguous cases, achieving a classifier accuracy of +90%. The successful application of this method serves as a motivation to apply it to collective terms in future work.

A sample analysis was conducted to trace GIL usage cross-linguistically and diachronically per register. Our assumption that German spearheads GIL usage, displaying a higher GIL usage compared to Spanish, was confirmed. We also confirmed the assumption that over the past ten years, there is a diachronic GIL increase overall in both languages. In terms of register, we showed that cross-linguistically, GOV texts make the strongest use of GIL features. WEB texts display the highest number of different GIL features and highest proportions of

explicit GIL features involving grapho-stylistic disturbances such as *amig@* or *Freund*innen* facilitating explicit reference to (non-) binary gender identities.

The GILDEES resource is a valuable contribution to the study of GIL use in Spanish and German, but can also be used for other register-based and /or diachronic contrastive studies.

7. Acknowledgements

This research was funded by the Internationalization Fund of Saarland University. We thank the reviewers for their valuable comments and suggestions. We also gratefully acknowledge the annotators of the corpus: Maria Basova, Léon Maurice Jost, and Valentina Fajardo.

Limitations

Corpus annotation bears many difficulties. Apart from errors in automatic annotations, which we have tried to maintain at a minimum through effortful manual annotation and correction work, attributions of ambiguous cases are the biggest source for errors. There are cases in which even a human annotator would struggle to decide whether the term refers to a person (Example 4), or whether the label epicene (referring to a single person) or collective (referring to a group of people) would be more applicable. While Example 1 is unambiguously refers to an activity, and Example 2 refers unambiguously to a person, Example 3 stays ambiguous regarding the type of personal reference (individual or collective).

1. *Ich habe heute **Aufsicht** in der Pause.*
   *EN: I am doing **recess supervison** today.*
   (non-person).

2. *Die **Aufsicht** geht den Gang entlang.*
   *EN: The **supervisor** walks down the hallway.*
   (single person).

3. *Die **Aufsicht** signalisierte Zustimmung.*
   *EN: The **supervisory authority** signaled its approval.*
   (ambiguous: collective person/single person).

4. *Wir haben das Rundum-Sorglos-Paket: Mittagessen, **Hausaufgabenbetreuung**, sauberes Haus.*
   *EN: We have the all-inclusive-carefree package, lunch, **homework support**, clean house.*
   (person?)

Another limitation lies in deciding whether general reference to persons using epicenes and collective terms is intentional GIL usage. The assumption that GIL is used to linguistically represent more than just male gender identities implies a certain degree of intentionality. With inherently neutral forms like epicenes and collective nouns, this intentionality is impossible to prove and needs to be inferred in context. For the present study, we follow the axiom that all neutral references represent inclusive ways of reference, irrespective of the author's intention, and therefore count them as GIL features. Their unknown intentionality status, however, suggests that their inclusiveness operates on a different scale. We intend to address this issue in future, more conceptual work.

Finally, due to copyright restrictions, the corpus in its original textual form cannot be made publicly available and may only be shared with designated reviewers for research evaluation purposes; therefore, for legally compliant public dissemination, it is available in a derived format that preserves metadata and linguistic annotations while masking words which do not belong to the immediate context (preceding 4 and following 5 words) of the GIL forms prevent reconstruction of the original copyrighted texts.

8. Bibliographical References

Sonia Aguilar Gavira, Remedios Benítez Gavira, et al. 2019. Guía para un uso inclusivo del lenguaje.

Marco Baroni, Adam Kilgarriff, Jan Pomikálek, and Pavel Rychlỳ. 2006. Webbootcat. instant domain-specific corpora to support human translators. In *Proceedings of the 11th annual conference of the european association for machine translation*.

Friederike Braun, Anja Gottburgsen, Sabine Sczesny, and Dagmar Stahlberg. 1998. Können Geophysiker Frauen sein? Generische Personenbezeichnungen im Deutschen. *Zeitschrift für germanistische Linguistik*, 26(3):265–283.

Regula Bühlmann. 2002. Ehefrau Vreni haucht ihm ins Ohr... Untersuchung zur geschlechtergerechten Sprache und zur Darstellung von Frauen in deutschschweizer Tageszeitungen. *Linguistik Online*, 11(2).

Alexis Conneau, Kartikay Khandelwal, and Naman Goyal. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Raúl Cremades and Ignacio Fernández-Portero. 2022. Actitudes del alumnado universitario ante el lenguaje inclusivo y su debate en los medios de comunicación. 89:89–116.

Joke Daems. 2023. Gender-inclusive translation for a gender-inclusive sport: strategies and translator perceptions at the international quadball association. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 37–47, Tampere, Finland. European Association for Machine Translation.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Eline Demey. 2002. Leser und leserinnen gesucht! zum generischen gebrauch von personenbezeichnungen in deutschen stellenanzeigen und zeitungsartikeln. *Deutsche Sprache*, 30(1):28–49.

Anna-Katharina Dick, Matthias Drews, Valentin Pickard, and Victoria Pierz. 2024. GIL-GALaD: Gender inclusive language - German auto-assembled large database. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7740–7745. ELRA and ICCL.

Gabriele Diewald and Anja Steinhauer. 2022. *Handbuch geschlechtergerechte Sprache: Wie Sie angemessen und verständlich gendern*. Duden.

Daniel Elmiger, Eva Schaeffer-Lacroix, and Verena Tunger. 2017. Geschlechtergerechte sprache in schweizer behördentexten: Möglichkeiten und Grenzen einer mehrsprachigen Umsetzung. *Osnabrücker Beiträge zur Sprachtheorie*, 90:61–90.

Laure Gautherot. 2017. Vom Sprachfeminismus zum gendergerechten Sprachgebrauch in der BRD. *Kwartalnik Neofilologiczny*, (1):39–53.

Dirk Goldhahn, Thomas Eckart, Uwe Quasthoff, et al. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *LREC*, volume 29, pages 31–43.

Kathryn Greene and Donald L. Rubin. 1991. Effects of gender inclusive/exclusive language in religious discourse. 10(2):81–98.

Ingrid Guentherodt, Marlies Hellinger, Luise F. Pusch, and Senta Trömel-Plötz. 1980. Richtlinien zur vermeidung sexistischen sprachgebrauchs.(principes pour éviter les usages sexistes dans la langue). *Linguistische Berichte Braunschweig*, (69):15–21.

Susanne Günthner. 2019. Sprachwissenschaft und Geschlechterforschung: Übermittelt unsere sprache ein androzentrisches weltbild? In *Handbuch Interdisziplinäre Geschlechterforschung*, pages 571–579. Springer.

Pascal Gygax, Ute Gabriel, Oriane Sarrasin, Jane Oakhill, and Alan Garnham. 2008. Generically intended, but specifically interpreted: When beauticians, musicians, and mechanics are all men. *Language and cognitive processes*, 23(3):464–485.

Pascal Mark Gygax, Daniel Elmiger, Sandrine Zufferey, Alan Garnham, Sabine Sczesny, Lisa Von Stockhausen, Friederike Braun, and Jane Oakhill. 2019. A language index of grammatical gender dimensions to study the impact of grammatical gender on the way we perceive women and men. *Frontiers in psychology*, 10:1604.

Katherine J Hampares. 1976. Sexism in Spanish lexicography? *Hispania*, 59(1):100–109.

IPinfo. Ipinfo ip geolocation api. Accessed 2025-04-20.

Christiane Kaufmann and Gerd Bohner. 2014. Masculine generics and gender-aware alternatives in Spanish. *IZGOnZeit. Onlinezeitschrift des Interdisziplinären Zentrums für Geschlechterforschung (IZG)*, pages 8–17.

Manuel Lardelli, Timm Dill, Giuseppe Attanasio, and Anne Lauscher. 2024. Sparks of fairness: Preliminary evidence of commercial machine translation as English-to-German gender-fair dictionaries. In *Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies*, pages 12–21, Sheffield, United Kingdom. European Association for Machine Translation (EAMT).

M. ª Antonia Martínez Linares. 2022. Sobre los dobletes de género y cuestiones gramaticales conexas. 89:71–88.

Kimberly Matheson and Connie M. Kristiansen. 1987. The effect of sexist attitudes and social structure on the use of sex-biased pronouns. *The Journal of Social Psychology*, 127(4):395–398.

Antonia María Medina Guerra. 2016. Las alternativas al masculino genérico y su uso en el español de españa. *Estudios de lingüística aplicada*, 34(64).

Ben Papadopoulos. 2022. A brief history of gender-inclusive spanish. 48(1):31–48.

Janet B Parks and Mary Ann Roberton. 2008. Generation gaps in attitudes toward sexist/nonsexist language. *Journal of Language and Social Psychology*, 27(3):276–283.

Adam Paszke, Sam Gross, Francisco Massa, and et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, and et al. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Andrea Piergentili, Dennis Fucci, Beatrice Savoldi, Luisa Bentivogli, and Matteo Negri. 2023. Gender neutralization for an inclusive machine translation: from theoretical foundations to open challenges.

Manuel Cabello Pino. 2022. Los morfemas de género emergentes (-x y -e) y su tratamiento en la prensa española. 89:57–70.

Luise Pusch. 1979. Der Mensch ist ein Gewohnheitstier, doch weiter kommt man ohne ihn. Eine Antwort auf Kalverkämpers Kritik an Tromel-Plotz' Artikel über Linguistik und Frauensprache. *Linguistische Berichte Braunschweig*, (63):84–102.

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. Association for Computational Linguistics.

Philip Rauth. 2025. Freund:innenliche Chef:innen. Eine Korpusstudie zu Gendermovierungen. In Alexander Werth, editor, *Die Movierung: Formen, Funktionen, Bewertungen*, pages 332–364. De Gruyter, Berlin/Boston.

Kenneth Reitz. 2016. *Requests documentation*. EEUU: Requests Documentation.

Beatrice Savoldi, Eleonora Cupin, Manjinder Thind, Anne Lauscher, Andrea Piergentili, Matteo Negri, and Luisa Bentivogli. 2025. mGeNTE: A multilingual resource for gender-neutral language and translation. ADS Bibcode: 2025arXiv250109409S.

Dagmar Stahlberg, Sabine Sczesny, and Friederike Braun. 2001. Name Your Favorite Musician: Effects of Masculine Generics and of their Alternatives in German. 20(4):464–469.

Christian Stecker, Jochen Müller, Andreas Blätte, and Christoph Leonhardt. 2021. The evolution of gender-inclusive language. Evidence from the German Bundestag, 1949-2021.

Delia Esther. Suardiaz. 1973. *Sexism in the Spanish language*. University of Washington studies in linguistics and language learning ; v. 11. [Distributed by the Department of Linguistics, University of Washington], Seattle.

Carla Sökefeld. 2021. Gender (un) gerechte Personenbezeichnungen: derzeitiger Sprachgebrauch, Einflussfaktoren auf die Sprachwahl und diachrone Entwicklung. 46(1):111–141.

Senta Trömel-Plötz. 1978. Linguistik und Frauensprache.(linguistique et langage féminin). *Linguistische Berichte Braunschweig*, (57):49–68.

Thomas Wolf, Lysandre Debut, Victor Sanh, and et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Miriam Zapf. 2024. *Gender, Sprache, Kognition: Eine linguistische Untersuchung zu genderinklusivem Sprachgebrauch im Spanischen*. Linguistik in Empirie und Theorie/Empirical and Theoretical Linguistics. Springer.

9. *Appendix A.*

```
<s id="23">
1→MASK→MASK→ADV→ADV→PronType=Ind→2→advmod→_→_
2→MASK→MASK→DET→ADV→Degree=Cmp|PronType=Ind→9→nsubj→_→_
3→als→als→ADP→KOKOM→_→5→case→_→_
4→die→der→DET→ART→Case=Nom|Definite=Def|Gender=Fem|Number=Sing|PronType=Art→5→det→_→_
5→Hälfte→Hälfte→NOUN→NN→Case=Nom|Gender=Fem|Number=Sing→2→nmod→_→_
6→der→der→DET→ART→Case=Gen|Definite=Def|Number=Plur|PronType=Art→7→det→_→_
7→Teilnehmenden→Teilnehmend→NOUN→NN→Case=Gen|Number=Plur→5→nmod→_→imp|neut|nom|part
8→sind→sein→AUX→VAFIN→Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin→9→cop→_→_
9→Männer→Mann→NOUN→NN→Case=Nom|Gender=Masc|Number=Plur→0→root→_→_
10→.→.→PUNCT→$.→_→9→punct→_→_
</s>
```

Figure 7: Derived corpus publication format

A Diachronic Comparable Corpus of Spanish Digital News (2017–2026) for the Study of Stylistic Convergence in the GenAI Era

Hugo Sanjurjo-González

University of Deusto
Avda. de las Universidades, 24, 48007, Bilbao, Spain
hugo.sanjurjo@deusto.es

Abstract

This study introduces a comparable corpus of Spanish digital news (2017–2026) designed to analyze potential linguistic shifts coinciding with the widespread adoption of Generative AI. We propose an analytical framework structured across three levels: lexical statistics, semantic topology, and neural classification. By implementing a protocol of NER-masking, we isolate structural discourse markers from topical content to identify the stylistic patterns of the contemporary period. Our results suggest a measurable structural shift within the analyzed corpus, indicating a trend toward a more standardized professional register. While macro-statistical metrics like Shannon entropy remain stable —indicating statistical consistency— Zipf-Mandelbrot distributions and SVD mapping reveal a concentration of unique vocabulary into more predictable clusters. In this scenario, the 2023–2026 subcorpus exhibits a discernible topological displacement compared to the 2017–2021 baseline. The study identifies a 'Gray Zone' where highly structured technical reporting and hybridized production become indistinguishable, suggesting a structural stylistic convergence within this digital environment. These findings provide a methodological baseline for analyzing discursive stabilization in professional domains without assuming definitive authorship.

Keywords: Generative AI, Forensic Linguistics, Spanish Corpus, Diachronic Change

1. Introduction

The mass adoption of Large Language Models (LLMs) —rooted in the Transformer architecture (Vaswani et al., 2017) and scaled through systems like GPT-3 (Brown et al., 2020)— has marked a significant shift in digital discourse production. The data points toward a stylistic drift in digital news, suggesting the emergence of structural patterns that align with the generative logic of modern LLMs.

This transition poses a risk of linguistic homogenization, where generative models act as a central force that standardizes syntax and flattens lexical variance (Moon et al., 2025; Sourati et al., 2025; Ahuja et al., 2024). In Spanish, this is exacerbated by causal LLMs imposing rigid Subject-Verb-Object templates that override the language's natural syntactic fluidity and subject-omission flexibility (Busto-Castiñeira et al., 2025). Consequently, detection now requires a forensic examination of global topological configurations rather than isolated surface-level markers.

To address this, we introduce a diachronic corpus of European Spanish news, partitioned into a human baseline (2017–2021) and a hybridized period (2023–2026). We implement Named-Entity Recognition (NER) de-lexicalization to decouple structural markers from topical variance, stylistic patterns of the analyzed periods. Our framework decomposes this evolution into three dimensions:

- Statistical Linguistics: Quantifying informational dynamics through Shannon entropy and Zipf–Mandelbrot distributions (Mandelbrot, 1953) to detect systemic predictability.
- Lexical Topology: Mapping the spatial behavior of hapax legomena via density-based clustering (Ester et al., 1996) to contrast the high lexical dispersion characteristic of the 2017–2021 baseline against the 2023-2026 subcorpus.
- Neural Separability: Evaluating cohort distinguishability using a Spanish-specific Transformer (Cañete et al., 2020) and analyzing the "Gray Zone" where styles converge.

We hypothesize that contemporary journalism is evolving toward a lexical standardization pattern, a structurally cohesive but statistically more predictable configuration.

2. Related Work

Recent research documents the transition toward a linguistic ecosystem permeated by Generative Artificial Intelligence. Liang et al. (2024) identified tell-tale vocabulary spikes, while Anderson et al. (2024) demonstrated lower informational uncertainty in AI outputs, supporting the theory that generative systems follow probabilistic paths of least resistance. This structural predictability, or neural text degeneration (Holtzman et al., 2019), stems from the tendency to maximize probability over linguistic innovation.

Structural analyses in English language reveal an overall standardization: increased syntactic rigidity, higher density of logical connectors, and reduced sentence variability (Casal & Kessler, 2023). This linguistic finish (Rafique et al., 2024) and uniform punctuation (Desaire et al., 2023) result in stylistically refined but predictable texts, leading to a distributional convergence in the tails of the lexicon (Gray et al., 2024).

Proceedings of the 19th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC 2026, pages 53–61
11 May 2026.

In the Spanish domain causal decoders often impose English-like Subject-Verb-Object structures. This conflicts with Spanish's natural syntactic fluidity and subject-omission flexibility (Busto-Castiñeira et al., 2025). García-Díaz et al. (2024) confirmed these shifts in Spanish media, reporting measurable changes in adjective density and n-gram distributions.

Most Natural Language Processing approaches to AI-generated text detection rely on binary, static datasets produced under controlled prompting conditions (e.g., AuTextification, MGTBench). While valuable for benchmarking, such corpora abstract away from the actual editorial processes of real-world media production. Large-scale resources like MarIA (Gutiérrez-Fandiño et al 2022) provide high-quality reference points but lack the longitudinal perspective necessary to capture linguistic evolution. Following established frameworks on register and genre stability (Biber & Conrad, 2019), the selection of a journalistic corpus allows for a controlled environment to measure diachronic change while minimizing cross-genre noise.

In contrast, the present study adopts a diachronic and in situ approach. By analyzing a comparable corpus extracted from a digital newspaper across two distinct eras—a pre-generative AI baseline (2017–2021) and a hybridized editorial ecosystem (2023–2026)—we move beyond binary authorship detection. This design enables the investigation of whether distributional convergence and semantic compactness emerge as systemic properties of professional language. By utilizing NER-based de-lexicalization to isolate structural features from topical bias (Stamatatos, 2009), this approach filters out thematic noise to better capture stylistic evolution.

3. Materials and Methods

3.1 Corpus Construction and Stratification

The corpus was constructed through historical snapshots of 20minutos[1] from the Wayback Machine[2], using a stratified quarterly sampling protocol to ensure seasonal representativeness. We established a baseline subcorpus (2017–2021) as a human-authored reference and an experimental subcorpus (2023–2026) to capture the current hybrid production landscape—a spectrum likely encompassing varying degrees of human authorship and AI mediation—while excluding 2022 as a transitional buffer zone. The final balanced corpus consists of 1,602 news instances (n=801 per subcorpus) after random under sampling.

Metric / Category	**2017–2021**	**2023–2026**	**Δ (%)**
Avg. Words (±σ)	470.0 (±315)	496.6 (±284)	-9.8% (σ)
Emoji Usage	0.318	0.047	-85.2%
Lexical Richness (TTR)	0.537	0.532	-0.9%
Pandemic Bias	26.09%	3.00%	-23.09%
Spanish Politics	20.85%	19.10%	-1.75%

Table 1 - Descriptive Statistics and Corpus Comparability Audit

Preliminary analysis (Table 1) reveals structural stabilization (contracted σ) and stylistic sobriety (sharp decline in informal markers), while stable Type-Token Ratio (TTR) and punctuation density suggest that diachronic shifts reside in syntactic topology rather than surface metrics. Agenda consistency was confirmed via Term Frequency – Inverse Document Frequency (TF-IDF), showing an 80% lexical overlap (keywords: España 'Spain', *años* 'years', *según* 'according to', *además* 'moreover').

3.2 Data Normalization and Content-Independent Masking

To isolate the syntactic skeleton from chronological leaks or formatting artefacts, all documents underwent a multi-layered normalization and masking protocol:

- Structural Cleaning: Removal of HTML (HyperText Markup Language) tags, whitespace collapsing, and orthotopic standardization of punctuation to prevent software-based fingerprinting.
- Thematic and Temporal Masking: Neutralization of chronological markers ([YEAR]), AI-related terminology ([AI]), and high-variance topical clusters. This includes public health crises ([HEALTH]) and recurring soft-news clusters—labeled as [RECURRING_TOPIC]—which encompass service journalism and wellness terminology (e.g., lifestyle or health-trend anchors) to mitigate the influence of shifting editorial agendas on stylistic metrics.
- De-lexicalization (NER): Using spaCy's es_core_news_lg, all entities were replaced with labels ([PER], [LOC], [ORG], [MISC]), and numerical values were abstracted to [NUM].

This protocol ensures that subsequent classification relies on the discursive architecture.

[1] https://www.20minutos.es/

[2] https://web.archive.org/

Post-masking audits reveal that the hybridized period (2023–2026) exhibits a significant increase in logical connectors (*como* 'such as' +26%, *también* 'also' +21%) and complex subordinators (*aunque* 'although'), signaling a shift toward the increased connective density characteristic of the current informational style.

4. Methodology

The proposed methodology adopts a multi-level framework to analyze journalistic language across three complementary dimensions: statistical, lexical-topological, and neural. The objective is to identify systematic patterns of stylistic evolution that differentiate the human baseline (2017–2021) from the hybridized period (2023–2026).

4.1 Level I: Macro-Statistical Analysis (Information Theory)

In this stage, the corpus is treated as a stochastic system to analyze the statistical properties of language of the text. Using information-theoretic metrics, we quantify the predictability and structural distribution of journalistic language:

- Shannon Entropy (H): We compute entropy to measure information density and lexical uncertainty. A higher H value indicates a more diverse and less predictable distribution. This metric detects whether the transition to hybridized period (2023–2026) involves a flattening of information or a loss of lexical spontaneity.
- Zipf-Mandelbrot Law: Word frequency distributions are modelled to calculate the slope parameter (s). This parameter identifies the gravity of the linguistic core versus the long tail of rare words. We analyze whether the hybridized period (2023–2026) exhibits a more standardized distribution (a more rigid s) compared to the high-variance tail of the human-authored subcorpus.

4.2 Level II: Lexical-Topological Mapping (u-SVD & Clustering)

This stage focuses on the topological behavior of Hapax Legomena (terms occurring only once). Following the hypothesis that rare words carry the most authentic traces of authorship, we analyze their spatial organization:

- Unfolded Singular Value Decomposition (u-SVD): High-dimensional word embeddings are projected into a low-dimensional topological map. Unlike standard SVD, u-SVD better preserves the latent semantic divergence between terms, allowing us to visualize the geometry of the unique vocabulary.
- Density-Based Spatial Clustering (DBSCAN): We utilize DBSCAN to analyze the spatial distribution of low-frequency vocabulary. This facilitates a comparison between the high lexical dispersion characteristic of the human-authored baseline and the denser semantic clusters observed in the hybridized subcorpus (2023–2026).

By correlating these topological maps with the entropy metrics derived in Level I, we assess whether the Hapax Legomena shift from a high-entropy, scattered distribution in the human baseline toward lower-entropy, denser semantic clusters in the hybridized period. This transition would indicate a move from organic lexical diversity toward more calculated and predictable linguistic structures.

4.3 Level III: Neural Classification

The final stage validates the systematic discriminability of the two periods, proving that detected shifts represent a fundamental change in the syntactic signature rather than mere topical correlations:

- Dataset Versions: Analysis is performed on both unmasked and masked datasets to verify the robustness of the stylometric footprint.
- Neural Architecture (BETO): We fine-tuned BETO (Cañete et al, 2020), a Spanish-optimized BERT for supervised classification. Training on the masked corpus tests if the hybridized signature remains robust even after removing all contextual and thematic references.

5. Results and Discussion

5.1 Level I: Macro-Statistical Evidence (Zipf & Shannon)

The first level of analysis examines the structural complexity of the corpus through Shannon Entropy (H) and the Zipf-Mandelbrot Law. Contrary to the initial hypothesis suggesting a flattening or simplification of language in the hybridized period (2023–2026), the results reveal a slight but significant increase in informational density.

The analysis of informational density across the 2017–2026 timeline shows a remarkably stable trajectory. As evidenced in Table 2, mean lexical complexity remains consistent.

Subcorpus	Mean Entropy (H)	T-test Results
2017-2021	8.2049	t -1.56
2023-2026	8.2261	p=0.117

Table 2: Comparative Shannon Entropy (H) Results

The statistical non-significance (p>0.05) in Shannon Entropy is a finding in itself: it points to a high-fidelity mimicry between the two periods. If generative models are being integrated into the newsroom, they have successfully adopted the informational density of professional journalism. The noise and complexity of a 2026 article are indistinguishable from those of 2017 at a macro-statistical level. This suggests that the hybridized period is not simplifying the language, but rather populating pre-existing journalistic templates with an equivalent lexical density. Consequently, if a distinctive stylistic fingerprint exists in the 2023–2026 period, it must be sought not in the quantity of information, but in its topological distribution (Level II) and latent structural patterns (Level III).

While the quantity of information remains stable, its distribution reveals a different story. By fitting the frequency ranks to the Zipf-Mandelbrot Law ($f(r)=C/(r+b)^s$), we observe a clear flattening of the linguistic curve.

Metric	**2017-2021**	**2023-2026**
Slope (s)	0.8446	0.8217
Lexical Balance (1/s)	11.839	12.169

Table 3: Zipf-Mandelbrot Fit Parameters.

The decrease in the slope parameter (s) from 0.84 to 0.82 indicates that the hybridized period relies less on a few dominant 'anchor' words and more on a distributed variety of terms. This results in an increased lexical balance (1/s=12.16). These results challenge the common trope of the 'repetitive machine'; in our corpus, journalistic production from the 2023–2026 period exhibits a lower level of redundancy in its high-frequency ranges than the 2017–2021 baseline. This indicates that contemporary text generation—regardless of its human or synthetic origin—has achieved a level of lexical distribution that matches or even exceeds the structural variety of the human baseline era at a macro-statistical level. This phenomenon suggests a pattern of synthetic complexity, where the contemporary informational style incorporates a wider variety of connectors and formal lemmas (e.g., *aunque* 'although', *además* 'moreover', *también* 'also') compared to the baseline human journalistic practice. While the latter often operates under the 'Principle of Least Effort' (Zipf, 1949) and tight production deadlines, the current period exhibits a shift toward a more connective-dense and structured architecture.

At this foundational level, this shift in Zipf-Mandelbrot parameters (Table 3) provides the first empirical indication of synthetic sophistication. While 2017-2021 subcorpus is constrained by cognitive and temporal efficiency, the 2023-2026 subcorpus exhibits a 'flatter' distribution. This suggests that the discursive footprint is not characterized by the use of specific 'forbidden words', but by a systemic redistribution of lexical frequency—a texture of standardized complexity that we will further explore in the following level.

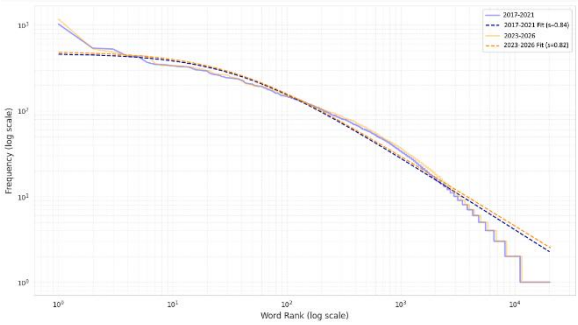


Figure 1: Zipf-Mandelbrot Law Comparison

The log-log plot (Figure 1) illustrates the rank-frequency distribution for both corpora. The hybridized period exhibits a slightly shallower slope (s=0.82), indicating a move toward higher lexical balance and a more uniform distribution of the vocabulary compared to the human baseline (s=0.84).

5.2 Level II: Topological Analysis of Low-Frequency Vocabulary

The second level moves from global metrics to local stylistic and semantic shifts, mapping how the distributional properties of the journalistic corpus have shifted.

5.2.1 Variations in Term Weighting: Comparative TF-IDF Analysis

The TF-IDF variation analysis (Figure 2) provides evidence of a systematic shift in priorities. After neutralizing thematic noise through masking, the delta in word importance reveals:

- The 2017–2021 Baseline: Significant declines are observed in terms traditionally associated with urgent, event-driven news, such as *crisis* 'crisis', *virus*[3] 'virus', and *publicación* 'publication'. News production in this baseline period appears more anchored in immediate, reactive reporting, reflecting a lexicon of disruption that has diminished in the current hybridized era.

[3] *Virus* was retained due to its polysemy and generic usage, unlike univariate terms (e.g., coronavirus) neutralized to prevent thematic bias.

The 2023–2026 Period Shift: The term *contexto* ('context') exhibits the highest positive delta, alongside structural or abstract terms such as *año* 'year', *destacar* 'to highlight', and *contenido* 'content'. This shift reinforces the hypothesis of a "meta-journalistic" framework in contemporary production, which prioritizes logical framing and discursive synthesis over raw event reporting.

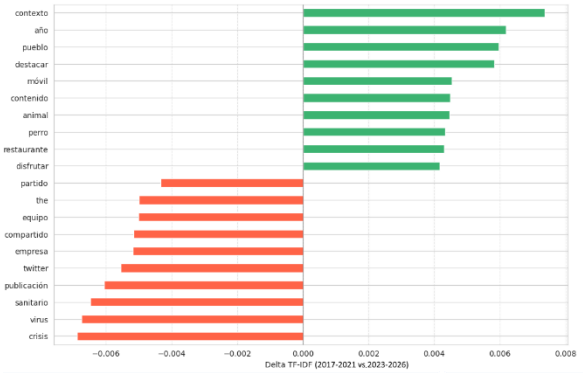


Figure 2: Top TF-IDF Weight Fluctuations.

5.2.2 Semantic Proximity and u-SVD

To confirm if this represents a fundamental change in architecture, we utilized Document-Level Embeddings using paraphrase-multilingual-MiniLM-L12-v2 (Reimers & Gurevych, 2019) processed through un-weighted Singular Value Decomposition (u-SVD) and K-Means clustering (K=10).

Cluster ID	Mean Similarity	Median	Std. Deviation	N (2017-2021/2023-2026)
Cluster 3	0.0904	0.0815	0.1104	61 / 46
Cluster 1	0.0866	0.0787	0.1078	72 / 40
Cluster 5	0.0726	0.0634	0.1096	43 / 67
Cluster 2	0.0223	0.0164	0.0909	155 / 163

Table 4. Semantic Proximity Results by Cluster (2017-2021 vs. 2023-2026).

By applying a K-Means clustering (K=10), we ensured that comparisons were made within consistent thematic neighborhoods. The cross-period cosine similarity results (Table 4) reveal a profound discursive divergence. The analysis suggests that the hybridized period exhibits a distinct topological configuration compared to the human baseline.

These findings are articulated across three strategic axes:

- Semantic divergence: Despite addressing identical topics, the similarity between human baseline and hybridized period texts falls below 0.10 across all clusters. This suggests a topological displacement where the deep semantic structure of the news now occupies an entirely different region of the latent space, placing them in different regions of the latent semantic space.
- The 2023–2026 Stylistic Pattern: A distinct structural shift has been identified. While the 2017–2021 baseline prioritizes linear, event-driven narratives focused on immediate chronology, the hybridized period exhibits a tendency toward structural abstraction. The dominance of the term *contexto* 'context' suggests a shift in the journalistic framework: the immediacy of the event is increasingly complemented—or replaced—by a synthetic discursive organization. This suggests that contemporary production (regardless of its human or synthetic origin) favors logical framing over traditional raw reporting.
- The results show a low overlap between subcorpora in the latent semantic space. Even with an 80% overlap in top vocabulary tokens, the syntactic organization is so fundamentally different that the data from our corpus suggests a significant structural transition in news production.

This structural shift provides the underlying signal that allows neural classifiers to distinguish between human and synthetic authorship with a level of precision that traditional statistical metrics fail to achieve.

5.2.3 Dispersion Analysis and Topology of Unique Vocabulary (Hapax Legomena)

Using DBSCAN (ϵ=0.20,min_samples=5), we categorized unique terms into stochastic noise and semantic clusters. The results reveal a Normalization of Rarity:

- 2017-2021: Exhibits higher stochastic noise with idiosyncratic terms (*protocolo* 'protocol', *contaminación* 'contamination').

- 2023-2026: Shows a higher concentration of clustered terms (82.33%).

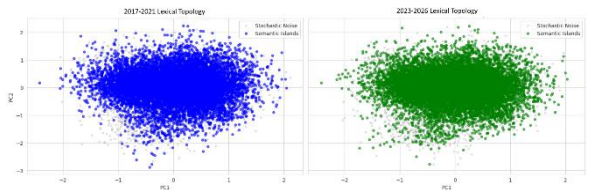


Figure 3: DBSCAN Topology of Unique Vocabulary.

The density analysis (Figure 3) suggests that the semantic clusters in the hybridized period are more cohesive (lower mean distance to centroid).

- Human baseline rareness: Characterized by "Event-Driven Rareness" (e.g., *superar* ‘to overcome’, él ‘he’, *preaviso* ‘advance notice’, *Sabanés* [a Spanish surname]), tied to specific administrative or local contexts.
- Hybridized period rareness: Characterized by "Stylistic Rareness" (e.g., *muere* ‘dies’, *selecto* ‘elite’, *underground* ‘underground’). This subcorpus exhibits low-frequency terms with high semantic cohesion, suggesting a more calculated lexical distribution in contemporary production.

This finding clarifies the semantic distribution patterns observed: while both periods maintain a similar lexical density (Entropy), the 2023–2026 subcorpus exhibits a more constrained spatial distribution of its vocabulary. Paradoxically, the rare terms in the hybridized period are more structurally predictable; they appear in highly cohesive clusters rather than the organic, idiosyncratic dispersion of the 2017–2021 baseline.

5.3 Neural Classification

We deployed a supervised classifier based on BETO to validate the three-level analysis across two parallel experiments.

5.3.1 Quantitative Performance: Structure vs. Context

The model demonstrated high robustness in both scenarios, with an increase in precision when contextual entities were present (see Table 5).

The increase in accuracy (73.52%) in the unmasked version suggests that specific entities (names of politicians, prices, dates) provide chronological anchors that help the model distinguish the eras. However, the 71.02% achieved with masks is the most significant finding, as it indicates a permanent structural change in the prose that persists even when the subject matter is hidden.

Dataset	**Best Epoch**	**Validation Loss**	**Accuracy**	**F1-Score**	**Gray Zone (N)**
Masked	3	0.5576	71.02%	70.84%	34 texts
Unmasked	2	0.5757	73.52%	75.50%	12 texts

Table 5. Comparative Performance of BETO Classifier.

5.3.2 Error Dynamics: Analysis of Classification Uncertainties

A comparative analysis of the confusion matrices reveals a shift in the model's perception.

Experiment	False Positives (2017–2021 as 2023-2026)	False Negatives (2023-2026 as 2017–2021)
Masked	69	16
Unmasked	55	38

Table 6. Error Distribution Analysis.

The classification performance in Table 6 reveals a significant shift in model behavior under masking conditions. In the Masked scenario, Accuracy drops to 71.02%, primarily driven by a surge in False Positives (N=69). Conversely, the Unmasked scenario shows a higher Accuracy (73.52%) and a notable reduction in False Negatives (16). This indicates that while thematic entities act as predictive markers, their removal exposes a deeper structural convergence between the two periods.

5.3.3 Qualitative Synthesis of the Gray Zone

The Gray Zone (where P≈0.50) serves as a case study of structural convergence where stylistic boundaries between periods collapse. Analysis of these ambiguous texts identifies two distinct phenomena:

- Technical Formalization: Journalistic production focused on "service news"—such as automotive specifications or energy auctions—exhibits a shift toward modular logic. The use of lists, rigid technical data, and dry formatting creates a stylistic overlap. In these cases, the high degree of structural optimization in the 2017–2021 baseline mimics the standardized discursive patterns that have become dominant in the 2023–2026 subcorpus.

- Informative Stabilization: Contemporary texts concerning specialized topics—such as medical monographs or clinical symptoms—frequently bypass neural classification by successfully maintaining the discursive conventions of professional objectivity. By bridging the Contextual Gap identified in Level II, this structural "smoothness" becomes indistinguishable from traditional technical reporting. In these cases, the 2023–2026 production aligns with the longstanding standards of medical and scientific journalism, suggesting a point of formal stabilization where the period's signature is absorbed by the genre's own rigidity.

6. Conclusions

6.1 Main findings

This study provides a three-level characterization of the linguistic evolution within the analyzed corpus. Our findings suggest that 2023-2026 subcorpus exhibits shifts that are not only thematic but also structural and topological when contrasted with the 2017–2021 baseline.

The transition is characterized by a smoothing of the language. While superficial entropy suggests maintained variety, the adherence to Zipf's Law and the reduction in the alpha parameter (1.10→1.07) indicate that journalistic output has become more statistically predictable. The 2023-2026 era exhibits a standardization of rarity: unique words are no longer idiosyncratic outliers of human expression but are organized into dense, cohesive stylistic blocks that follow the probabilistic logic of generative models.

Semantic proximity analysis reveals high degree of semantic divergence in the human baseline and hybridized period (similarity <0.10). While human baseline traditionally relies on event-driven, linear narratives, texts from hybridized period prioritize meta-journalistic framing. Using terms such as *contexto* 'context' and other logical anchors, which serve to compensate for the model's lack of direct, situated reporting.

Neural classification using BETO achieved an accuracy of 73.52%, proving that a robust synthetic footprint exists even after rigorous NER-masking. However, the qualitative analysis of the Gray Zone identifies a notable convergence in technical registers:

- Journalism as algorithm: Technical and data-saturated news authored by humans (False Positives) is increasingly indistinguishable from hybridized period due to its modular and formulaic structure.
- Stylistic Continuity: In contrast, the model fails to identify a distinctive structural signal in social and emotional contexts (e.g., obituaries or human-interest stories). In these clusters, the 2023–2026 production maintains a high similarity with the 2017–2021 baseline. This suggests that either the traditional style in these areas is inherently formulaic—relying on established 'sentimental tropes'—or that the hybridized signature characteristic of this period effectively preserves the conventional formalisms of the genre.

In conclusion, the results of this three-level analysis suggest a structural-discursive divergence between the 2017–2021 and 2023–2026 subcorpora. Although language is inherently dynamic and subject to temporal shifts (Hamilton et al., 2016), the contemporary period is characterized by a notable trend toward standardized informational styles, particularly in technical and data-dense news.

Rather than a total rupture, we observe a topological overlap where journalistic production naturally aligns with the synthetic logic prevalent in current digital environments. The presence of a Gray Zone in our neural classification (Table 6) confirms that the boundary between highly structured traditional reporting and emerging standardized patterns has become increasingly porous. This suggests that the distinctive signature of the current period lies in a formal stabilization of the journalistic craft, where professional routines and automated structures have converged into a shared, era-defining discursive architecture.

6.2 Limitations

Despite the robustness of the three-level analysis, several limitations must be acknowledged to contextualize the findings:

- Temporal Proxy vs. Ground Truth: The primary limitation is the use of a temporal boundary as a proxy for AI integration. Since newsrooms do not explicitly disclose the extent of LLM usage for each article, this study identifies stylistic shifts in the era of AI rather than definitively labelling individual texts as AI-written. The 2023-2026 subcorpus likely contains a spectrum of human-only, AI-assisted, and AI-generated content.
- Linguistic and Geographical Scope: The corpus is strictly limited to a European Spanish national newspaper. Consequently, the findings regarding lexical density and the contextual gap may be influenced by specific Spanish journalistic traditions and may not be directly generalizable to regional press, other languages, or different journalistic

cultures (e.g., Anglo-Saxon or Asian media).

- The Event-Driven Bias: Although the unmasked model attempted to control for this, certain black swan events (e.g., geopolitical crises or specific legal changes in the 2023-2025 period) might introduce unique vocabulary that the classifier could mistake for an 2023-2026 signature. While masking entities mitigates this, the differences in semantic distribution observed could still be partially influenced by the shifting nature of global news cycles.
- Classifier Opacity: While BETO provides a high degree of accuracy, neural models remain black boxes to some extent. The identification of the Gray Zone is a qualitative interpretation of statistical confidence; further research using Explainable AI (XAI) tools like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) would be required to pinpoint the exact tokens triggering the classification.

6.3 Future work

This study establishes a foundational diachronic framework for analyzing emergent patterns of standardization in professional digital discourse. Future work should focus on the longitudinal integration of this corpus into forensic linguistics pipelines, enabling the detection of stylistic drift across different journalistic traditions. By treating this corpus as a benchmark for hybridity, our methodology provides the necessary evidence to calibrate tools that go beyond simple authorship attribution, moving towards a deeper understanding of the technological impact on professional linguistic registers.

7. Acknowledgments

The author would like to thank the anonymous reviewers for their insightful comments and constructive suggestions, which significantly helped to improve the clarity and rigor of this manuscript.

8. Bibliographical References

Ahuja, S., Gumma, V., & Sitaram, S. (2024). Contamination report for multilingual benchmarks. *arXiv*. https://doi.org/10.48550/arXiv.2410.16186

Anderson, B., Ganehandran, G., & Thompson, R. (2024). The Entropy of Artificial Language: Stylometric analysis of LLM-generated vs. Human-written text. *Journal of Computational Linguistics and Stylometry*, 12(2), 145-168.

Casal, J. E., & Kessler, M. (2023). Can linguists distinguish between ChatGPT-generated and human-written research abstracts? A corpus-based analysis. *Research Methods in Applied Linguistics*, 2(3), 100068.

Desaire, H., Chua, A. E., Isom, M., Hua, R., & Schorno, R. (2023). Distinguishing ChatGPT from human writing: Machine learning on specific features. *Cell Reports Physical Science*, 4(6), 101426.

García-Díaz, V., Valencia-García, R., & Colomo-Palacios, R. (2024). Stylometric evolution in Spanish digital media: The impact of ChatGPT on journalistic standards. *International Journal of Information Management Data Insights*, 4(1), 100215.

Gray, J., Rogers, A., & Marcus, G. (2024). The Centripetal Force of AI: Homogenization of the Digital Commons. *Nature Machine Intelligence*, 6, 212-225.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 1489–1501, Berlin, Germany, August. Association for Computational Linguistics.

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2024). Monitoring AI-modified content at scale: A case study on the surge of “delve” and other telltale words in scientific abstracts. *arXiv preprint* arXiv:2403.07183.

Moon, K., Green, A. E., & Kushlev, K. (2025). Homogenizing effect of large language models (LLMs) on creative diversity: An empirical comparison of human and ChatGPT writing. *Computers in Human Behavior: Artificial Humans*, 6, 100207. https://doi.org/10.1016/j.chbah.2025.100207

Pastor-Galindo, J., Zago, M., Nespoli, P., Bernal, S., & Huertas Celdrán, A. (2023). The Style of GPT: A comparative analysis of AI-generated news vs. traditional press in Spanish. *Expert Systems with Applications*, 230, 120531.

Rafique, M., Ahmed, S., & Shafi, J. (2024). The “Polishing” Effect: How LLMs standardize discourse and pragmatic politeness. *AI & Society*, 39(1), 89-104.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, November. Association for Computational Linguistics. URL: http://arxiv.org/abs/1908.10084

Sourati, Z., Karimi Malekabadi, F., Ozcan, M., McDaniel, C., Ziabari, A. S., Trager, J., Tak, A. N., Chen, M., Morstatter, F., & Dehghani, M. (2025). The shrinking landscape of linguistic

diversity in the age of large language models. *arXiv*. https://doi.org/10.48550/arXiv.2502.11266

Align and Shine: Building High-Quality Sentence-Aligned Corpora for Multilingual Text Simplification

Kenji Hilasaca, Nouran Khallaf, Serge Sharoff
Centre for Translation, Localisation and Interpreting Studies
School of Languages, Cultures and Societies
University of Leeds, UK
pjbd103@leeds.ac.uk, n.khallaf@leeds.ac.uk, s.sharoff@leeds.ac.uk

Abstract

Text simplification plays a crucial role in improving the accessibility and comprehensibility of written information for diverse audiences, including language learners and readers with limited literacy. Despite its importance, large-scale, high-quality datasets for training and evaluating text simplification models remain scarce for languages other than English. This paper reports an experimental study on the collection and processing of crowd-sourced simplification data from comparable corpora to construct a corpus suitable for both training and testing text simplification systems across multiple languages (Catalan, English, French, Italian and Spanish). We report mechanisms for sentence-level alignment from document-level data. The resulting dataset of the aligned sentence pairs is publicly available.

Keywords: Sentence alignment; Crowdsourcing; Multilinguality; Readability

1. Introduction

Automatic text simplification plays a crucial role in improving the accessibility and comprehensibility of written information for diverse audiences, including language learners and readers with limited literacy (Saggion, 2017). Data needed for training automatic text simplification tools are based on aligned sentences. This alignment at the sentence level, rather than at the document level, is essential for supervised learning approaches, as it enables models to learn specific simplification operations and their contextual application. A sentence-aligned corpus is also essential for evaluation of zero-shot approaches.

Despite its importance, large-scale, high-quality publicly available datasets for training and evaluating text simplification models remain scarce for languages other than English, which has such datasets as ASSET (Alva-Manchego et al., 2020) and Wikipedia-derived ParallelSEW (Coster and Kauchak, 2011). This paper reports an experimental study on collecting and processing of crowd-sourced simplification data to construct a corpus suitable for both training and testing text simplification systems across multiple languages. We report mechanisms for sentence-level alignment from document-level data. The resulting dataset, together with the aligned sentence pairs, is publicly available.[1] This is the first large open-source corpus for text simplification in Catalan and Spanish. The parallel corpus is also consistent across the languages with respect to its genre, which will help with cross-lingual evaluation of text simplification models.

2. Related studies

Early work on machine translation highlighted both the value and the limitations of domain-specific resources such as the European Parliament corpus (Koehn, 2005) and the United Nations corpus (Ziemski et al., 2016). The limitations on the amount and diversity of texts motivated large-scale mining from comparable corpora, a line of research that ultimately contributed to the pre-training data pipelines for Large Language Models (Sharoff et al., 2023). Within this paradigm, Wikipedia articles connected via iWiki links proved a particularly productive source for extracting translation pairs (Adafre and de Rijke, 2006; Schwenk et al., 2019), owing to their broad topical coverage and cross-lingual parallelism. The sentence alignment methods have also evolved consid-

[1] https://github.com/kenjihilasak/Align-and-Shine

Proceedings of the 19th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC 2026, pages 62–71
11 May 2026.

erably, from early character-based approaches (Gale and Church, 1993) through more advanced statistical models (Varga et al., 2007) to contemporary neural architectures (Jiang et al., 2020).

Similar efforts extracted complex-simple pairs from web corpora (Brunato et al., 2016) or machine-translated Wikipedia alignments to construct multilingual resources (Cardon and Grabar, 2020). While these were consolidated into a single corpus (Ryan et al., 2023), the result suffers from noise and inconsistent annotation due to domain mismatch.

Our contribution addresses this gap directly: we conduct alignment experiments on a novel, reliable dataset drawn from a single source domain across multiple languages, using modern sentence-alignment tools.Furthermore, while recent advancements have introduced aligners, such as CATS (Štajner et al., 2018), as well as robust multilingual models like Bertalign (Niklaus et al., 2026), our current focus is on evaluating the impact of distinct semantic representation paradigms within the hybrid SentAlign framework, leaving the comparison against these newer aligners for future work.

3. Methodology

Our study proposes a two-phase methodology for identifying parallel sentences in document-aligned simplification corpora. First, we evaluate traditional surface-level baselines against modern semantic embedding methods using a manually annotated gold standard. This evaluation phase allows us to identify the best-performing embedding model and tune the optimal cosine similarity threshold (τ) to filter out noise and verbatim copies for each language. In the second phase, we apply these optimal, language-specific configurations to a large Web-derived comparable corpus to extract a large-scale dataset suitable for text simplification. To accurately capture text simplification operations, which inherently involve compression, summarization or explanation, our alignment pipeline utilizes an asymmetric search strategy that supports flexible N-to-M mappings.

3.1. Baseline Approaches

We evaluate two established language-independent baselines to quantify the benefits of neural embeddings.

The **Gale and Church** algorithm (Gale and Church, 1993) assumes correlated sentence lengths in parallel texts. While this assumption is often violated in text simplification due to sentence splitting or compression, it is evaluated here because it forms the heuristic pre-selection stage of the primary framework (SentAlign). Testing it independently serves as a necessary ablation baseline to demonstrate the value added by the subsequent neural semantic anchoring stage

Hunalign (Varga et al., 2007) augments length heuristics with a dictionary built from parallel texts. While effective for translation alignment, its reliance on lexical overlap limits performance in text simplification tasks.

3.2. Experiments with SentAlign

To overcome the limitations of length-based methods, we use **SentAlign** (Steingrimsson et al., 2023), a hybrid alignment algorithm combining neural embeddings with a three-stage pipeline:

1. **Heuristic Pre-selection:** SentAlign reduces the search space by generating candidate alignments using the Gale and Church (1993) algorithm, which relies on character length ratios.
2. **Semantic Anchoring:** Candidates are validated using the chosen embedding model (Section 3.3). Those exceeding a high-confidence cosine similarity threshold become *anchors*. This establishes fixed points in the document map that partition the text into smaller segments.
3. **Global Optimization:** The algorithm aligns segments between anchors using Dijkstra's shortest path algorithm, with costs derived from the cosine similarity matrix. We configure this stage to prioritize the simplified document as the reference, enabling an asymmetric search that retrieves the closest semantic equivalent(s) in the complex document for each simple sentence. This easily accommo-

Table 1: Initial Wikipedia / Vikidia corpus. #Docs is the document count (paired Wikipedia–Vikidia articles by topic). #Words and #Sentences are totals across all documents in each subset. IQR gives the inter-quartile (25% to 75%) range of sentence lengths in words in each subset.

		Wikipedia			Vikidia		
Language	**#Docs**	**#Words**	**#Sent's**	**IQR**	**#Words**	**#Sent's**	**IQR**
Catalan	179	396,277	16,813	(15, 29)	19,394	1,000	(13, 23)
English	2,585	8,281,625	340,924	(16, 29)	424,306	22,462	(13, 22)
Spanish	3,875	7,946,169	301,241	(16, 33)	607,990	27,825	(14, 26)
French	33,438	46,618,143	1,945,046	(15, 29)	6,643,567	320,372	(14, 25)
Italian	3,902	6,790,163	263,271	(16, 32)	537,723	25,202	(14, 26)

dates flexible N-to-M mappings (e.g., 1-to-2, 1-to-0, or 2-to-1), successfully capturing simplification operations such as splitting, deletion, and summarization.

3.3. Comparison of Semantic Representations

Our experiment focuses on evaluating the impact of different semantic representations during the *Semantic Anchoring* and *Global Optimization* stages of SentAlign. We compare three multilingual encoders:

- **LaBSE** (Feng et al., 2022): A translation-optimized model, traditionally strong in cross-lingual alignment. We assess its transferability to monolingual text simplification tasks.
- **BGE-M3** (Xiao et al., 2024): Designed for Retrieval-Augmented Generation (RAG) and semantic search, BGE-M3 handles multi-granularity inputs. We hypothesize it may excel in simplification due to its focus on retrieving the most relevant sentences, and its ability to manage structural and length disparities between complex and simple text.
- **SONAR** (Duquenne et al., 2023): Developed under the *No Language Left Behind* (NLLB) project, SONAR supports over 200 languages using a distinct architecture. We test whether its massive multilingual capacity improves alignment for lower-resource languages (e.g., Catalan in our case) compared to BERT-based models.

4. Evaluation Setup

We start with the initial corpus, which has been crawled from Vikidia[2], a website that maintains Wikipedia-style content aimed at "children and anyone seeking easy-to-read content". For each Vikidia document, we added the corresponding Wikipedia article in the same language to form comparable document pairs. Stub articles (with little content at the moment) have been discarded. The total amount of data across all languages is listed in Table 1. The entries have been aligned between the two versions when they had identical headings, hence we have the same document count for each language. However, the Vikidia entries are much shorter, so the sentence count for Vikidia is 10-15 times smaller for our languages apart from French, as Vikidia is much more popular in the French-speaking world. The IQR column shows that the sentences in Vikidia are consistently shorter, but not much shorter than their Wikipedia counterparts.

To assess the performance of the different alignment configurations, we perform an intrinsic evaluation focusing on the accurate retrieval of simplified sentence pairs.

4.1. Gold Standard Creation

To create a reliable ground truth for our intrinsic evaluation, we randomly selected 15 document pairs for each language. The sentence alignment for these documents was performed manually from scratch by one annotator per language.

To ensure high inter-annotator agreement and to capture the nature of text simplification more accurately, the annotators followed our

[2] https://www.vikidia.org/

guidelines. Two sentences (or groups of sentences) were considered aligned if they exhibited a clear semantic correspondence, specifically encompassing the following accepted simplification operations:

- **Lexical Simplification (1-to-1):** Substituting complex vocabulary or idioms with accessible equivalents.
- **Syntactic Simplification:** Restructuring complex grammar (e.g., passive to active voice).
- **Sentence Splitting (1-to-N):** Breaking a long complex sentence into multiple shorter ones.
- **Summarization (N-to-1):** Condensing peripheral details from multiple sentences into a single core sentence.
- **Block Summarization (N-to-N):** Reorganizing complex paragraphs into a different number of simplified sentences while maintaining semantic equivalence.
- **Deletion:** Omitting overly technical or tangential sentences entirely (deliberately left unaligned).

4.2. Evaluation Metrics and Thresholding

For each alignment tool and embedding variant, we compute the similarities across the entire paired document. The performance is then measured using standard retrieval metrics: **Precision**, **Recall**, and **F_1-score**. These metrics are calculated by comparing the system's output against our manually annotated Gold Standard using the official SentAlign evaluation script[3].

Because the raw output includes pairs with varying degrees of semantic overlap, we implement a filtering mechanism to isolate actual simplifications:

- **Lower-bound Threshold (τ):** For each language and model configuration, we establish a specific threshold (τ) tuned via grid search to maximize the F_1-score on the test sample.
- **Upper-bound Threshold (0.95):** To explicitly capture text where simplification operations have occurred (e.g., paraphrasing, lexical substitution, or splitting), we discard sentence pairs with a cosine similarity score > 0.95. This removes exact or near-exact copies (with minor variations in punctuation) that provide no learning signal for text simplification models, and tend to plague such alignments as WikiLarge (Cardon and Grabar, 2020). The upper bound threshold was found to be acceptable for all languages and for all embedding frameworks.

Only the sentence pairs whose scores fall within this $[\tau, 0.95]$ range are considered acceptable alignments and passed to the evaluation script. We consolidate the results using these optimal thresholds, yielding two evaluation settings based on how forgiving the scoring is when the algorithm makes a "partial match" compared to the Gold Standard. **Strict Evaluation** requires an exact group match (e.g., if the Gold Standard establishes a 2-to-1 summarization like [1, 2] $\rightarrow$ [3], the algorithm only gets credit if it predicts exactly [1, 2] $\rightarrow$ [3]). Conversely, **Lax Evaluation** allows for partial overlap, rewarding the algorithm if it successfully finds a valid, albeit incomplete, semantic connection (e.g., predicting [1] $\rightarrow$ [3] when the true label is [1, 2] $\rightarrow$ [3]).

5. Results and Discussion

5.1. Quantitative Analysis

Table 2 presents the comprehensive alignment performance across all five languages. We report the two surface-feature baselines alongside the SentAlign framework instantiated with three embedding spaces (LaBSE, BGE-M3, and SONAR). For each neural model, we include both raw output (without threshold filtering) and optimised output (using the best threshold τ per model and language), enabling a direct comparison between the intrinsic structure of each vector space and its tuned performance.

Baseline methods. There is a stark contrast between traditional and embedding-based ap-

[3] https://github.com/steinst/SentAlign/blob/master/evaluation/evaluate.py

Table 2: Comprehensive alignment results comparing raw outputs (None) and optimized threshold outputs (τ). **Strict** requires exact matches, while **Lax** rewards partial semantic overlaps. Best F_1 scores per language are highlighted in bold.

			Strict Evaluation			Lax Evaluation		
Language	**Algorithm**	**Threshold (τ)**	**P**	**R**	**F_1**	**P**	**R**	**F_1**
Catalan	Gale-Church	None	0.000	0.000	0.000	0.000	0.000	0.000
(ca)	Hunalign	None	0.000	0.000	0.000	0.000	0.000	0.000
	LaBSE	None	0.215	0.950	0.351	0.223	0.952	0.361
	LaBSE	**0.61**	**0.526**	0.833	**0.645**	**0.547**	0.839	**0.662**
	BGE	None	0.140	0.667	0.232	0.182	0.722	0.291
	BGE	0.58	0.276	0.617	0.381	0.366	0.681	0.476
	SONAR	None	0.160	0.433	0.233	0.215	0.507	0.302
	SONAR	0.56	0.302	0.267	0.283	0.396	0.323	0.356
English	Gale-Church	None	0.000	0.000	0.000	0.000	0.000	0.000
(en)	Hunalign	None	0.020	0.090	0.030	0.020	0.090	0.030
	LaBSE	None	0.218	0.672	0.330	0.277	0.722	0.400
	LaBSE	0.67	0.570	0.414	0.480	0.688	0.460	0.552
	BGE	None	0.111	0.711	0.192	0.139	0.755	0.235
	BGE	**0.73**	0.562	0.492	**0.525**	**0.634**	0.522	**0.573**
	SONAR	None	0.188	0.367	0.249	0.264	0.449	0.332
	SONAR	0.53	0.325	0.289	0.306	0.421	0.345	0.379
Spanish	Gale-Church	None	0.000	0.000	0.000	0.000	0.000	0.000
(es)	Hunalign	None	0.070	0.270	0.110	0.070	0.270	0.110
	LaBSE	None	0.135	0.694	0.227	0.171	0.742	0.279
	LaBSE	**0.67**	**0.418**	0.426	**0.422**	**0.509**	0.475	**0.491**
	BGE	None	0.074	0.500	0.129	0.103	0.581	0.175
	BGE	0.71	0.330	0.343	0.336	0.464	0.423	0.443
	SONAR	None	0.146	0.333	0.203	0.215	0.424	0.285
	SONAR	0.58	0.253	0.222	0.236	0.263	0.229	0.245
French	Gale-Church	None	0.000	0.000	0.000	0.000	0.000	0.000
(fr)	Hunalign	None	0.030	0.090	0.040	0.030	0.090	0.040
	LaBSE	None	0.166	0.660	0.266	0.201	0.701	0.312
	LaBSE	**0.62**	**0.451**	0.489	**0.469**	**0.510**	0.520	**0.515**
	BGE	None	0.061	0.340	0.104	0.126	0.516	0.202
	BGE	0.70	0.292	0.223	0.253	0.569	0.360	0.441
	SONAR	None	0.111	0.255	0.155	0.190	0.369	0.251
	SONAR	0.54	0.190	0.202	0.196	0.240	0.242	0.241
Italian	Gale-Church	None	0.000	0.000	0.000	0.000	0.000	0.000
(it)	Hunalign	None	0.070	0.350	0.110	0.070	0.350	0.110
	LaBSE	None	0.129	0.831	0.223	0.147	0.849	0.250
	LaBSE	**0.64**	**0.448**	0.727	**0.554**	**0.488**	0.744	**0.589**
	BGE	None	0.072	0.662	0.130	0.090	0.711	0.160
	BGE	0.78	0.494	0.494	0.494	0.597	0.541	0.568
	SONAR	None	0.119	0.416	0.186	0.138	0.451	0.211
	SONAR	0.55	0.213	0.351	0.265	0.236	0.375	0.290

proaches. Gale-Church and Hunalign fail to produce viable alignments in the text simplification setting. Both methods rely on sentence-length correlation and lexical overlap, assumptions that break down when confronted with the heavy paraphrasing, sentence splitting, and compression that characterise Wikipedia-to-Vikidia aligned corpus. Consequently, neither baseline yields F_1-scores above chance, confirming that surface-level features are insufficient for this task.

LaBSE. Among the neural approaches, **LaBSE** demonstrates consistent superiority across the Romance languages. In the optimised Strict evaluation, it achieves the highest F_1-scores in Catalan (0.645), Spanish (0.422), French (0.469), and Italian (0.554). Its sta-

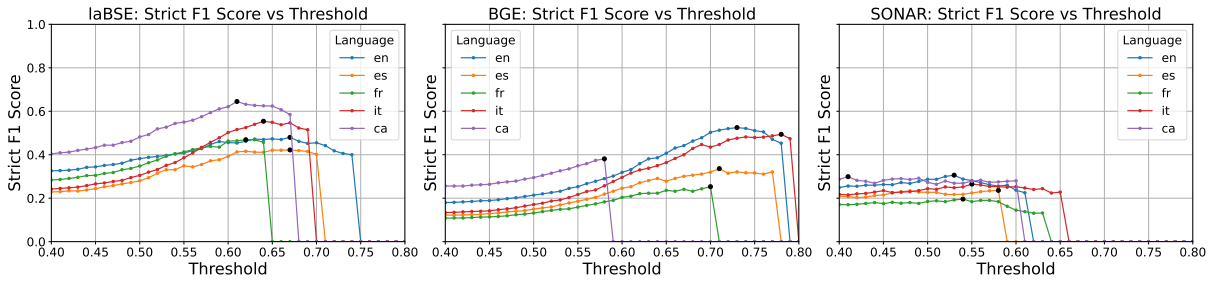


Figure 1: Plotting Strict F_1-scores across cosine similarity thresholds (τ) for the five languages for each embedding method. The black dots indicate the optimal threshold that maximizes the F_1-score for each language and for each method. Notice the overall height superiority of LaBSE, the rightward shift of the BGE peaks reflecting its similarity distribution, and the compressed performance of SONAR.

bility across both high- and lower-resource settings suggests that LaBSE's translation-ranking training objective produces a semantic space well suited to detecting monolingual paraphrases, a capability that directly benefits the simplification alignment task.

BGE-M3. **BGE-M3** exhibits a clear domain-specific advantage in English, where it is the only model to surpass LaBSE (Strict F_1: 0.525 vs. 0.480). This is consistent with its architecture, which is heavily optimised for English-centric retrieval tasks. However, performance degrades markedly in French (Strict F_1: 0.253) and Spanish (0.336), pointing to limited cross-lingual stability. The model's strong English performance should therefore be interpreted in light of this imbalance rather than as evidence of general robustness.

SONAR. Despite its extensive multilingual pre-training, **SONAR** trails both LaBSE and BGE across all languages, reaching a maximum Strict F_1 of only 0.306 in English. We attribute this to a mismatch between SONAR's training objective and the demands of the present task: while SONAR excels at cross-lingual alignment, its vector space appears to lack the fine-grained monolingual resolution required to reliably distinguish a genuine simplification from a merely topically related sentence within the same language.

Strict vs. Lax evaluation. Across all optimized models, **Lax F_1-scores** consistently exceed their Strict counterparts. LaBSE's English score, for instance, rises from 0.480 (Strict) to 0.552 (Lax). This gap confirms that the SentAlign framework captures partial simplification operations, such as 1-to-N sentence splits, that are penalised under exact-match criteria despite constituting semantically valid alignments. The Lax metric thus provides a more faithful upper-bound estimate of true alignment quality.

5.2. Impact of Thresholding

The inclusion of unfiltered metrics in Table 2 highlights the critical role of threshold selection in semantic alignment. In the absence of filtering, models default to near-exhaustive recall at the cost of precision: LaBSE in Catalan, for example, achieves a raw recall of 0.950 while precision collapses to 0.215, as the algorithm aligns almost every target sentence regardless of semantic relevance. Applying the optimal τ corrects this imbalance, raising precision to 0.526 and restoring a competitive F_1-score.

Figure 1 plots the Strict F_1-score as a function of τ for all three models and five languages on a unified scale. The shape and position of the curves are informative about the underlying geometry of its vector space:

LaBSE produces the highest performance across most languages, with optimal thresholds clustered between $\tau = 0.60$ and 0.67. Its curves rise gradually before dropping sharply to zero beyond the optimum, a cliff-edge pattern indicating that overly conservative thresholds aggressively discard valid paraphrases. This

sensitivity reinforces the necessity of careful per-language tuning.

BGE-M3 shows optimal thresholds shifted substantially rightward, predominantly between 0.70 and 0.78. This reflects systematically higher cosine similarity scores in BGE-M3's vector space, requiring a stricter τ to separate true matches from background noise. The model also exhibits the greatest cross-lingual variance: English and Italian peak comparably to LaBSE, while French remains markedly depressed, consistent with the quantitative results in Table 2.

SONAR occupies the bottom of the performance range across all languages. Its curves not only peak lower but decay earlier, collapsing before $\tau = 0.65$. This premature decay reflects the model's difficulty in producing high-confidence semantic links for monolingual simplification alignment, corroborating the quantitative findings and suggesting a fundamental mismatch between SONAR's pre-training regime and the demands of this task.

5.3. Qualitative Analysis and Limitations

To complement our quantitative evaluation, we conducted a manual inspection of the alignment outputs across the five languages to understand the models' behavior and identify typical errors. We categorize the aligned pairs into three distinct regions based on their similarity scores, revealing both the strengths of the pipeline and the inherent challenges of mining parallel data from independently edited wikis.

The Upper Bound: Verbatim copies (> 0.95). A distinct class at the extreme high end of the distribution consists of sentence pairs that are essentially identical in wording. These arise when Vikidia retains a Wikipedia sentence verbatim. Qualitative inspection confirms that these pairs typically involve proper nouns, numerical data, or definitional statements that resist paraphrasing. As noted in Section 3, pairs scoring above 0.95 are explicitly excluded from the final corpus because they carry no simplification signal and would teach a generation model to simply copy the source text.

The Sweet Spot: Genuine simplification. In the high-to-mid scoring range, the models successfully capture true simplification operations without semantic drift. In these optimal alignments, the Vikidia segments typically contain shorter sentences, exhibit reduced syntactic complexity, substitute technical vocabulary with more accessible alternatives, and omit parenthetical qualifying clauses, all while keeping the propositional core intact. This confirms that contextual embeddings are well-calibrated for detecting semantic equivalence even when structural changes are drastic (e.g., 1-to-N sentence splits).

The Threshold Limit: Topical overlap vs. Propositional equivalence. The most typical errors in our pipeline occur close to the τ threshold limit. Across all five languages, sentence pairs in this lower-scoring boundary often share a common topic and named entities—creating a "topical illusion"—but diverge in propositional content. Both sentences discuss the same subject, yet they assert different facts or approach the topic from distinct perspectives. This reflects a fundamental property of the Vikidia corpus: a substantial proportion of its sentences were not produced by directly simplifying their Wikipedia counterparts, but were instead written independently by editors. Consequently, the pipeline's most common error is aligning sentences that are topically related but semantically mismatched, which strongly justifies the necessity of our strict, empirically tuned thresholds to filter out this noise.

5.4. Whole corpus alignment

With better understanding of the best alignment parameters, we have applied the best models (LaBSE with the respective τ for all languages except English, where BGE was used) to each document pair of the full corpus from Table 1. The results are presented in Table 3. On average, 5% of the original sentences in the Vikidia corpora find high-confidence equivalents in the respective Wikipedia articles. This strict filtering significantly reduces

Table 3: Aligned Wikipedia / Vikidia corpus. Meaning preservation needs to be close to 1. For the Simplification metrics, Δ SDepth and Δ NDense represent the average change (Target - Source) within each alignment block.

	Wikipedia		Vikidia		Meaning	Simplification	
Language	**#Words**	**#Sent's**	**#Words**	**#Sent's**	**BERTScore**	**Δ SDepth**	**Δ NDense**
Catalan (LaBSE)	13,658	501	8,455	464	0.902	-0.98	+4.07%
English (BGE)	115,326	5,058	79,372	4,569	0.913	-0.78	+2.18%
Spanish (LaBSE)	204,928	7,883	153,718	7,209	0.915	-0.42	+1.62%
French (LaBSE)	2,591,525	123,280	2,046,642	108,301	0.901	-0.23	+1.20%
Italian (LaBSE)	260,142	11,131	185,681	9,846	0.908	-0.44	+1.31%

data volume but guarantees a noise-free corpus, which is crucial in text simplification to prevent model hallucinations and teach genuine simplification rather than loose semantic similarities.

To validate the quality of the final aligned corpus, we perform an automated linguistic assessment focusing on two core dimensions of text simplification: Meaning Preservation and Structural Simplification. For this analysis, we evaluate the optimal alignments generated by our best-performing models per language as established in our previous tests.

Meaning Preservation: We utilize BERTScore (Zhang et al., 2020) to measure the semantic equivalence between the complex source and the simplified target. Crucially, to accurately evaluate N-to-M alignments (such as sentence splits or multi-sentence summarizations), we concatenated the sentences within each aligned group before computing the score. This ensures the contextual embedding model evaluates the full semantic unit. As shown in Table 3, a high average BERTScore F_1 across all languages (consistently > 0.90) confirms that the SentAlign pipeline, when constrained by selecting optimal thresholds (τ) and upper bound filter (< 0.95), successfully extracts aligned pairs that are very similar in their meaning.

Structural Simplification: To verify that the target sentences are genuinely simpler, we parsed the sentences using SpaCy to compute two syntactic metrics: Maximum Tree Depth and Noun Phrase (NP) Density. For alignments containing multiple sentences, NP Density was calculated by aggregating the total number of NPs divided by the total number of tokens across the aligned sentence group, while Tree Depth was taken as the maximum depth among the constituent sentences within that specific alignment.

The results in Table 3 reveal the structural mechanisms of the simplifications. The negative values in Δ Max Tree Depth confirm that the Vikidia target texts consistently employ flatter, less complex grammatical structures. Interestingly, Δ NP Density exhibits a slight increase across all languages (e.g., $+2.18\%$ in English). This is a well-documented artifact of text distillation: as peripheral words (adverbs, complex adjectives, and subordinate clauses) are pruned to shorten the sentence, the core informative entities (Noun Phrases) occupy a higher overall percentage of the remaining token count, resulting in a denser, fact-focused syntax.

6. Conclusions

This study provides the first systematic comparison of semantic embedding spaces for sentence-level alignment in multilingual text simplification, a gap previously unaddressed in the literature. We demonstrate that LaBSE's translation-ranking objective transfers robustly to monolingual paraphrase detection across Romance languages, while BGE-M3's retrieval-optimized architecture is slightly better for English. These findings are directly relevant to NLP researchers building simplification corpora, dataset curators working with comparable sources, and developers of accessibility tools targeting low-resource languages, all of whom require principled, reproducible methods for extracting high-precision parallel data from comparable sources without costly manual annotation.

7. Limitations

While our pipeline successfully extracts high-precision parallel corpora, it has notable limitations. First, our strict thresholding strategy (τ) prioritizes precision over recall, discarding approximately 95% of the original sentences. Although this ensures a noise-free dataset, it inevitably filters out valid but highly abstract simplifications. Second, embedding performance is highly language-dependent; for instance, BGE-M3 excels in English but struggles with Romance languages. Finally, our framework currently operates exclusively at the sentence level, meaning fine-grained lexical substitutions or morphological adaptations are captured only implicitly within the aligned blocks. Extending this methodology to explicitly model token-level and sub-word alignments remains a primary focus of our ongoing work.

8. Ethics Statement

This research complies with standard ethical guidelines for NLP. We utilize publicly available, crowdsourced data from Wikipedia and Vikidia, strictly adhering to their Creative Commons (CC-BY-SA) licenses. Given the encyclopedic nature of the texts, the dataset contains no personally identifiable information (PII) or sensitive personal data. By publicly releasing this multilingual sentence-aligned corpus, we aim to support the development of accessibility tools that assist children, language learners, and individuals with cognitive disabilities, thereby contributing to the democratization of information.

9. Acknowledgments

This document is part of a project that has received funding from the European Union's Horizon Europe research and innovation program under Grant Agreement No. 101132431 (iDEM Project). The University of Leeds was funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant Agreement No. 10103529). The views and opinions expressed in this document are solely those of the author(s) and do not necessarily reflect the views of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

10. Bibliographical References

Sisay Fissaha Adafre and Maarten de Rijke. 2006. Finding similar sentences across multiple languages in Wikipedia. In *Proceedings of the Workshop on NEW TEXT Wikis and blogs and other dynamic text sources*.

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679, Online. Association for Computational Linguistics.

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, and Giulia Venturi. 2016. PaCCSS-IT: A parallel corpus of complex-simple sentences for automatic text simplification. In *Proc EMNLP*, pages 351–361, Austin, Texas. Association for Computational Linguistics.

Rémi Cardon and Natalia Grabar. 2020. French biomedical text simplification: When small and precise helps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 710–716, Barcelona, Spain (Online). International Committee on Computational Linguistics.

William Coster and David Kauchak. 2011. Simple English Wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669, Portland, Oregon, USA. Association for Computational Linguistics.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. Sonar: sentence-level multimodal and language-agnostic representations. *arXiv preprint arXiv:2308.11466*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960, Online. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proc. MT Summit X*.

Christina Niklaus, Isabel Espinosa-Zaragoza, Víctor García-Muñoz, and Paloma Moreda. 2026. A comparative study of sentence alignment methods for Spanish text simplification. *Language Resources and Evaluation*, 60(2):29.

Michael J Ryan, Tarek Naous, and Wei Xu. 2023. Revisiting non-English text simplification: A unified multilingual benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4898–4927, Toronto, Canada. Association for Computational Linguistics.

Horacio Saggion. 2017. *Automatic text simplification*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. WikiMatrix: Mining 135m parallel sentences in 1620 language pairs from Wikipedia. *arXiv preprint arXiv:1907.05791*.

Serge Sharoff, Reinhard Rapp, and Pierre Zweigenbaum. 2023. *Building and Using Comparable Corpora for Multilingual Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Springer Nature.

Sanja Štajner, Marc Franco-Salvador, Paolo Rosso, and Simone Paolo Ponzetto. 2018. CATS: A tool for customized alignment of text simplification corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Steinthor Steingrimsson, Hrafn Loftsson, and Andy Way. 2023. SentAlign: Accurate and scalable sentence alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 256–263, Singapore. Association for Computational Linguistics.

D. Varga, P. Halacsy, A. Kornai, V. Nagy, L. Nemeth, and V. Tron. 2007. Parallel corpora for medium density languages. In *Recent Advances in Natural Language Processing IV. Selected papers from RANLP-05*, pages 247–258. John Benjamins.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proc LREC*, Portorož, Slovenia.

Bi-Text Mining Across German Dialects: On the Role of Synthetic Training Data for Dialect Adaptation

Jing Wang[1] **Barbara Plank**[1,2] **Robert Litschko**[1,2]
[1]MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
[2]Munich Center for Machine Learning (MCML), Munich, Germany
robert.litschko@lmu.de

Abstract

Cross-dialect bi-text mining relies on robust multilingual sentence representations to identify semantically equivalent sentence pairs across languages. While recent multilingual bi-encoder models achieve strong performance on standardized written languages, their behavior on dialectal varieties is largely unknown. In this study, we use Tatoeba to evaluate the performance of four widely-used bi-encoders on dialect-to-standard German translation retrieval, covering German documents and queries written in three dialects: Low German, Bavarian, and Alemannic. Motivated by the lack of resources, we examine the extent to which synthetic translations (from dictionaries and large language models; LLMs) can serve as weak supervision for dialect adaptation. Our results reveal that bi-encoders, when applied in a zero-shot setting, exhibit deficiencies in capturing semantic similarity between German and dialects, while fine-tuning on synthetic data substantially improves their retrieval effectiveness, with larger gains obtained from LLM-translated training data. We further analyze retrieval performance on Bavarian across varying dialect word proportions and observe a drop when dialect words make up more than 60% of the text.

Keywords: bi-text mining, synthetic data augmentation, dialect retrieval, German dialects.

1. Introduction

Most existing multilingual embedding models are optimized using large-scale parallel or comparable corpora involving standardized written languages, such as English, Chinese and Standard German (Zhang et al., 2024; Feng et al., 2022; Chen et al., 2024; Reimers and Gurevych, 2019, *inter alia*). These training regimes implicitly assume relatively stable orthography, consistent lexical conventions, and sufficient coverage across language varieties. Dialectal data, however, are often in strong contrast to these assumptions. Unlike many low-resource languages, dialects typically do not form independent standardized systems but exist in a continuum with the standard language, sharing large portions of vocabulary while simultaneously exhibiting significant culture-specific language use, such as regionally preferred lexical variants, local idioms, pragmatic particles, and discourse conventions. For example, the Standard German imperative phrase *"Beeil dich nicht"* (*"Don't hurry"*) may be realized in Swabian as *"No ned huddla"*,[1] and the first-person pronoun *ich* may appear as *i/ig* in Swiss German. Dialect-specific spelling variations and idioms reduce surface-level similarity with standard German and increase the lexical gap between languages (Berger et al., 2000). In this work, we study the alignment of text representations between standard German and dialects.

Large-scale dialect-Standard German parallel data is scarce, and dialects are largely absent from established bi-text mining benchmarks (e.g., prior BUCC shared tasks (Zweigenbaum et al., 2018, 2017)), which focus largely on standard languages. At the same time, dialect-aware machine translation evaluation (Deutsch et al., 2025; Vamvas et al., 2025) and representation learning (Philippy et al., 2025; Artemova and Plank, 2023) are becoming increasingly active research areas. However, their application on German dialects remains underexplored: beyond the Bavarian NMT case study of Her and Kruschwitz (2024), much of the prior work focuses on translation at the lexical level and the creation of dialect dictionaries (Bui et al., 2026; Chiarcos et al., 2025; Litschko et al., 2025a). Practically, this motivates dialect-standard German bi-text mining as a scalable way to extract aligned supervision from comparable sources for both evaluation and model adaptation. Moreover, recent work has demonstrated that synthetic query–document pairs generated by large language models (Jeronymo et al., 2023) or dictionary-based translations (Alam et al., 2024) can effectively augment semantically correct training data when authentic resource is limited. While these methods have shown promise for standard languages, their applicability to dialectal varieties remains uncertain.

Motivated by these gaps and trends, we study dialect-to-standard German retrieval. To this end, we create datasets from Tatoeba, using German documents and queries in Bavarian (bar), Low German (nds), and Alemannic (als). Our work is most similar to Artemova and Plank (2023), who compared how well the cosine similarity between Ger-

[1]The verb "huddla" is derived from the noun "huddel" and culturally grounded. See Appendix C for further information on the etymology of the word.

Proceedings of the 19th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC 2026, pages 72–83
11 May 2026.

man and Alemannic/Bavarian sentence embeddings aligns with human similarity judgments on a Likert scale. We evaluate bi-encoders on sentence-level retrieval and quantify performance gains obtained from fine-tuning on synthetic data. We also ablate their performance with respect to varying proportions of dialect terms. Taken together, we investigate bi-encoders under a more realistic and challenging evaluation protocol. In this work, we address the following research questions:

- **RQ1**: How well do state-of-the-art bi-encoders perform in bi-text mining when queries are written in German dialects, compared to when they are written in English?
- **RQ2**: To what extent does training on translated data from dictionaries and LLMs improve the retrieval performance of bi-encoders?
- **RQ3**: How robust is the performance of bi-encoders with respect to different ratios of dialect code mixing?

To summarize our contributions, we (1) propose an evaluation protocol for dialect-aware translation retrieval encompassing three German dialects; (2) provide a comprehensive evaluation of multilingual bi-encoders in both zero-shot and fine-tuned settings, offering insights into their strengths and limitations when applied to dialectal data; (3) we study synthetic data augmentation strategies for dialect retrieval and conduct ablation analyses on factors affecting task difficulty, while also discussing common quality issues in the dialect subsets of web-crawled datasets and their implications for evaluation reliability. Our code and data can be found at: https://github.com/mainlp/dialect-bitext-mining.

2. Related Work

NLP Research for German Dialects. Recent work in German dialect NLP spans resource building (Burghardt et al., 2016; Litschko et al., 2025a), annotation and dialect identification (Zampieri et al., 2017; Blaschke et al., 2024; Peng et al., 2024), information retrieval (Litschko et al., 2025b), and machine translation (Her and Kruschwitz, 2024; Aepli et al., 2023). In cross-dialect retrieval, Litschko et al. (2025b) explicitly cast dialect variation as an information access problem and introduce WikiDIR, a cross-dialect retrieval test collection derived from multiple German dialect Wikipedia. Though their framing aligns with our setting, there are key differences: we focus on bi-text mining with dense retrieval rather than keyword-oriented matching. Our study extends beyond evaluation and includes training of bi-encoders under synthetic supervision. We additionally study the robustness of bi-encoders under varying proportions of dialect tokens. Despite the growing interest in German dialect NLP, publicly available parallel corpora remain still limited. Blaschke et al. (2023) provide a systematic overview of resources for German dialects including corpora that contain dialect text aligned to Standard German (or multilingual) translations. In Section 3.2, we discuss quality aspects of parallel dialect data.

Dense Retrieval with Bi-Encoders. In dense retrieval, bi-encoders are used to independently embed queries and documents into continuous vector representations, which are then ranked based on similarity measures (e.g., cosine similarity) (Karpukhin et al., 2020; Reimers and Gurevych, 2019). Bi-encoders are trained using contrastive objectives that pull the embeddings of query-document pairs closer together while pushing apart those of queries and non-relevant documents (Karpukhin et al., 2020; Oord et al., 2019). Popular benchmarks for evaluating multilingual bi-encoders, such MMTEB (Enevoldsen et al., 2025), do not focus on dialectal variation, leaving the performance of the bi-encoders on dialects largely underexplored. To address this gap, we evaluate recent models on German dialects from the Tatoeba dataset (Tiedemann, 2020). Related to our work, Vamvas et al. (2024) investigate how continual pre-training of multilingual pre-trained language models affects sentence retrieval performance in Swiss German (gsw). The authors focus on unsupervised retrieval, where sentences are matched at the token-level using BERTScore (Zhang et al., 2020). In contrast, we study multilingual bi-encoders (single-vector retrieval) and evaluate their retrieval effectiveness after fine-tuning on synthetic training data.

Synthetic Data Augmentation. Since large-scale dialect-standard parallel data is scarce, a practical way to obtain supervision for training dense retrievers is to generate synthetic training data. Large language models (LLMs) have been shown to be effective synthetic data generators in information retrieval (IR) tasks (Thakur et al., 2024; Jeronymo et al., 2023; Harsha et al., 2025), as well as in machine translation (MT) for low-resource languages (Scalvini et al., 2025; de Gibert et al., 2025). In the context of MT, Kim et al. (2025) find that synthetic data generated by GPT-4o improves MT quality for low-resource languages. We extend this analysis to dialect bi-text mining and evaluate if GPT-4o translations improve the retrieval effectiveness of bi-encoders. Next to LLM-based generation, prior work has also explored dictionaries-based data augmentation for low-recourse MT (Alam et al., 2024; Nag et al., 2020). A key advantage of lexical trans-

	Tatoeba	WikiMatrix	Wikimedia	Total
nds-de	17,984	75,591	—	93,575
bar-de	90	41,991	3,351	45,432
als-de	1,714	—	1,149	2,863
de-eng	322,413	1,573,438	180,809	2,076,660

Table 1: The amount of available parallel sentences for each language pairs. Wikimedia statistics are based on the v20230407 release and the Swiss German (gsw) subdialect of Alemannic (als).

lation is that it allows us to precisely control the proportion of dialect words (Section 5.3).

3. Evaluation Protocol

3.1. Available Dialect Datasets

In our experiments, we pair Low German (nds), Bavarian (bar) and Alemannic (als) dialect queries with German (de) documents, yielding three language pairs {nds, bar, als}–de. We include two varieties of Alemannic: Swiss German and Swabian. These dialects are among the few varieties for which sentence-level parallel data with Standard German is publicly available in sufficient quantity to support systematic evaluation (Blaschke et al., 2023). Existing datasets vary in their coverage of our selected dialect pairs and number of instances. In this study, we use Tatoeba (Tiedemann, 2020), WikiMatrix (Schwenk et al., 2021) and Wikimedia, which have been made available by the OPUS project (Tiedemann, 2012).Table 1 provides an overview of each dataset and the number of available instances. Tatoeba is a crowd-sourced collection of user-provided translations, which has over 12.6M sentences in 426 languages and is widely used in low-resource and multilingual machine translation research. WikiMatrix is a large-scale automatically mined parallel corpus extracted from aligned Wikipedia articles using LASER (Artetxe and Schwenk, 2019), which contains 135M parallel sentences for 16,720 different language pairs in total (Schwenk et al., 2021). The extracted text is split into sentences and de-duplicated. Parallel texts in Wikimedia originate from Wikipedia articles that have been translated with computer-assisted translation tools and human oversight (Laxström et al., 2015).

3.2. Dataset Quality

According to Schwenk et al. (2021), the bi-text mining method adopted by WikiMatrix may lead to a drawback of increased misalignment risk, especially for low-resource languages. In particular, previous research on the quality of web-crawled multilingual datasets (Kreutzer et al., 2022) also shows that the ratio of correct samples from WikiMatrix is at a surprisingly low level. We assessed the data quality of the nds-de and bar-de pairs by examining 1,000 samples of each language pair. Our analysis revealed that a large proportion of the pairs is misaligned: 33.2% for nds-de and 47.7% for bar-de. We also find many instances where the Standard German sentence appears on the Bavarian side (28.9%). As an additional check, we compare how well translation pairs in WikiMatrix, Wikimedia, and Tatoeba can be retrieved using BM25 (Robertson and Zaragoza, 2009). Our results on WikiMatrix and Wikimedia (Table 2) are substantially higher than those on Tatoeba (Table 3), indicating a higher risk of introducing lexical shortcuts during evaluation. Taken together, our analyses reveal substantial lexical overlaps and misaligned pairs as factors that could bias the retrieval results. We therefore exclude both WikiMatrix and Wikimedia from the evaluation test set and proceed with Tatoeba (Sections 3.3).

3.3. Evaluation Data

Dialect-to-German Evaluation. We focus on the translation retrieval involving four language pairs {nds, als, bar, en}-de. Here, en-de serves as a reference point to compare the results against a high-resource language pair. Based on our analysis in Section 3.2, we select Tatoeba as a high quality dataset. In Tatoeba, texts consists of a mix of phrases and sentences. In the following, we refer to dialect translations as queries, and their German counterparts as documents. Following Litschko et al. (2019), we use 1K different queries for each dialect and 100K German documents. The document pool is shared between all language pairs and includes all 4K "relevant documents" (i.e., translations) and 96K randomly sampled nonrelevant documents. These negatives are Standard German texts sampled from Tatoeba's German-English subset. Given that Tatoeba contains only 90 Bavarian-German sentence pairs, we supplement the dataset with an additional 910 bar-de instances. These are generated by translating the German sentences from the German-English subset of Tatoeba into Bavarian. Models are evaluated using Mean Reciprocal Rank (MRR), Recall@10, and Precision@1.

	als-de (Wikimedia)			nds-de (Wikimedia)			bar-de (WikiMatrix)		
Model	MRR@10	R@10	P@1	MRR@10	R@10	P@1	MRR@10	R@10	P@1
BM25	0.451	0.537	0.410	0.221	0.334	0.175	**0.715**	**0.764**	**0.690**

Table 2: BM25 results of MRR@10, Recall@10, Precision@1 on Wikimedia bi-text for als-de and WikiMatrix sentence pairs for {nds, bar}-de. For als-de, we use 1K Wikimedia translation pairs of Swiss German and Standard German, and augment them with 99K German documents from WikiMatrix.

Dialect-Standard Mixtures. Dialects are frequently mixed with varying degrees of standard language terms. To investigate the retrieval performance with respect to different proportions of dialectal terms, we curate a dataset of 39 German sentences, each consisting of 10 words. These sentences are sampled from the German side of Tatoeba's English-German subset. We first tokenize each sentence into a 10-word list and prompt GPT-4o to generate a list of translations Bavarian. The model is constrained to output the same number of tokens (see Appendix A). Based on the word-aligned sentence pairs, we then separately substitute 20%, 40%, 60%, 80% and 100% of the original tokens in the German sentence with the translated Bavarian variants to generate 5 subsets with different portions of dialect words.

3.4. Synthetic Training Data

Weak Supervision. Motivated by the lack of large-scale training data for dialect bi-text mining, we evaluate two methods to obtain synthetic training data: dictionary-based word-by-word substitutions; LLM-based dialect translations. The synthetic data is not assumed to be fully correct or noise-free (Kim et al., 2025). Instead, it is used to simulate realistic low-resource training conditions, where manually curated parallel data is difficult to acquire on a large demand.

Dictionary-based Translations. In this approach, we use the Bavarian dialect variation dictionary (Litschko et al., 2025a) to generate synthetic training data through word-level code-switching. The dictionary is based on human annotations and provides Bavarian spelling variations for 5,124 German lemmas. To ensure a high vocabulary coverage, we use WikiMatrix as the source of parallel query-document pairs. We perform word-by-word substitution on both query and document sides: on the query side, we generate German-like documents by replacing dialect words (where available in the lexicon) with their Standard German equivalents; on the document side, we create dialect-like queries by substituting Standard German words with their Bavarian variants. This process expands each original de-bar WikiMatrix instance into multiple de-de$_{\text{bar}}$ and bar-bar$_{\text{de}}$ pairs. We limit the number of synthetic instances to at most 30 per sentence pair. The resulting dataset contains 32,458 instances with an average length of 26.4 tokens. We use our dictionary-based training data to evaluate whether models trained on Bavarian code-switched instances generalize to other dialects (**cross-dialect transfer**).

LLM-generated Translations. In this approach, we use GPT-4o-2024-08-06 (OpenAI, 2024) to translate Standard German sentences into dialects. We randomly select Standard German sentences from Tatoeba's German-English subset. To avoid data leakage, we ensured that none of the German sentences appear in any of the test splits. We then instruct the model to generate translations for each source-target language pair using the following prompt:

Translation Prompt

Translate the following Standard German sentence into natural, fluent `{target dialect}`. Only output the translation. Try to aim for diverse translations.

The output sentence is paired with the original Standard German sentence to form a synthetic parallel query-document pair. The synthetic subsets from each dialect–Standard German pair are merged into a single mixed parallel dataset (**multi-source training**). Based on prior work (Zhou et al., 2024; Lim et al., 2024), we expect that jointly training bi-encoders on multiple dialects will improve their performance on each individual dialect. The resulting dataset contains 27K translation pairs, with an average length of 7.7k tokens.

4. Experimental Setup

Models. In our experiments, we select four state-of-the-art multilingual sentence encoders that have been pretrained on large-scale cross-lingual data: LaBSE (Feng et al., 2022), gte-multilingual-base (Zhang et al., 2024), BGE-M3 (Chen et al., 2024) and Qwen3-Embedding-0.6B (Zhang et al., 2025). For retrieval, we rank documents according to their cosine similarity to the query. Although these models demonstrate strong retrieval performance

Model	als-de			nds-de			bar-de		
	MRR@10	R@10	P@1	MRR@10	R@10	P@1	MRR@10	R@10	P@1
Lexical Retrieval Baseline									
BM25	0.219	0.328	0.173	0.129	0.210	0.096	0.416	0.535	0.360
Zero-shot Evaluation									
LaBSE	**0.526**	**0.657**	**0.461**	**0.611**	**0.782**	**0.520**	**0.676**	**0.776**	**0.625**
GTE	0.427	0.554	0.363	0.532	0.701	0.449	0.592	0.693	0.54
BGE-M3	0.421	0.540	0.358	0.564	0.731	0.479	0.638	0.734	0.590
Qwen3	0.374	0.498	0.315	0.390	0.545	0.320	0.575	0.682	0.516
Avg.	0.437	0.563	0.374	0.524	0.690	0.442	0.620	0.721	0.570
Fine-tuning on dictionary-based translations									
LaBSE	**0.569**	**0.703**	**0.502**	0.656	0.815	0.570	0.692	0.790	0.639
GTE	0.510	0.650	0.444	0.617	0.794	0.526	0.659	0.754	0.616
BGE-M3	0.561	0.682	0.493	**0.711**	**0.849**	**0.637**	**0.726**	**0.819**	**0.677**
Qwen3	0.399	0.540	0.334	0.381	0.567	0.293	0.561	0.673	0.509
Avg.	0.510	0.644	0.443	0.591	0.756	0.507	0.659	0.759	0.610
Fine-tuning on LLM-generated translations									
LaBSE	0.811	0.898	0.762	0.853	0.956	0.784	0.896	0.962	0.860
GTE	0.793	0.891	0.729	0.851	0.960	0.775	0.889	0.958	0.850
BGE-M3	**0.849**	**0.921**	**0.797**	**0.880**	**0.969**	**0.822**	**0.936**	**0.981**	**0.908**
Qwen3	0.826	0.917	0.768	0.849	0.951	0.786	0.917	0.965	0.886
Avg.	0.820	0.907	0.764	0.858	0.959	0.791	0.909	0.967	0.876

Table 3: Results of evaluating bi-encoders without any training, and when trained with synthetic supervision. Results are reported in terms of Mean Reciprocal Rank at 10 (MRR@10), Recall at 10 (R@10) and Precision at 1 (P@1). We additionally report BM25 results (baseline). Best results are highlighted in **bold**.

on standard English–German setting, their performance on dialect-aware bi-text retrieval remains unclear. We include BM25 as a lexical baseline and as a proxy to measure task-level difficulty, providing a reference point for how much a task can be solved through lexical matching. We report our results using MRR@10, Recall@10 and Precision@1.

We evaluate bi-encoders in both zero-shot and fine-tuned settings. Training is based on Sentence-BERT (Reimers and Gurevych, 2019) using InfoNCE loss (Oord et al., 2019) with in-batch negatives. We select a batch size of 64 and maximum input sequence length 128 tokens. Both zero-shot evaluation and fine-tuning process are conducted on a single A100 GPU with 80 GB. On average, each run for training took 0.48 GPU-hours per model.

5. Results and Discussion

5.1. Zero-shot Evaluation

Table 3 (upper half) reports the results of our zero-shot retrieval experiments. Comparing the zero-shot retrieval results on dialects against those obtained on en-de Table 4 reveals a large gap. On average, bi-encoder models reach an MRR@10 reaches 0.861 on en-de. On dialect–Standard German language pairs, however, retrieval performance drops to MRR@10 scores ranging from 0.437 to 0.620. The lower zero-shot retrieval performance on dialect-Standard German pairs, compared to English-German (**RQ1**), highlights that current models are much better in aligning texts written in standard and high-resource languages. However, their performance deteriorates under dialect variation.

All dense retrieval models substantially outperform BM25, demonstrating the advantage of semantic matching over lexical overlap. Among them, LaBSE achieves the highest overall performance, leading in all three language pairs: it achieves MRR@10 of 0.526 for als-de, 0.611 for nds-de, and 0.676 for bar-de. The BGE-M3 and GTE models perform competitively, while Qwen3 performs the weakest, suggesting limited capacity to handle dialectal variation despite its strong multilingual foundation.

Among the dialect pairs, all models perform best on bar-de. BM25 achieves a performance of 0.416 MRR@10, while bi-encoders achieve a MRR@10 score of 0.721 on average (+0.305 MRR@10). On

Model	en-de		
	MRR@10	R@10	P@1
BM25	0.038	0.070	0.029
LaBSE	**0.918**	**0.990**	**0.867**
GTE	0.881	0.971	0.821
BGE-M3	0.901	0.978	0.850
Qwen3	0.806	0.938	0.721
Avg.	0.861	0.960	0.797

Table 4: Zero-shot retrieval results on the en-de portion of Tatoeba.

the other hand, the performance on als-de and nds-de is notably lower, with BM25 scores of 0.219 and 0.129, respectively, and bi-encoder results of 0.437 and 0.524, both below those for bar-de. This shows that lexical overlap directly relates to retrieval difficulty (**RQ3**). However, it is important to note that test instances bar-de consists mostly of synthetic instances. In Appendix B we quantify the evaluation gap between retrieval on authentic and synthetic data, and show that models obtain stronger performance on translated data.

5.2. Fine-tuning Evaluation

Table 3 (bottom half) shows our results obtained by fine-tuning bi-encoders on synthetic data. Across the board, fine-tuning on synthetic data yields substantial gains compared to zero-shot retrieval (**RQ2**). On average, fine-tuning on dictionary-based translations improves the zero-shot performance by +0.073 MRR@10 on als-de, +0.067 on nds-de, and +0.039 on bar-de. Fine-tuning on LLM-generated translations leads to much more pronounced improvements, with bi-encoders achieving MRR@10 gains of +0.383, +0.334, and +0.289 for the three dialects, respectively. These results are encouraging and show that weak supervision in the form of Bavarian code-switched data benefits the retrieval performance also on other dialects. The larger gains obtained with models trained on LLM-generated translations can likely be attributed to the fact that these translations provide full-text semantic equivalence, closely matching the test data. In contrast, dictionary-based synthesis is primarily word-level variant substitution rather than context-aware sentence translation. More crucially, the vocabulary coverage of our dictionary is limited, so that many content words in the synthetic dialect sentence remain unchanged Standard German words.[2] As a result, model training may collapse to capturing exact token matches.

[2]Our dictionary-based translation pairs still have a token overlap of 81.6%, while LLM-generated translation pairs show only 19.3% token overlap.

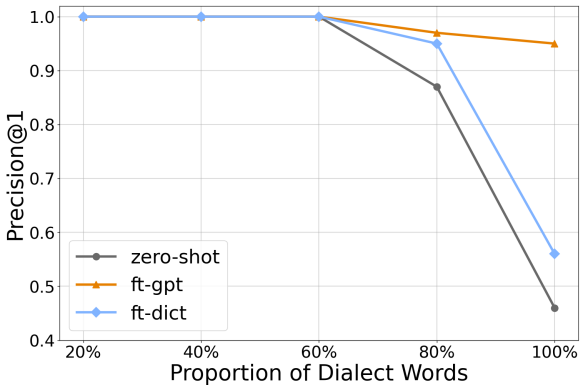


Figure 1: Results for LaBSE on bar-de with different ratios of Bavarian words. We compare zero-shot retrieval (zero-shot) to fine-tuning on LLM-translated (ft-gpt) and dictionary-translated data (ft-dict).

5.3. Robustness to Dialect Mixing

On the example of Bavarian, we now investigate how the dialect retrieval performance fluctuates with respect to different proportions of dialect tokens. In our experiments, we leave German documents unchanged and apply code switching on the query side. Figure 1 shows the retrieval results for LaBSE. We observe that retrieval performance remains consistently high when only 20%–60% of German tokens are replaced with their Bavarian translations. However, as the proportional of Bavarian words increases from 60%, performance deteriorates to varying extents. The zero-shot model exhibits the largest performance drop at 100% replacement, with P@1 reducing to 0.462. We also observe a sharp decline when LaBSE is fine-tuned on dictionary-translated data (0.564 P@1). In contrast, fine-tuning on LLM-generated data results in the most stable performance, with P@1 at 0.949. We find these trends to be consistent across bi-encoder models (see Appendix A). The results suggest that sufficient token overlap can compensate for a model's lack of dialect understanding (**RQ3**). Fine-tuning bi-encoders on LLM-generated data improves representation alignment and yields the most robust results.

6. Conclusion

This study evaluates four bi-encoder models for dialect-to-standard German retrieval. While all models outperform the lexical BM25 baseline, their retrieval performance lags behind when compared to English-to-German retrieval. We further show that fine-tuning on synthetic data consistently improves results, especially for LLM-generated translations. Our ablation on Bavarian-German reveals that the retrieval effectiveness starts to drop when the proportion of dialect words exceeds 60%.

7. Ethical considerations and limitations

Due to data scarcity, we did not evaluate model performance when trained on authentic dialect data. LLM-generated translations may introduce hallucinations or subtle meaning shifts (Vazquez et al., 2025), and dictionary-based substitutions often result in ungrammatical outputs and weak semantic equivalence (Alam et al., 2024). We quantify this difference in Appendix B.

Our focus lies on the alignment of dialect text representations with their corresponding Standard German translations. In practice, written dialects is used in social media, regional Wikipedia, and informal communication. Consequently, the scarcity of parallel data is reflective of the limited domains in which written dialects can be found. Future work should explore cross-modal alignment between dialects and Standard German in the speech domain.

Acknowledgements

We thank the anonymous reviewers for their invaluable feedback. This work is funded by the ERC Consolidator Grant DIALECT 101043235.

8. Bibliographical References

Noëmi Aepli, Chantal Amrhein, Florian Schottmann, and Rico Sennrich. 2023. A benchmark for evaluating machine translation metrics on dialects without standard orthography. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1045–1065, Singapore. Association for Computational Linguistics.

Md Mahfuz Ibn Alam, Sina Ahmadi, and Antonios Anastasopoulos. 2024. A morphologically-aware dictionary-based data augmentation technique for machine translation of under-represented languages.

Ekaterina Artemova and Barbara Plank. 2023. Low-resource bilingual dialect lexicon induction with large language models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 371–385, Tórshavn, Faroe Islands. University of Tartu Library.

Mikel Artetxe and Holger Schwenk. 2019. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

Adam L. Berger, Rich Caruana, David A. Cohn, Dayne Freitag, and Vibhu Mittal. 2000. Bridging the lexical chasm: statistical approaches to answer-finding. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Verena Blaschke, Barbara Kovačić, Siyao Peng, Hinrich Schütze, and Barbara Plank. 2024. MaiBaam: A multi-dialectal Bavarian Universal Dependency treebank. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10921–10938, Torino, Italia. ELRA and ICCL.

Verena Blaschke, Hinrich Schuetze, and Barbara Plank. 2023. A survey of corpora for Germanic low-resource languages and dialects. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 392–414, Tórshavn, Faroe Islands. University of Tartu Library.

Minh Duc Bui, Manuel Mager, Peter Herbert Kann, and Katharina von der Wense. 2026. Meenz bleibt meenz, but large language models do not speak its dialect.

Manuel Burghardt, Daniel Granvogl, and Christian Wolff. 2016. Creating a lexicon of Bavarian dialect by means of Facebook language data and crowdsourcing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2029–2033, Portorož, Slovenia. European Language Resources Association (ELRA).

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Christian Chiarcos, Janine Siewert, Tabea Gröger, and Christian Fäth. 2025. Towards a cross-dialectal dictionary for Low German (Low Saxon). In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Long and Short Papers*, pages 282–294, Hannover, Germany. HsH Applied Academics.

Ona de Gibert, Joseph Attieh, Teemu Vahtola, Mikko Aulamo, Zihao Li, Raúl Vázquez, Tiancheng Hu, and Jörg Tiedemann. 2025. Scaling low-resource MT via synthetic data generation with LLMs. In *Proceedings of the*

2025 Conference on Empirical Methods in Natural Language Processing, pages 27674–27692, Suzhou, China. Association for Computational Linguistics.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. Wmt24++: Expanding the language coverage of wmt24 to 55 languages & dialects.

Duden. 2023. hudeln. Accessed: 2026-03-05.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, et al. 2025. Mmteb: Massive multilingual text embedding benchmark. In *The Thirteenth International Conference on Learning Representations*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding.

Hermann Fischer. 1911. *Schwäbisches Wörterbuch*, volume 3. Verlag der Laupp'schen Buchhandlung, Tübingen.

Chetan Harsha, Karmvir Singh Phogat, Sridhar Dasaratha, Sai Akhil Puranam, and Shashishekar Ramakrishna. 2025. Synthetic data generation using large language models for financial question answering. In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 76–95, Abu Dhabi, UAE. Association for Computational Linguistics.

Wan-hua Her and Udo Kruschwitz. 2024. Investigating neural machine translation for low-resource languages: Using Bavarian as a case study. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 155–167, Torino, Italia. ELRA and ICCL.

Vitor Jeronymo, Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo, Jakub Zavrel, and Rodrigo Nogueira. 2023. Inpars-v2: Large language models as efficient dataset generators for information retrieval.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Seungone Kim, Juyoung Suk, Xiang Yue, Vijay Viswanathan, Seongyun Lee, Yizhong Wang, Kiril Gashteovski, Carolin Lawrence, Sean Welleck, and Graham Neubig. 2025. Evaluating language models as synthetic data generators. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6385–6403, Vienna, Austria. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Niklas Laxström, Pau Giner, and Santhosh Thottingal. 2015. Content translation: Computer-assisted translation tool for wikipedia articles. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*.

Seonghoon Lim, Taejun Yun, Jinhyeon Kim, Jihun Choi, and Taeuk Kim. 2024. Analysis of multi-source language training in cross-lingual transfer. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 712–725.

Robert Litschko, Verena Blaschke, Diana Burkhardt, Barbara Plank, and Diego Frassinelli. 2025a. Make every letter count: Building dialect variation dictionaries from monolingual corpora.

Robert Litschko, Goran Glavaš, Ivan Vulic, and Laura Dietz. 2019. Evaluating resource-lean cross-lingual embedding models in unsupervised retrieval. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1109–1112, New York, NY, USA. Association for Computing Machinery.

Robert Litschko, Oliver Kraus, Verena Blaschke, and Barbara Plank. 2025b. Cross-dialect information retrieval: Information access in low-resource and high-variance languages.

Ernst Martin and Hans Lienhart. 2012. *Wörterbuch der elsässischen Mundarten*. Walter de Gruyter.

Friedrich (ed.) Maurer, Friedrich Stroh, Rudolf Mulch, and Roland Mulch. 1973–1977. *Südhessisches Wörterbuch*, volume H–ksch of *Hessische Historische Kommission Darmstadt*. Marburg.

Josef Müller, Heinrich Dittmaier, Karl Meisen, and Matthias Zender. 1928–1971. Rheinisches wörterbuch, digitalisierte fassung im wörterbuchnetz des trier center for digital humanities.

Sreyashi Nag, Mihir Kale, Varun Lakshminarasimhan, and Swapnil Singhavi. 2020. Incorporating bilingual dictionaries for low resource semi-supervised neural machine translation.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding.

OpenAI. 2024. Gpt-4o system card.

Siyao Peng, Zihang Sun, Huangyan Shan, Marie Kolm, Verena Blaschke, Ekaterina Artemova, and Barbara Plank. 2024. Sebastian, Basti, Wastl?! recognizing named entities in Bavarian dialectal data. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14478–14493, Torino, Italia. ELRA and ICCL.

Fred Philippy, Siwen Guo, Jacques Klein, and Tegawende Bissyande. 2025. LuxEmbedder: A cross-lingual approach to enhanced Luxembourgish sentence embeddings. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11369–11379, Abu Dhabi, UAE. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*, volume 4. Now Publishers Inc.

Barbara Scalvini, Iben Nyholm Debess, Annika Simonsen, and Hafsteinn Einarsson. 2025. Rethinking low-resource MT: the surprising effectiveness of fine-tuned multilingual models in the LLM age. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 609–621, Tallinn, Estonia. University of Tartu Library.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Nandan Thakur, Jianmo Ni, Gustavo Hernandez Abrego, John Wieting, Jimmy Lin, and Daniel Cer. 2024. Leveraging LLMs for synthesizing training data across many languages in multilingual dense retrieval. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7699–7724, Mexico City, Mexico. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, volume 2012, pages 2214–2218.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Jannis Vamvas, Noëmi Aepli, and Rico Sennrich. 2024. Modular adaptation of multilingual encoders to written Swiss German dialect. In *Proceedings of the 1st Workshop on Modular and Open Multilingual NLP (MOOMIN 2024)*, pages 16–23, St Julians, Malta. Association for Computational Linguistics.

Jannis Vamvas, Ignacio Pérez Prat, Not Battesta Soliva, Sandra Baltermia-Guetg, Andrina Beeli, Simona Beeli, Madlaina Capeder, Laura Decurtins, Gian Peder Gregori, Flavia Hobi, Gabriela Holderegger, Arina Lazzarini, Viviana Lazzarini, Walter Rosselli, Bettina Vital, Anna

Rutkiewicz, and Rico Sennrich. 2025. Expanding the wmt24++ benchmark with rumantsch grischun, sursilvan, sutsilvan, surmiran, puter, and vallader.

Raul Vazquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sanchez Vega, Alessandro Raganato, Jindřich Libovický, Jussi Karlgren, Shaoxiong Ji, Jindřich Helcl, Liane Guillou, Ona De Gibert, Jaione Bengoetxea, Joseph Attieh, and Marianna Apidianaki. 2025. SemEval-2025 task 3: Mu-SHROOM, the multilingual shared-task on hallucinations and related observable overgeneration mistakes. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2472–2497, Vienna, Austria. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial evaluation campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–15, Valencia, Spain. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models.

Shijia Zhou, Huangyan Shan, Barbara Plank, and Robert Litschko. 2024. Mainlp at semeval-2024 task 1: Analyzing source language selection in cross-lingual textual relatedness. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1842–1853.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2017. Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the Tenth Workshop on Building and Using Comparable Corpora*, pages 60–67, Vancouver, Canada. Association for Computational Linguistics.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2018. Overview of the third bucc shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

A. Dialect Mixing Ablation Study

For our ablation study (Section 5.3), we used GPT-4o to translate German sentences into Bavarian. We instructed the model to translate each given sentence word-by-word into Bavarian, following a structured output format. Input sentences are provided as a tokenized list of words.

Translation Prompt

You are translating German words into Bavarian.
Task:
Translate the following 10 German source words into exactly 10 Bavarian words, in the SAME order.
Rules:
- Return exactly 10 output words.
- Each German word must map to exactly ONE Bavarian word.
- No output word may contain spaces.
- Do not add punctuation. Do not change the order.
- Every output word must be different from the corresponding German source word (case-insensitive).
- Return ONLY a JSON object in this exact format:

```
{ "translations": ["w1", "w2",
"w3", "w4", "w5", "w6", "w7",
"w8", "w9", "w10"] }
```

Source words:

```
{list of German words}.
```

Only output the translation. Try to aim for diverse translations.

Figures 2 to 4 show the results for GTE, BGE-M3, and Qwen3 evaluated on varying proportions of dialect words. Overall, the trends are consistent with those reported for LaBSE (Section 5.3). That is, models demonstrate strong retrieval results if the proportion of dialect words is 60% or less.

B. Comparison Between Synthetic and Authentic Translations

In our main experiments evaluating Bavarian-standard German retrieval, we relied heavily on LLM-generated translations, with 910 out of 1,000

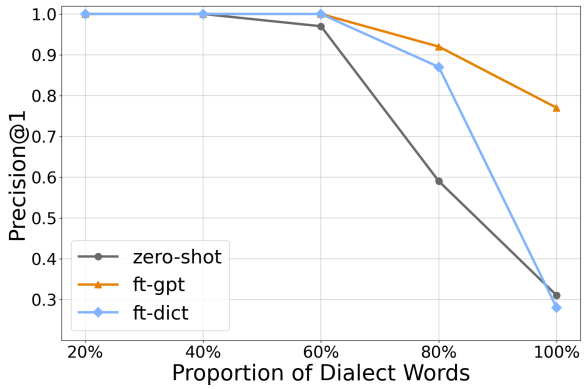


Figure 2: Results for GTE on bar-de with different ratios of Bavarian words. We compare zero-shot retrieval (zero-shot) to fine-tuning on LLM-translated (ft-gpt) and dictionary-translated data (ft-dict).

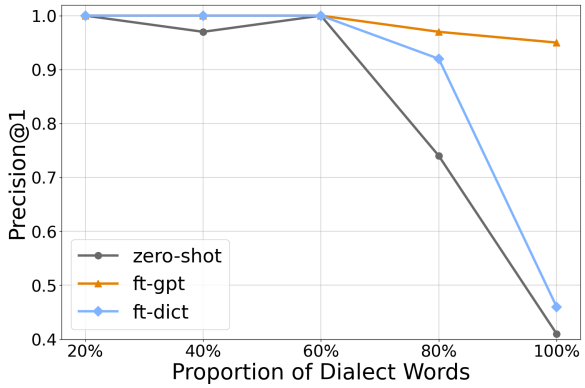


Figure 3: Results for M3 on bar-de with different ratios of Bavarian words. We compare zero-shot retrieval (zero-shot) to fine-tuning on LLM-translated (ft-gpt) and dictionary-translated data (ft-dict).

Bavarian queries being machine-translated (see Section 3.3). This was necessary due to the limited availability of human translations in Tatoeba, where we only had access to 90 authentic examples. However, this approach may introduce a bias, as bi-encoders trained on LLM-translated data may learn to recognize the characteristic "dialect style" of the GPT-4o model rather than genuine features of the Bavarian dialect. As a result, the reported performance gains may be inflated, and do not necessarily reflect the models' ability to capture authentic dialectal characteristics. To quantify this effect, we conduct an additional side-by-side comparison on the 90 authentic translation pairs, which we also translate using GPT-4o.

Table 5 shows the results of our best-performing bi-encoder (BGE-M3) evaluated on authentic translation pairs (top) and LLM-translated pairs (bottom). As expected, the model performs consistently better when evaluated on LLM-translated data. Fine-tuning and evaluating M3 on LLM-translated data

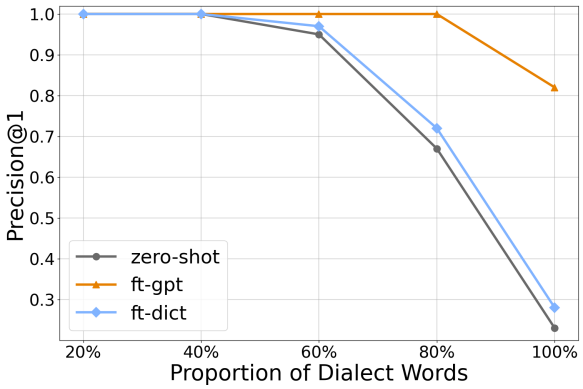


Figure 4: Results for Qwen3 on bar-de with different ratios of Bavarian words. We compare zero-shot retrieval (zero-shot) to fine-tuning on LLM-translated (ft-gpt) and dictionary-translated data (ft-dict).

Model	MRR@10	R@10	P@1
Zero-shot	0.288	0.425	0.219
Fine-tuned (LLM)	0.765	0.890	0.685
Fine-tuned (Dict)	0.406	0.548	0.329
Zero-shot	0.495	0.630	0.424
Fine-tuned (LLM)	0.867	0.987	0.781
Fine-tuned (Dict)	0.571	0.767	0.493

Table 5: Results for evaluating BGE-M3 on 90 parallel human-translated (top) and LLM-translated data (bottom).

yields the best results (0.781 P@1). Evaluating the same model on human translations reduces its retrieval effectiveness by 0.096 P@1 points. The performance gap is largest when the model is evaluated in a zero-shot fashion.

C. Explanation of "No ned huddla!"

The expression "No ned huddla!" ("nur nicht hudeln!") is taken from the Swabian Tatoeba dataset,[3] and means to not rush and be careless in a certain task (Duden, 2023). The Alemannic Wikipedia[4] traces the term "hudlâ" (dialect spelling variation) to traditional baking, where workers used damp cloths (="huddles") to quickly clean hot coals from ovens before baking bread, requiring swift action to prevent the cloth from burning. The etymological link is corroborated by documented regional dictionaries: the *Südhessisches Wörterbuch* (Maurer et al., 1973–1977, col. 760) records "hudeln" as "den angeheizten Backofen mit dem Huddel auswischen" (to wipe the heated oven with the

[3] https://tatoeba.org/en/sentences/show/6974751

[4] https://als.wikipedia.org/wiki/Wort:Schw%C3%A4bische_Vokabeln

Huddel); the *Rheinisches Wörterbuch* (Müller et al., 1928–1971, col. 885) defines “aushuddeln” as “den Backofen a., mit dem Huddel, dem Wischlumpen nach der Herausnahme der Kohlen auswischen” (to wipe out the oven with the rag after removing the coals); the *Wörterbuch der elsässischen Mundarten* (Martin and Lienhart, 2012, p. 304) lists “hudle” as “Den Backofen reinigen mit einem nassen Lumpen” (to clean the oven with a wet rag); and the *Schwäbisches Wörterbuch* (Fischer, 1911, p. 1851) lists the noun “hudel” as “Lumpen, mit dem der Bäcker den Backofen reinigt” (rags used by the baker to clean the oven).

Parallel Corpora of Scholarly Documents for English-French Machine Translation

Ziqian Peng[1,3], **Lichao Zhu**[2], **Rachel Bawden**[3], **Maud Bénard**[2], **Éric de la Clergerie**[3], **Mathilde Huguin**[4], **Natalie Kübler**[2], **Paul Lerner**[1], **Alexandra Mestivier**[2] and **François Yvon**[1]

[1] ISIR, CNRS & Sorbonne Université, Paris, France [2] ALTAE, Université Paris Cité, Paris, France
[3] Inria, Paris, France [4] INIST, CNRS, Nancy, France
contact@anr-matos.fr

Abstract

The growing ability of large language models (LLMs) to process long-range context opens new perspectives for document-level machine translation (MT), especially in scholarly communication. In fact, translating scholarly texts requires to integrate both local and long-range contextual information to ensure the consistency and coherence across the full document. However, document-level parallel corpora for such text types remain scarce, limiting both evaluation and domain adaptation of MT systems for this task. To address this gap, we introduce PARAEPS (Earth and Planetary Sciences Bilingual Corpus) and PARANLP (Natural Language Processing Bilingual Corpus), two new parallel corpora covering 14k abstracts and 105 full-length articles in two scientific domains to be used for fine-tuning and evaluation purposes. We compare the performance of eight MT systems on these test sets and find that fine-tuning on document-level data closes the gap between open systems based on Large Language Models (LLMs) and commercial systems. We also find that the performance of recent LLMs can worsen when translating full articles instead of translating them on a per paragraph basisfine-tuning. These experiments underscore the need for corpora such as PARAEPS and PARANLP.

Keywords: Machine Translation, Parallel Corpus, Scientific Documents, Long-context Modelling, Large language models

1. Introduction

The development of large language models (LLMs) creates new possibilities for document-level machine translation (MT), owing to their ability to process long-range dependencies (Karpinska and Iyyer, 2023; Peng et al., 2024a; Wang et al., 2024, 2025; Zhu et al., 2025). However, document-level parallel corpora remain scarce,[1] particularly for scholarly documents. For such texts, existing resources are mostly limited to sentence-aligned parallel texts (Roussis et al., 2022; Esalati et al., 2024) without document boundary information or restricted to the medical domain (Ive et al., 2016; Névéol et al., 2018). Although some resources preserve the document structure (Abdul Rauf and Yvon, 2024), they generally comprise only abstracts (Kleidermacher and Zou, 2025). Consequently, the training of MT systems for scientific texts mainly relies on short texts, thus failing to represent the actual complexity of scholarly articles, which often contain complex formulas, citations and long-range contextual dependencies. Moreover, given the lack of document-level test sets, the evaluation of recent LLMs to translate scholarly articles is often restricted to reference-free metrics (Zhu et al., 2025) or human judgments (Kleidermacher and Zou, 2025).

In this work, we present the curation of additional resources for document-level translation in two scientific fields: Earth and Planetary Sciences and PARANLP for Natural Language Processing. Our corpora, PARAEPS and PARANLP, consist of both abstracts and full-length articles; each contains training, validation and test sets of parallel abstracts, and a test set of complete parallel articles. These parallel articles are constructed using one of the three different approaches: 1) human translations opportunistically collected by the authors, 2) human post-edits of machine translated texts, 3) combination of human translations, automatic post-editions and machine translations derived from comparable articles published in both English and French.

Using our test sets of full articles, we compare the performance of six MT systems translating on a per paragraph basis, and of two LLMs translating chunks of varying sizes. Experimental results show that 1) fine-tuning on paragraph-level datasets closes the gap between the performance of medium-size LLMs and commercial MT system such as DeepLPro,[2] for the translation of abstracts and 2) there is no systematic performance gain when using recent LLMs to translate scientific texts at the article level as opposed to applying them at

[1] A situation that is changing, at least for Web pages (O'Brien et al., 2025).

[2] https://deepl.com

Proceedings of the 19th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC 2026, pages 84–95
11 May 2026.

the paragraph or sentence level. These results illustrate the utility of our corpora and the need to improve recent models for MT tasks addressing the increasing needs of scholarly communication across languages.[3] Our main contributions are as follows:

- the collection and construction of parallel scholarly documents in two domains, including 14k abstracts and 105 full-length articles.
- an original pipeline that constructs parallel articles from comparable articles extracted from scholarly publications in the NLP domain; as these parallel texts are partly machine-generated, partly human generated, we dub this corpus a silver reference corpus;
- benchmark results of two commercial systems and recent LLMs, fine-tuned or not on these newly created corpora.

We openly release our corpora and the code of the corpus construction pipeline under a permissive license.[4]

2. Related Work

Parallel corpora are central for training and evaluating MT systems, but most existing resources for scholarly documents are only aligned at the sentence level (Roussis et al., 2022; Esalati et al., 2024; Roussis et al., 2024) disregarding document boundary information, or restricted to specific domains (e.g., for the medical sciences (Ive et al., 2016; Névéol et al., 2018; Abdul Rauf and Yvon, 2024), or both (for instance the Taus Corona Crisis corpus[5] and the Mlia Covid corpus[6])). Even when document structure is preserved, documents mostly correspond to abstracts (Kleidermacher and Zou, 2025), failing to represent (a) very long-range dependencies (e.g., between the Introduction and Conclusion sections) and (b) translation issues that are specific to scholarly texts such as the translation of captions, table cells, or the insertion of citations in the discursive flow.

The lack of availability of full-length parallel documents also means that the evaluation of MT systems in their ability to translate scholarly content mainly relies on reference-free metrics computed at the paragraph level (Zhu et al., 2025), human evaluation (Kleidermacher and Zou, 2025) or the analysis of translated abstracts (Sebo and de Lucia, 2024).

Recently, several parallel corpora comprising long documents have been introduced in literary domains (Jiang et al., 2022; Wang et al., 2024a,b). In addition, O'Brien et al. (2025) recently introduced DocHPLT, a massive collection of parallel documents extracted from the Internet Archive for multiple language pairs. However, they do not focus on scholarly texts, nor do they provide sufficiently informative domain tags. Another recent resource is ACADATA (Lacunza et al., 2025), a collection of parallel abstracts across 12 languages harvested from public academic web sites and archives. These documents are however relatively short (about 1000 characters), and lack gold domain tags. Finally, the ACL 60-60 initiative aimed to produce reference translations for NLP abstracts in multiple languages, as well as a large number of automatically generated translations of papers and talks.[7] One outcome of this effort was the release of development and test data for the IWSLT 2023 shared task (Salesky et al., 2023): each set contains a post-edited version of the translations (in 11 languages) of 10 presentations delivered during the ACL 2022 conference[8].

Our work complements these developments by introducing two document-level parallel corpora of scholarly documents, including full-length parallel-articles, aimed to mitigate the scarcity of resources required to study discourse-aware MT.

3. Dataset Creation

We create new parallel resources for English–French translation in two fields of study: Earth and Planetary Sciences (EPS) and natural language processing (NLP). Each of the two subsets, PARAEPS and PARANLP, comprises four data splits: train, dev and test sets of *abstracts* (TRAIN, DEV and TEST) and a second test set (TEST-LONG) of *full articles*. We collect texts from multiple sources and use various techniques to construct the datasets, including manual translation, post-editing and, in the case of the NLP domain, some partial automatic translation to complement manually translated examples. We describe the sources and dataset creation process for PARAEPS and PARANLP in Sections 3.1 and 3.2 respectively; statistics for both datasets can be found in Table 1.

3.1. EPS Dataset (PARAEPS)

For the EPS domain, we collected 11k abstracts and 29 articles in both English and French. This section presents data collection, text processing, alignment and the construction of the data splits.

[3] https://www.helsinki-initiative.org/.
[4] https://anr-matos.github.io/pages/resources.html
[5] https://md.taus.net/corona
[6] http://eval.covid19-mlia.eu/task3/
[7] https://acl6060.org/
[8] https://2022.aclweb.org/

Split	#docs	# sents	#toks/doc ($\mu \pm \sigma$) en	fr
PARAEPS				
TRAIN	10,577	83,036	347 ± 180	474 ± 233
DEV	400	3,273	344 ± 144	483 ± 192
TEST	391	3651	401 ± 203	544 ± 271
BSGF	*132*	*1311*	*472 ± 196*	*622 ± 263*
CRAS	*100*	*677*	*277 ± 118*	*388 ± 168*
CRG	*59*	*364*	*260 ± 101*	*360 ± 135*
THESES$_{EPS}$	*100*	*1299*	*512 ± 211*	*707 ± 281*
TEST-LONG	29	5,133	7,773 ± 2,755	10,539 ± 3,613
MERSENNE	*19*	*3,532*	*7,673 ± 2,828*	*10,652 ± 3,814*
STUDENT	*10*	*1,601*	*7,962 ± 2,600*	*10,322 ± 3,186*
PARANLP				
TRAIN	2,723	24,085	287 ± 159	429 ± 234
DEV	96	1,024	353 ± 135	523 ± 210
TEST	346	2,022	176 ± 124	264 ± 181
rTAL	*246*	*1,015*	*121 ± 46*	*184 ± 68*
THESES$_{NLP}$	*100*	*1,007*	*310 ± 150*	*463 ± 216*
TEST-LONG	76	14,467	6,064 ± 2,783	8,388 ± 3,820
NLPG$_{OLD}$	4	533	*4,477 ± 2,999*	*6,679 ± 4,436*
NLPS$_{ILVER_{EN-FR}}$	36	7,025	*6,028 ± 2,399*	*8,433 ± 3,331*
NLPS$_{ILVER_{FR-EN}}$	36	6,909	*6,275 ± 3,045*	*8,532 ± 4,145*

Table 1: Statistics for PARAEPS and PARANLP and their data splits. TRAIN, DEV and TEST splits are composed of abstracts. TEST-LONG is composed of full articles. English (en) and French (fr) token counts are based on TOWERBASE tokens.

3.1.1. Data Collection and Processing

Abstracts We collected over 11k parallel scientific abstracts (89k sentences, and 2.2M and 2.6M words in English and French respectively) from seven sources (listed in Table 2 with the statistics after quality filtering), by extracting the plain texts from the HTML pages, and aligning them across the two languages. We use `langdetect`[9] to filter out noisy abstracts written in other languages, and we also disregard abstracts of source-to-target length ratio is smaller than 0.5.

We applied NFC normalization using unicodedata,[10] before segmenting abstracts in sentences using Trankit (Nguyen et al., 2021), which reliably disambiguates the multiple interpretations of the dot ('.') symbol in scholarly documents (e.g. in numbers or abbreviations, in addition to the sentence-final punctuation mark). We aligned the resulting bilingual segments using a slightly modified version of BertAlign (Liu and Zhu, 2022), which robustly supports many-to-many sentence alignments.[11] We use the value 0.001 for the `skip` parameter, which improves zero-to-one alignments. We also introduce a new parameter `len_slack` (with value 0.15), which prevents to apply a length penalty for parallel segments having length ratio close to 1; this tends to reduce the number of spurious many-to-any alignments.

Then we filtered the aligned sentences using quality estimation scores from TransQuest (Ranasinghe et al., 2020). Additional details concerning the filtering are provided in Section 3.1.2, as we also use the alignment scores when selecting data for the different data splits.

Full Articles We collected parallel articles from two sources. Firstly, ten English articles and their translations were obtained from a specialised translation course. The original articles were either sourced from the ISTEX database[12] (Maurel et al., 2019) or were Open Access. The translations, which were produced by master's students, were the result of either translation from scratch or post-edition of MT (both standard approaches used by translation specialists) followed by proofreading. We refer to these texts as the STUDENT collection.

Secondly, we also collected 19 articles (five in English and fourteen in French) published in the "Compte rendus Géosciences" journal,[13] which were then automatically translated and post-edited by a professional translator using the MateCat platform (Federico et al., 2014). We refer to these texts as the MERSENNE collection. Unlike the STUDENT collection, we had to extract parallel texts from MERSENNE. We used pandoc[14] to extract the plain text from the HTML documents, removing empty lines, the symbol `\xa0` and carrying out NFC normalization. We extract the blocks of text (hereafter referred to as paragraphs),[15] equations and tables.[16] Since the English and French versions contain the same number of paragraphs, we trivially align them, before segmenting into sentences and aligning the sentences using the same pipeline as for the abstracts. Non-aligned sentences were manually realigned. We validated the resulting alignment using TransQuest (Ranasinghe et al., 2020): all paired sentences had an alignment score of at least 0.75, indicated good alignment throughout.

[9] https://pypi.org/project/langdetect/

[10] https://docs.python.org/3/library/unicodedata.html

[11] https://github.com/ANR-MaTOS/bertalign

[12] https://www.istex.fr/

[13] https://comptes-rendus.academie-sciences.fr/geoscience

[14] https://pandoc.org/

[15] In practice, these also correspond to section titles and captions in addition to paragraphs.

[16] We store equations and tables as complementary information to be used in future work.

[17] We collected abstracts from the following scientific journals: *Hydrogeology Journal*, *Mineralogy and Petrology*, *Swiss Journal of Geosciences*, *Geodinamica Acta*, *Journal of South American Earth Sciences*, etc. in ISTEX, which is a data portal of multilingual scientific data.

	Source of abstracts	#segments	#abstracts (all)	#abstracts		
				TRAIN	DEV	TEST
PARAEPS	BSGF (Bulletins de la Société Géologique de France)	1,311	132	-	-	132
	CanMin (Canadian Mineralogist)	8,140	793	793	-	-
	CJES (Canadian Journal of Earth Sciences)	37,525	4,624	4,524	100	-
	CRAS (Comptes Rendus de l'Académie des Sciences - Earth and Planetary Sciences, 1995-2001)	9,620	2,026	1,826	100	100
	CRG (Comptes rendus Géoscience)	364	59	-	-	59
	ISTEX (Infrastructure de services pour la fouille de textes)[17]	15,190	2,117	2,017	100	-
	THESES (Database of PhD abstracts)	17,810	1,617	1,417	100	100
	TOTAL	89,960	11,368	10,577	400	391
PARANLP	ISTEX (Infrastructure de services pour la fouille de textes)	8,099	1,309	1,309	-	-
	rTAL (revue TAL)	1,015	246	-	-	246
	THESES (Database of PhD abstracts)	18,987	1,610	1,414	96	100
	TOTAL	30,543	3,165	2,723	96	346

Table 2: Data sources and statistics for the abstracts in the PARAEPS and PARANLP TRAIN, DEV and TEST splits, after quality filtering.

3.1.2. Dataset splits

PARAEPS consists of TEST-LONG, containing the 29 parallel articles, and TRAIN, DEV and TEST splits composed of parallel abstracts.

EPS-TEST-LONG is composed of the full articles from the MERSENNE and STUDENT collections aligned at the sentence level. For MERSENNE articles, we also preserve paragraph-level boundary information, comprising 915 paragraphs.

EPS-TEST is composed of abstracts from four of the collections listed in Table 2: BSGF, CRAS, CRG and THESES. To ensure the quality of the test set, we computed the average alignment score for sentence pairs within each abstract, and empirically excluded abstracts with an alignment score below 0.5. We then kept the remaining abstracts from BSGF and CRG due to their small size, and selected the 100 most recent abstracts from CRAS and THESES.

EPS-DEV contains a total of 400 abstracts. Only considering parallel abstracts with an alignment score above 0.5, we randomly sample 100 abstracts from the 200 most recent abstracts from each of the CJES, CRAS, ISTEX and THESES collections.

EPS-TRAIN contains the remaining 10,577 parallel abstracts after filtering the most unreliable alignments (i.e. those whose average alignment score is below 0.4 when all sentences are aligned, or below 0.5 if at least one sentence is unmatched).

The right side of Table 2 displays the distribution of abstracts in EPS-TEST, EPS-DEV, EPS-TRAIN, broken down by data source.

3.2. NLP Dataset (PARANLP)

For the NLP domain, we collected 3k parallel abstracts and 76 parallel articles. The complete articles are derived from four human translated articles and 36 comparable human-written articles, each turned into two parallel texts as described below. The corresponding statistics are reported in the bottom part of Table 1.

3.2.1. Data Collection and Processing

Abstracts We collected abstracts from NLP publications for which both English and French versions are available. These raw texts include 246 parallel abstracts extracted from the French NLP journal *revue TAL* (rTAL),[18] 1358 NLP abstract retrieved from various journal articles available in the ISTEX archive, and 1701 abstracts from PhD dissertations (THESES).[19] We processed the abstracts using the same pipeline as for PARAEPS, i.e. first segmenting them with Trankit, then performing sentence alignment using BertAlign. Furthermore, we filtered the resulting alignments using `TransQuest` scores.[20] These abstracts are then split into three parallel corpora NLP-TRAIN, NLP-DEV and NLP-TEST, as detailled in Section 3.2.2. NLP-TEST was aligned using hunalign[21] (Varga et al., 2005) in the early stage of our data preparation process. To ensure its quality, we manually reviewed all the sentence pairs having a TransQuest score lower than 0.3. The statistics of PARANLP after quality filtering are in the bottom of Table 2.

[18] https://www.atala.org/revuetal

[19] https://theses.fr/

[20] We use the same filtering heuristics as for PARAEPS.

[21] https://github.com/danielvarga/hunalign

	NLPSILVER$_{FR-EN}$				NLPSILVER$_{EN-FR}$						
	total	mean	min	max	total	mean	min	max	correct	total	TER
Copy	2205	61	17	132	1677	46	13	104	41	126	24.6
APE	3478	96	38	246	4031	111	43	256	51	199	27.5
MT	1226	34	3	176	1317	36	5	184	12	42	21.5
all	6909	191	103	408	7025	195	112	477	109	367	26.2

Table 3: An analysis of the composition of our silver NLP corpus in number of sentences, with translations being produced through (i) copying the original aligned human-written translations (Copy), (ii) postediting them (APE) or (iii) machine-translating from scratch (MT). The three columns on the right correspond to a quality assessment, based on 3 articles that were manually post-edited by one author. For these, we report the number of correct translations out of all sentence pairs (total), and the TER score based on a comparison between the silver and the post-edited versions.

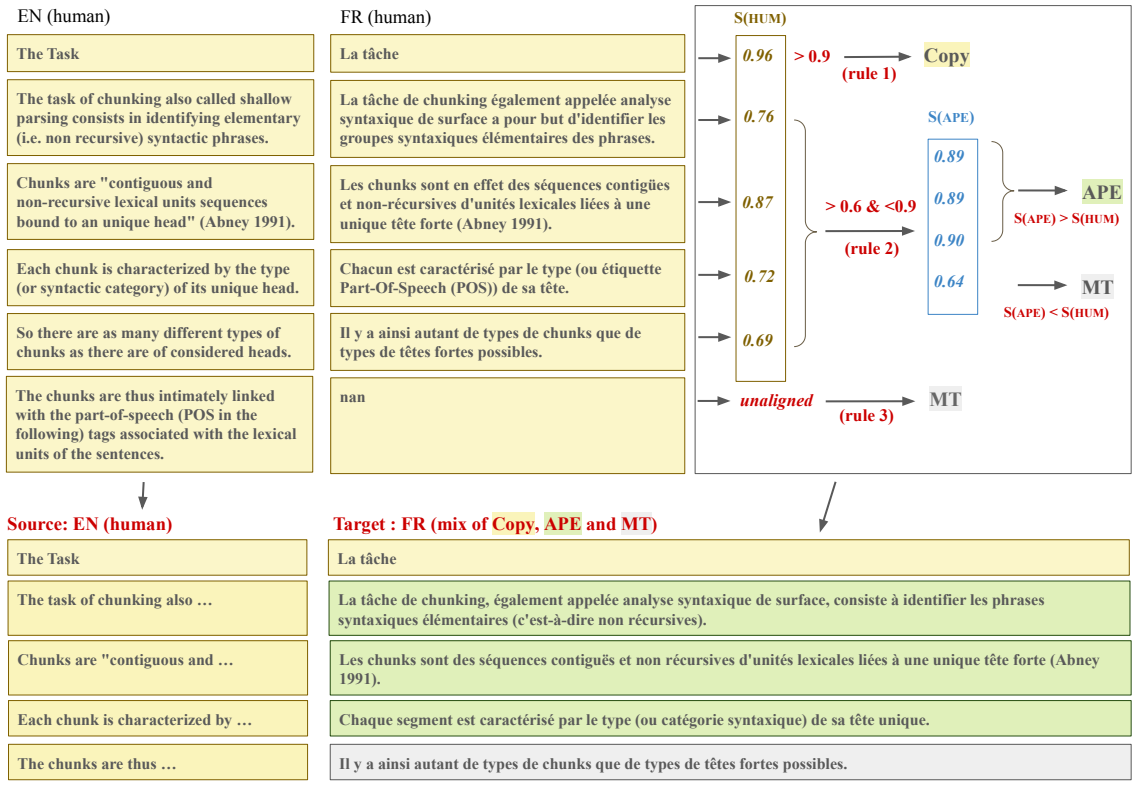


Figure 1: The pipeline to construct parallel articles from comparable articles, with examples from NLPSILVER$_{EN-FR}$. We denote S(HUM) and S(APE) the cosine similarity between an English (EN) sentence and its human-written translation (HUM) or automatic post-edition (APE), based on LaBSE embeddings.

Full articles We collect full parallel articles from two sources: (i) regular human translations (NLPGOLD), consisting of two English articles translated into French and two French articles translated into English, collected opportunistically by the authors, which we complement with (ii) a larger set of silver translations (NLPSILVER), which we assembled using a combination of automatic and manual operations, as described below.

In addition to full human translations (NLPGOLD), we also construct NLPSILVER, which is derived from articles from the ACL Anthology.[22] While the ACL Anthology mostly stores English articles, it also contains a small number of French articles, published in French journals and conferences. A fraction of these also have an English counterpart, as some French speaking authors wish to publish their research in both languages, possibly using MT and post-editing to speed up the process. Given that there is no guarantee that the English and French versions are strictly parallel, we apply a more complex alignment process than for the previously described datasets. We first extracted the text version of all English and French papers, segmented them into sentences,

[22] https://aclanthology.org/

filtering short segments or segments containing mostly non-alphabetic characters (likely equations or OCR errors, for older articles) and embedded them in a joint multilingual space using LaBSE (Feng et al., 2022).[23] We then indexed the embeddings using FAISS (Johnson et al., 2021), and, for each French sentence, searched for its closest English neighbours. We manually inspected the French articles with a large number of sentences whose neighbours were found in the same English documents, resulting in a final list of 36 pairs of articles.

We convert these articles from pdf to markdown using `pymupdf4llm`[24], remove noisy texts (page numbers, pdf headers, algorithms, etc.) using regular expression and manual verification, extract and keep aside tables, resulting in clean markdown files, which are finally converted into plain texts for sentence segmentation.

For each of pair of articles, we first performed sentence alignment using the same pipeline as described previously (using Trankit and Bertalign). We then derived a fully parallel *English-French* version, denoted as NLPSILVER$_{\text{EN-FR}}$, as follows. We process each English article, and for each source segment e and its aligned French counterpart f and apply the following rules, which are also illustrated in Figure 1:[25]

1. if the alignment score between e and f is above 0.9, we keep the pair (e, f), assuming that they are mutual translations (Copy).
2. if the alignment score between e and f is between 0.6 and 0.9, we use automatic post-editing (APE) of (e, f) with TowerInstruct-13B-v0.1[26] to generate f'. If f' has a better alignment score with e than f, we keep f'; otherwise we retranslate e from scratch (MT).
3. if the alignment score between e and f is below 0.6, or if e is not aligned with any French counterpart, we retranslate e from scratch as for case 2 (MT).

We used the same heuristics to create a fully parallel *French-English* version, which we refer to as NLPSILVER$_{\text{FR-EN}}$, comprising all the human-written French texts and their corresponding translations derived from the pipeline of Figure 1. We provide the corresponding number of segments produced by each rule for each version in Table 3.

In order to evaluate the confidence of each type of rule, a native French speaker with expertise in NLP post-edited a random sample of three articles from the English-French version.[27] TER scores (Snover et al., 2006) were also computed using SacreBLEU (Post, 2018). The results of this analysis are in the right part of Table 3. Sentences produced through the *Copy* action (i.e. the translation was taken from the original article) are slightly more likely to be fully correct than segments generated with the other rules, but the edit distance to acceptable references is still non-negligible.

3.2.2. Dataset Splits

As for PARAEPS, PARANLP consists of TEST-LONG, containing full parallel articles, and TRAIN, DEV and TEST splits composed of parallel abstracts.

NLP-TEST-LONG is composed of the parallel articles just described (NLPGOLD, NLPSILVER$_{\text{EN-FR}}$ and NLPSILVER$_{\text{FR-EN}}$).

NLP-TEST contains 346 parallel abstracts, corresponding to all RTAL abstracts and 100 randomly selected abstracts from THESES$_{\text{NLP}}$.

TRAIN and DEV NLP-DEV is composed of 96 abstracts randomly sampled from THESES$_{\text{NLP}}$ (non-overlapping with those selected for TEST). NLP-TRAIN contains the remaining abstracts extracted from ISTEX and THESES$_{\text{NLP}}$.

4. Experiment Settings

We provide benchmarking experiments to illustrate the usefulness of the two datasets for both fine-tuning and evaluation of document-level MT for scholarly documents.

MT Engines We test the translation performance of two multilingual LLMs: TowerBase-7B[28] (TOWER) (Alves et al., 2024) and EuroLLM-9B[29] (EUROLLM) (Martins et al., 2025), when translating abstract test sets at the paragraph level, before and after fine-tuning on the corresponding training set (EPS-TRAIN and NLP-TRAIN). For comparison, we also evaluate the translation quality of two com-

[23] https://huggingface.co/sentence-transformers/LaBSE.

[24] https://github.com/pymupdf/pymupdf4llm

[25] e and f each correspond to one or more consecutive sentences resulting from BertAlign's many-to-many alignment.

[26] https://huggingface.co/Unbabel/TowerInstruct-13B-v0.1

[27] Using the MateCat platform, without knowledge of each segment origin.

[28] https://huggingface.co/Unbabel/TowerBase-7B-v0.1

[29] https://huggingface.co/utter-project/EuroLLM-9B

mercial systems: DeepLPro (DeepL)[30] and SystranPro (Systran).[31][32]

For the translation of complete articles, we compare the performance of Llama3.1-8B-Instruct[33] (Llama3) (Grattafiori et al., 2024), which has a context length of 128k tokens, and Qwen3-8B[34] (Qwen3) (Yang et al., 2025) with a context length greater than 32k on the MERSENNE subset of EPS-test-long.

Fine-tuning and Inference We perform supervised fine-tuning using QLoRA (Dettmers et al., 2023), following the quantization and LoRA configurations proposed by Moslem et al. (2023, Section 2.3) for Tower and EuroLLM. The QLoRA learning rate is $2e{-}4$ adjusted by a cosine schedule, with neither warm-up steps nor packing. The batch size is set to 8.

For NLP we fine-tune Tower and EuroLLM for two epochs on NLP-train, with two gradient accumulation steps to produce FT-Tower-NLP and FT-Euro-NLP respectively. Similarly, we fine-tune both models on EPS-train to produce FT-Tower-EPS and FT-Euro-EPS, although due to the fact that the training set is larger, we do so for one epoch only and set the gradient accumulation steps as 4.

Inference is performed without additional in-context examples, with bfloat16 and greedy search, using the Huggingface implementation, except for the inference of Llama3 and Qwen, which is carried out with vLLM (Kwon et al., 2023). For Qwen3, we use the suggested configuration for the non-thinking mode using a hybrid decoding method that combines top-k and top-p, with temperature, top-p, top-k values of 0.7, 0.8, and 20 respectively.

Prompts We prompt base models following their HuggingFace model cards, using the following two prompts for Tower and EuroLLM respectively:

(1) `English: SRC\nFrench:`

(2) `English: SRC French:`

We use the following prompt for fine-tuning and fine-tuned models:

(3) `Translate the following text from English into French.\nEnglish: SRC\nFrench: TGT`

For the translation of full articles with Qwen3, we use the following prompt:

(4) `Translate the following text from English into French.\nEnglish: SRC\nFrench:`

For full article translation with Llama3, we use the following system prompt:

(5) `You are a good translator! Translate the following text from English into French. Reply only with the translated text.`

and the following user prompt template:

`English: SRC\nFrench:`

Metrics To evaluate translation quality, we use standard BLEU (Papineni et al., 2002) and its document-level variant, denoted ds-BLEU (Peng et al., 2024b).[35,36] We also report the document-level COMET (d-COMET) score (Vernikos et al., 2022) with `wmt22-comet-da` (Rei et al., 2022). While ds-BLEU can be applied as-is to documents without requiring sentence alignment, to evaluate BLEU and d-COMET, we first have to realign abstracts, paragraphs, and articles at the sentence level.[37]

5. Results and Analysis

5.1. Paragraph-level MT

Tables 4 reports the translation quality of the six MT systems in translating abstracts from test sets in EPS-test and in NLP-test.

For EPS, DeepLPro and FT-Euro-EPS are ranked as the top two MT systems. DeepLPro achieves the best d-COMET scores for all test sets, while FT-Euro-EPS results in higher BLEU scores for three out of four subsets of EPS-test. We obtain performance gain through fine-tuning both LLMs on EPS-train, except for CRG when using the fine-tuned EuroLLM.

[30] https://deepl.com

[31] https://www.systransoft.com/

[32] DeepL and Systran were accessed in March 2026 for abstracts and October 2025 for MERSENNE articles.

[33] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[34] https://huggingface.co/Qwen/Qwen3-8B

[35] For ds-BLEU, each document is considered as a single segment (all sentences concatenated), 'sentence-level' BLEU is applied to each document and the average score is calculated over documents.

[36] We use SacreBLEU (Post, 2018) with the signature `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.4.0` to compute BLEU. For ds-BLEU, effective order is activated with `eff:yes`

[37] We do this by first aligning at the character level between the automatic translation and its reference using `edlib` (https://pypi.org/project/edlib/) then segmenting sentences with respect to the sentence boundaries of the reference.

	BSGF			CRAS			CRG			$\text{THESES}_{\text{EPS}}$		
	BLEU	ds-BLEU	d-COMET	BLEU	ds-BLEU	d-COMET	BLEU	ds-BLEU	d-COMET	BLEU	ds-BLEU	d-COMET
Systran	**42.8**	**44.3**	78.5	33.3	33.0	79.5	48.1	48.2	82.6	47.3	48.4	82.4
DeepL	41.2	43.2	**80.1**	33.5	33.2	**81.1**	47.5	47.9	**83.6**	45.8	47.2	**83.0**
Tower	36.1	38.2	75.6	33.2	33.1	78.6	48.5	48.3	81.7	43.9	45.5	80.6
FT-Tower-EPS	38.7	40.0	77.9	34.9	34.4	79.5	49.1	48.9	82.0	45.0	46.7	81.9
EuroLLM	40.7	42.1	78.6	35.8	35.5	79.9	**51.5**	**51.5**	82.8	47.5	49.2	82.7
FT-Euro-EPS	41.1	42.5	78.5	**36.6**	**35.8**	80.1	50.9	50.2	82.5	**48.0**	**49.5**	82.6

	rTAL			$\text{THESES}_{\text{NLP}}$		
	BLEU	ds-BLEU	d-COMET	BLEU	ds-BLEU	d-COMET
Systran	34.5	33.4	76.1	**43.0**	**43.1**	78.9
DeepL	34.2	33.6	**78.0**	41.7	42.4	**80.4**
Tower	32.1	31.1	75.5	40.0	40.1	77.7
FT-Tower-NLP	32.8	32.1	76.4	41.5	41.6	78.6
EuroLLM	34.6	33.5	76.4	**43.0**	42.7	79.2
FT-Euro-NLP	**35.0**	**33.8**	76.9	42.8	**43.1**	79.3

Table 4: BLEU, ds-BLEU and d-COMET scores for each subset of EPS-TEST (top) and NLP-TEST (bottom), corresponding to abstract translation. Best and second-best scores are bold and underlined, respectively

For NLP, DEEPLPRO and FT-EURO-NLP achieve the best and second-best performance for all metrics. As for EPS, we observe that fine-tuning is beneficial for both TOWER and EUROLLM.

5.2. MT of full articles

To investigate the capacity of LLMs with large context lengths to translate full scientific articles, we evaluate LLAMA3 and QWEN3 on the full-length articles from the MERSENNE subset of EPS-TEST-LONG. We also calculate the scores at the paragraph level in order to also compare the LLMs to the MT models from the previous experiments, whic have limited context window sizes.

Scores for paragraph-level translation are given in Table 5: EUROLLM and FT-EURO-EPS perform the best.

To study the effect of increasing the size of translation segments, we compare the translation quality (BLEU scores) of LLAMA3 and QWEN3 when translating MERSENNE at the sentence, paragraph, and article level. The results reported in Table 6 show that the translation quality of LLAMA3 slightly improves when translating paragraphs instead of sentences, although it degrades when translating full-length articles, despite the lengths of input articles fitting within the context window size. This is consistent with the findings of Wang et al. (2024) and Peng et al. (2025). The value of the brevity penalty observed suggests that under-translation is one of the reasons apart from the quality degradation. In contrast, the BLEU scores of QWEN3 increase when translating full articles with respect to paragraph-level translation, suggesting its robustness in long-context MT.

MT system	BLEU	ds-BLEU	d-COMET
Systran	57.3	58.0	87.3
DeepL	56.2	58.6	**88.0**
Tower	56.2	55.5	85.8
FT-Tower-EPS	57.1	58.4	87.0
EuroLLM	**61.7**	**62.7**	87.6
FT-Euro-EPS	59.7	60.7	87.6

Table 5: Scores for MT systems translating the MERSENNE subset of EPS-TEST-LONG (full articles) at the paragraph level. Best and second-best scores are bold and underlined, respectively.

Model	sent2sent	par2par	doc2doc
Llama3	51.2 (1.00)	51.8 (1.00)	49.2 (0.96)
Qwen3	48.4 (1.00)	48.3 (1.00)	50.8 (0.99)

Table 6: BLEU score (and brevity penalty) for LLMs translating MERSENNE articles at the sentence (sent2sent), paragraph (par2par), and full document (doc2doc) levels.

6. Conclusion

To address the scarcity of parallel documents in scientific fields for document-level MT, we constructed two English–French parallel corpora, PARAEPS and PARANLP, consisting of parallel abstracts and full-length parallel articles in Earth and Planetary Sciences and in Natural Language Processing respectively. Each corpus comprises TRAIN, DEV and TEST sets of abstracts and a second test of full articles (TEST-LONG). While most of the translations are produced through manual effort, the NLP-TEST-LONG is partly made up of silver translations constructed from comparable versions of the same article in English and French.

We present the original pipeline by which we derive silver parallel articles from those comparable human-written articles. Our experiments demonstrate the usefulness of our corpora, as fine-tuning LLMs on our dataset improves translation quality, and the parallel articles provide resources for the evaluation of article-level MT. Our future work will explore more precisely how document-level phenomena are handled by machine translation systems. This includes discourse phenomena that require extra-sentential context, but also terminological variation (in terms of consistency and logical use of term variants such as acronyms and reduced forms). Another short term goal will be to develop parallel scholarly corpora for a more diverse set of scientific domains, including other scientific domains that are characterised by domain-specific notation and formulae.

7. Ethical Considerations

BSGF, CRAS and CRG articles are released under a permissive CC BY 4.0 license. PARANLP abstracts are considered as part of the metadata of published documents and are therefore not copyrighted. The status of the CanMin and CJES abstracts is more restricted. Therefore, we open-source the scripts to collect and process the abstracts. This situation could change if specific permission of the respective publishers is received.

The parallel articles will also be released in accordance with the licenses of raw texts. *Comptes Rendus Géoscience* has been distributed since 2020 in partnership with the Mersenne Centre for Open Scientific Publishing based on a diamond open access policy, the journal articles and their translations distributed under a CC-BY 4.0 licence. The translated articles from the STUDENT collection are either Open Access or included in the ISTEX database.

The tools and LLMs used in our experiments are open-source or open-weights except for the commercial MT systems Systran and DeepL. We do see any ethical issues with this work.

8. Acknowledgements

This work was supported by the French national agency (ANR) as part of the MaTOS project under reference ANR-22-CE23-0033.[38] Rachel Bawden was also partly funded by her chair position in the PRAIRIE institute funded by ANR as part of the "Investissements d'avenir" programme under reference ANR19-P3IA-0001. The authors wish to thank Célia Vaudaine and Caroline Rossi for giving access to the MERSENNE corpus, to Maxime Bouthors for early work on the abstract corpus, and to Nicolas Dahan for an ex-post analysis of the data quality. The authors are also grateful to the anonymous reviewers for their insightful comments and suggestions.

[38] http://anr-matos.github.io/

9. Bibliographical References

Sadaf Abdul Rauf and François Yvon. 2024. Translating scientific abstracts in the bio-medical domain with structure-aware models. *Computer Speech and Language*, 87:101623.

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Marcello Federico, Nicola Bertoldi, Marco Trombetti, and Alessandro Cattelan. 2014. MateCat: an open source CAT tool for MT post-editing. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: Tutorials*, Vancouver, Canada. Association for Machine Translation in the Americas.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, and Bobbie Chern et al. 2024. The Llama 3 Herd of Models. Preprint arXiv:2407.21783.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565, Seattle, United States. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Lei Liu and Min Zhu. 2022. Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts. *Digital Scholarship in the Humanities*, 38(2):621–634.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. EuroLLM-9B: Technical Report. Preprint arXiv:2506.04079.

Yasmin Moslem, Rejwanul Haque, John D Kelleher, and Andy Way. 2023. Adaptive Machine Translation with Large Language Models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 80–90, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2024a. YaRN: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*.

Ziqian Peng, Rachel Bawden, and François Yvon. 2024b. À propos des difficultés de traduire automatiquement de longs documents. In *Actes de la 31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position*, pages 2–21, Toulouse, France. ATALA and AFPC.

Ziqian Peng, Rachel Bawden, and François Yvon. 2025. Investigating length issues in document-level machine translation. In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 4–23, Geneva, Switzerland. European Association for Machine Translation.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. TransQuest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5070–5081, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Dimitris Roussis, Sokratis Sofianopoulos, and Stelios Piperidis. 2024. Enhancing scientific discourse: Machine translation for the scientific domain. In *Proceedings of the 25th Annual*

Conference of the European Association for Machine Translation (Volume 1), pages 275–285, Sheffield, UK. European Association for Machine Translation (EAMT).

Paul Sebo and Sylvain de Lucia. 2024. Performance of machine translators in translating French medical research abstracts to English: A comparative study of DeepL, Google Translate, and CUBBITT. *PLOS ONE*, 19(2):1–13.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the seventh conference of the Association for Machine Translation in the America (AMTA)*, pages 223–231, Boston, Massachusetts, USA.

Dániel Varga, Péter Halaácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2005. Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005 Conference*, pages 590–596.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Longyue Wang, Zefeng Du, Wenxiang Jiao, Chenyang Lyu, Jianhui Pang, Leyang Cui, Kaiqiang Song, Derek Wong, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking and improving long-text translation with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7175–7187, Bangkok, Thailand. Association for Computational Linguistics.

Yutong Wang, Jiali Zeng, Xuebo Liu, Derek F. Wong, Fandong Meng, Jie Zhou, and Min Zhang. 2025. DelTA: An online document-level translation agent based on multi-level memory. In *The Thirteenth International Conference on Learning Representations*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, and Jiaxi Yang et al. 2025. Qwen3 technical report. Preprint arXiv:2505.09388.

Ziming Zhu, Chenglong Wang, Shunjie Xing, Yifu Huo, Fengning Tian, Quan Du, Di Yang, Chunliang Zhang, Tong Xiao, and Jingbo Zhu. 2025. LaTeXTrans: Structured LaTeX Translation with Multi-Agent Coordination. Preprint arXiv:2508.18791.

Sonia Zulfiqar, M. Farooq Wahab, Muhammad Ilyas Sarwar, and Ingo Lieberwirth. 2018. Is Machine Translation a Reliable Tool for Reading German Scientific Databases and Research Articles? *Journal of Chemical Information and Modeling*, 58(11):2214–2223.

10. Language Resource References

Esalati, Mersad and Dousti, Mohammad Javad and Faili, Heshaam. 2024. *Esposito: An English-Persian Scientific Parallel Corpus for Machine Translation*. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). ELRA and ICCL. PID https://huggingface.co/datasets/universitytehran.

Ive, Julia and Max, Aurélien and Yvon, François and Ravaud, Philippe. 2016. *Diagnosing High-Quality Statistical Machine Translation Using Traces of Post-Edition Operations*. International Conference on Language Resources and Evaluation - Workshop on Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem (MT Eval 2016).

Hannah Calzi Kleidermacher and James Zou. 2025. *Science Across Languages: Assessing LLM Multilingual Translation of Scientific Papers*. PID https://arxiv.org/abs/2502.17882.

Iñaki Lacunza and Javier Garcia Gilabert and Francesca De Luca Fornaciari and Javier Aula-Blasco and Aitor Gonzalez-Agirre and Maite Melero and Marta Villegas. 2025. *ACADATA: Parallel Dataset of Academic Data for Machine Translation*. PID https://huggingface.co/datasets/BSC-LT/ACAData.

Denis Maurel, Enza Morale, Nicolas Thouvenin, Patrice Ringot, and Angel Turri. 2019. Istex: A database of twenty million scientific papers with a mining tool which uses named entities. *Information*, 10(5).

Névéol, Aurélie and Jimeno Yepes, Antonio and Neves, Mariana and Verspoor, Karin. 2018. *Parallel Corpora for the Biomedical*

Domain. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA). PID https://github.com/biomedical-translation-corpora/corpora.

Dayyán O'Brien and Bhavitvya Malik and Ona de Gibert and Pinzhen Chen and Barry Haddow and Jörg Tiedemann. 2025. *DocHPLT: A Massively Multilingual Document-Level Translation Dataset*. PID https://huggingface.co/datasets/HPLT/DocHPLT.

Roussis, Dimitrios and Papavassiliou, Vassilis and Prokopidis, Prokopis and Piperidis, Stelios and Katsouros, Vassilis. 2022. *SciPar: A Collection of Parallel Corpora from Scientific Abstracts*. Proceedings of the Thirteenth Language Resources and Evaluation Conference. European Language Resources Association. PID https://live.european-language-grid.eu/catalogue/corpus/20067.

Salesky, Elizabeth and Darwish, Kareem and Al-Badrashiny, Mohamed and Diab, Mona and Niehues, Jan. 2023. *Evaluating Multilingual Speech Translation under Realistic Conditions with Resegmentation and Terminology*. Association for Computational Linguistics. PID https://iwslt.org/2023/multilingual.

Wang, Longyue and Du, Zefeng and Jiao, Wenxiang and Lyu, Chenyang and Pang, Jianhui and Cui, Leyang and Song, Kaiqiang and Wong, Derek and Shi, Shuming and Tu, Zhaopeng. 2024a. *Benchmarking and Improving Long-Text Translation with Large Language Models*. Findings of the Association for Computational Linguistics: ACL 2024. Association for Computational Linguistics. PID https://github.com/longyuewangdcu/Document-MT-LLM.

Wang, Longyue and Liu, Siyou and Lyu, Chenyang and Jiao, Wenxiang and Wang, Xing and Xu, Jiahao and Tu, Zhaopeng and Gu, Yan and Chen, Weiyu and Wu, Minghao and Zhou, Liting and Koehn, Philipp and Way, Andy and Yuan, Yulin. 2024b. *Findings of the WMT 2024 Shared Task on Discourse-Level Literary Translation*. Proceedings of the Ninth Conference on Machine Translation. Association for Computational Linguistics. PID https://www2.statmt.org/wmt24/literary-translation-task.html.

Validating a Pipeline to Create a Comparable Corpus of Government-Issued Travel Advisories from the Internet Archives

Laura Braun, Christian Oswald
University of the Bundeswehr Munich
Neubiberg, Bavaria, Germany
laura.braun@unibw.de, christian.oswald@unibw.de

Abstract

Government-issued travel advisories are used by citizens to get information about destination countries for tourism and other purposes such as temporary work stays or permanent relocation plans. However, qualitative evidence suggests that travel advisories may be influenced by considerations beyond current security situations. Systematic and rigorous quantitative analyses of advisories are scarce because relevant corpus data are not readily available and official government websites often provide practical obstacles. We validate a pipeline to generate a time-series cross-sectional dataset of government-issued travel advisories for three English-speaking issuing countries based on the Internet Archive's Wayback Machine. Using official government data sources that are prohibited to be scraped and used for research, we illustrate that our approach provides (near-)complete coverage. The resulting corpus and code are intended to support downstream research on comparative risk communication, international relations, and text analysis using natural language processing methods.

Keywords: travel advisories, corpus comparison, web archives, corpus construction, time-series cross-sectional data

1. Introduction

Travel advisories are designed to protect citizens from potential hazards and inform them about risks abroad. They also affect economic and political developments in the target countries of these advisories. Qualitative evidence suggests that travel advisories may also be in place for political or strategic reasons (Sharpley et al., 1996; Babey, 2019; Chu et al., 2021). However, systematic quantitative evidence for any such links is absent. Obstacles in obtaining data on travel advisories are numerous. Travel advisories are usually posted on government-administered websites. Website structures and URLs, as well as the structure of the content itself, can change over time. The content is regularly updated and replaced over time. Many of these websites also explicitly forbid the use of web scraping to obtain data and do not provide API access. Time-series cross-sectional data on government-issued travel advisories would be a valuable data source for research areas as diverse as tourism, communication, social sciences, or humanities and can be used for qualitative, quantitative, and text analysis approaches. We evaluate a transparent and reproducible pipeline to generate such a dataset for several issuing countries based on the Internet Archive's Wayback Machine (IAWM) using official government sources.

2. Background & Summary

Travel advisories and warnings show how governments communicate risks abroad. The web has been the main platform to inform the public about risks since the late 1990s (Löwenheim, 2007). Previous work links advisories to, for example, tourist behavior and policy communication in international relations (Sharpley et al., 1996; Murphy et al., 2007; Babey, 2019; Chu et al., 2021). However, these are mainly case studies and there is no global longitudinal dataset that spans various issuers, languages, and destinations. Previous research addressed this gap and designed a data collection pipeline to create time series by issuer and target (Braun and Oswald, 2025). We validate these efforts and estimate the validity of this approach and the resulting data to ensure that the travel advisory data can be reliably used in various research areas. With a multi-issuer corpus in place, research questions that were previously out of reach become possible to answer. The data allow to test, for example, convergence and divergence in risk communications between issuers, to track risk communication behavior over time, or whether the advisory represents the political and economical relationship between the issuer and the destination rather than the actual security situation on the ground.

The Internet Archive's Wayback Machine allows us to analyze web content over time as it changes (Weikum et al., 2011). URL-timestamp lists and stable snapshots allow us to reconstruct advisories that were already overwritten a long time ago. At the same time, web archives are neither neutral nor pure; they simply do not provide a fully systematic mirror of historical online content. Rather, they are socio-technical systems whose coverage results from a mix of broad crawls, institutional collections, hyperlink ecology, and individual archiving requests (Ben-David and Amram, 2018). English is the lan-

Proceedings of the 19th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC 2026, pages 96–107
11 May 2026.

guage with the highest visibility, with other European languages and Japanese following (AlNoamany et al., 2014). These patterns create an unbalanced snapshot density between issuers. In addition, the timing of snapshots varies and it can happen that an update may be captured days, weeks, or even months later. Delays between the actual content update on the web page and the time of the snapshot can hide interim updates. Access can also be affected by robot policies, site blocking, and even the location of the IAWM's server (Ben-David and Amram, 2018). All these aspects lie in the nature of the Internet Archive, but might influence the completeness of the collected data. However, the IAWM presents the best approach for obtaining cross-sectional time-series data of government-issued travel advisories in a reproducible and transparent manner in the absence of viable alternatives to obtain travel advisories at scale.

Table 1: Number of snapshots of index page with HTTP 200 responses

Country	Archive start	# IAWM 200
Australia	1997	1675
Canada (fr)	1997	1769
Canada (en)	1997	4902
China	2005	507
France	2000	2628
Germany	1998	1905
Hong Kong (ch)	2009	33
Hong Kong (en)	2009	256
India[1]	2016	241
Indonesia	2017	46
Japan	2003	2992
Mexico[2]	2003	193
Russia[3]	2016	109
United Kingdom	1997	10838
United States	1996	13342

3. Data and Corpus Construction

Our ultimate goal is to create a multilingual cross-sectional time-series corpus of government travel advisories that covers issuers on a globally representative scale. We build on a previously proposed TRAVELWARN-Crawler pipeline that can be adapted to additional issuers and languages (Braun and Oswald, 2025). We rely solely on the IAWM for our corpus construction. The IAWM offers dense coverage from the mid-2000s onward and an open CDX API for URL–timestamp enumeration, which makes it the most practical backbone for standardized retrieval over issuers and time (Murphy et al., 2007; Arora et al., 2016).

Country-specific URLs are gathered by manually identifying a list of seed URLs, which are the index pages that link to country pages (see the Australia index page example in Figures 8a and 8b). "Index page" refers to the main landing page of the travel advisory section, which then redirects to the country-specific warnings via links and, for some issuers such as the US and the UK, even presents a tabular overview with the latest update date for each country. We crawl these seeds for each issuer era and parse all links that point to country-level advisories into a database. An overview of manually identified URLs for each era and country of interest is provided in Table 3 in the Appendix. The pipeline then uses the country URLs with era boundaries. Using the CDX API for URL–timestamp enumeration, the crawler iterates over all snapshots for each country URL and parses content into the database whenever there is an unseen update date. This approach preserves overwritten content and supports longitudinal analysis.

Issuers differ in how they publish and present advisories on their index pages. Some expose the country link together with the latest update date, for example, US, Canada, France, and most of the times Australia (except around the year 2000). Others, such as the UK and Germany, do not. The Internet Archive tends to provide more snapshots of index pages than of country-specific URLs. For illustration, `https://www.smartraveller.gov.au/destinations/africa/sudan` was saved 98 times between December 3, 2019 and June 13, 2025, while `https://www.smartraveller.gov.au/destinations` was saved 545 times between December 9, 2019 and September 22, 2025. Using index page information allows us to generate validation data, although full recall is still not guaranteed.

We extract a minimal schema from each snapshot that supports replication and downstream text analysis. We store an index collection with at least the country name and country URL for each issuer and, where available, update dates or travel warning level information. We also store a text collection with country, source URL, warning date, advisory text, and where available additional metadata such as level and risk tags (Braun and Oswald, 2025). The design supports later comparisons of frequency and content between issuers.

Table 1 summarizes the total captures of index pages by all issuers of interest. The table highlights a strong imbalance in snapshot counts among issuers. The US index pages have by far the most snapshots. The UK also has many snapshots. Canada, Australia, France, and Germany are in an intermediate range. The French version of the Canadian pages has less than half the amount of snapshots as the English version. China has significantly fewer snapshots (Thelwall and Vaughan, 2004). These observations are in line with earlier

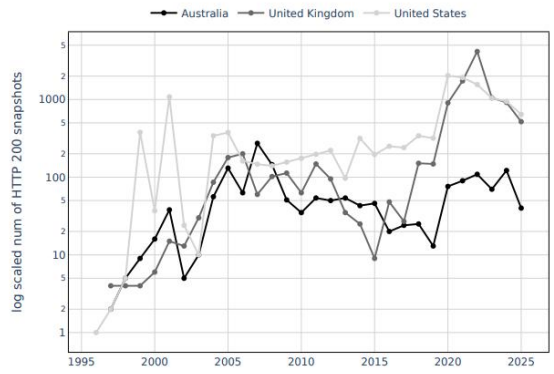


Figure 1: Yearly log-scaled number of HTTP 200 snapshots of travel advisory index pages

findings on the dominance of English language content and a Eurocentric and Western representation in web archives. We take this into account in the evaluation of the limitations of the IAWM as our data source. Table 3 in the Appendix provides a more detailed overview of the number of snapshots for each era, highlighting that the late 1990s and early 2000s have comparatively fewer captures.

Additionally, annually aggregated information on the number of HTTP 200 snapshots for the United States, UK, and Australia is plotted in Figure 1. We The values are shown on a logarithmic scale to account for the large differences in snapshot intensity between issuers and over time. Archiving for the US started in 1996 and soon reached a high snapshot density, reflecting an already high crawl intensity of its travel advisory index pages from the early 2000s onward. The UK and Australia first appeared in 1997, with fewer than ten snapshots in that year. The coverage for both subsequently increases, especially in the early 2000s, where all three converge toward roughly one hundred snapshots per year. All issuers show pronounced spikes in the number of snapshots during years affected by the Covid-19 pandemic, which may reflect general increased traffic to the warning pages. The number of snapshots increases to more than 4000 in 2022 for the UK, resulting in even several captures per day.

Note that the tables and line plot allow us to analyze the archiving behavior on the index pages only. We use country-specific pages for the data validation later in the paper, which in most cases have fewer snapshots than the corresponding index page. Analyzing the index page provides primarily a comparative estimate of how well a given issuing country is represented in the Internet Archive overall.

Table 2 provides a brief overview of the number of countries and territories for which each issuing country considered in this validation study

Table 2: Number of destination countries and territories covered by each issuer

Issuer	**# Countries & territories**
United Kingdom	230
Australia	221
United States	210

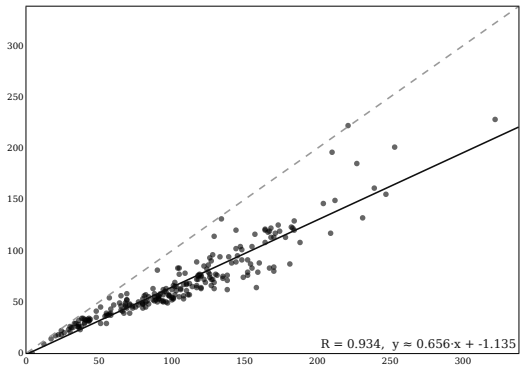
publishes travel advisories. The UK covers slightly more countries and territories than Australia and the US, as it includes a large number of island territories. The counts also include entities or states that no longer exist, e.g., Yugoslavia, as well as broader regional designations, such as “Africa (Central and Western)”, for which Australia issued a small number of advisories in the late 1990s. Given that the United Nations currently have 193 member and two non-member states, the pipeline seems to provide good coverage.

4. Validation

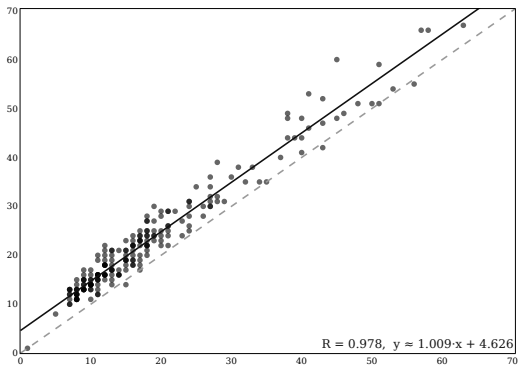
We validate the scraped data of three English-speaking issuing countries with two types of reference data. First, we derive validation targets from issuer index pages for Australia and the United States, which the Internet Archive captures more frequently than destination pages. Second, we use external sources where terms of use allow validation but not redistribution for the US and the UK.[4] We use items tagged as *Travel Advisory* from the Overseas Security Advisory Council (OSAC) catalog beginning in 2004 for the US. OSAC is a public-private partnership that seeks to help protect US interests overseas and is part of the Bureau of Diplomatic Security of the US Department of State. We use captures from the UK National Archives, similar to the IAWM, via a CDX endpoint and apply the same extraction logic as in Braun and Oswald (2025) for the UK. The National Archives also store the full timeline of published online advisories. Compared to the Internet Archive, their crawler nowadays runs systematically once every working day, but also has bigger gaps in earlier years. Note that the National Archives data can be used for validation purposes, but takedown and reclosure policies make scientific use and sharing of the data challenging.

We assess agreement and coverage. We fit an ordinary least squares regression model of our per country counts on the reference counts for matched years to measure agreement. Perfect agreement lies on the line $x = y$, and a larger R with a slope close to one indicates better alignment. We compare the annual number of advisories per destina-

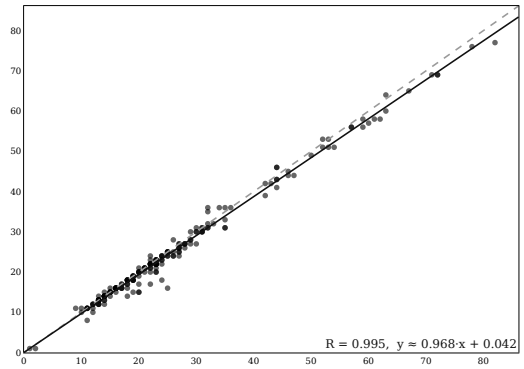
[4]We provide scripts to replicate the collection of the validation sources.



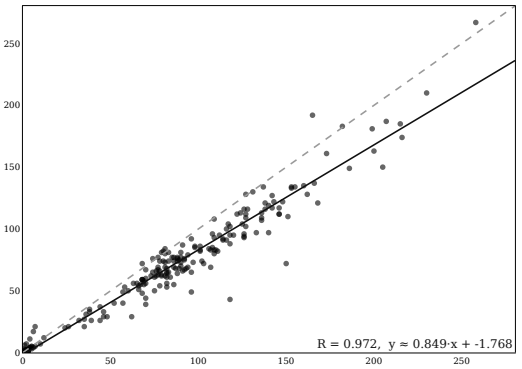
(a) UK vs. UK validation



(b) US vs. OSAC



(c) US vs. US index



(d) AU vs. AU index

Figure 2: Per country agreement between Internet Archive based counts of travel advisories and validation data. Each panel plots, for one selected country, the number of advisories in the reference data (horizontal axis) against the number recovered from the Internet Archive (vertical axis), together with the comparison line ($x = y$).

tion against the reference to measure coverage and summarize the distribution as violin plots with per year means. None of the reference data guaranties perfect recall. Our corpus can show overcoverage when it contains versions that are missing from the reference data. Validation is not affected by timing issues, as we base annual measurements only on the warning date rather than the snapshot date.

Figure 2a compares per-country advisory counts derived from the UK country pages in our Internet Archive corpus with counts based on the UK National Archives country pages. The regression line has a slope of about 0.66 and a negative intercept, which implies that we recover slightly more than two thirds of all advisories and that the shortfalls become larger for destinations with many updates. The annual coverage distribution in Figure 4a shows that this undercoverage fluctuates over time, with some earlier years showing wider negative tails, whereas later years, particularly 2022, cluster closer to the reference data. Figure 5 in the Appendix contrasts UK country pages as archived by the Internet Archive and by the UK

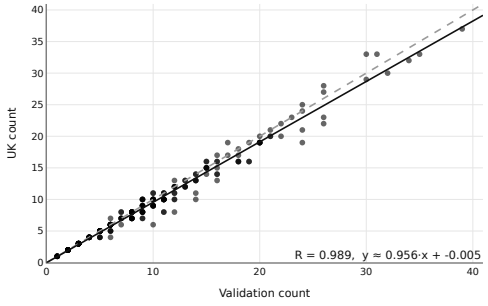


Figure 3: Agreement Internet Archive UK vs. The National Archives UK for the year 2022

National Archive only for the year 2022, a year in which the Internet Archive captures the index page almost daily, and with it seemingly also the country pages more often, as the almost perfect agreement indicates.

Although Figure 4a reveals a loss of updates for

years other than 2022, the correlation is high, and the coverage violins show no evidence of systematic exclusion of, e.g., a specific set of countries. Instead, missing observations are dispersed across destinations, so that relative differences between countries and overall temporal dynamics seem to be preserved even though absolute UK update counts are systematically underrepresented in our corpus.

The agreement between OSAC and our US corpus is strong with $R = 0.978$ and a slope close to one, as shown in Figure 2b. Notice the positive intercept of about $+4.6$, which indicates a small surplus on our side. Put differently, country level counts are systematically higher in our corpus, which in turn suggests that almost all destinations listed by OSAC are present in our text set and that we recover additional updates that do not appear in the OSAC catalog. The coverage plot in Figure 4b shows that this surplus is concentrated in recent years, especially between 2021 and 2023, where the OSAC catalog lists comparatively few entries. Screenshots of the OSAC catalog in the Appendix (Figure 6) highlight that filtering for 2023 only returns 91 entries compared to the year 2019 with more than 300 results. We do not have any insight into this gap, but it illustrates that collecting perfect data is challenging.

We also compare the US corpus to the US index derived target from the archive itself. Figure 2c shows an almost perfect fit with $R = 0.995$ and a slope of about 0.968. On average, about 3% of updates per country fall between captures of the index page. The coverage in Figure 4c is tightly centered near parity with a few expected outliers when a destination has only one advisory in the reference but none in the text set, which would translate to a coverage of -1. Given the large number of snapshots for US index and country pages and the moderate US update frequency compared to other issuers, we are confident to have a near-complete reconstruction for the US Department of State travel advisories.

Figure 2d shows high agreement between our scraped data and the Australian index target with $R = 0.972$ and a slope around 0.849. Undercoverage in our data grows for destinations with many updates, which fits the lower snapshot density of Australian country pages relative to index pages. The coverage violins in Figure 4d confirm wider negative tails in years with rapid update activity. The index design did not expose update dates around the year 2000 (see Figure 8a), which reduces the quality of the index-based target in that period. This explains a spike toward overcoverage in our text data when country pages carry advisory updates that the index does not surface.[5]

[5] Figure 2d does not reflect this design change, which may slightly lower R.

5. Discussion

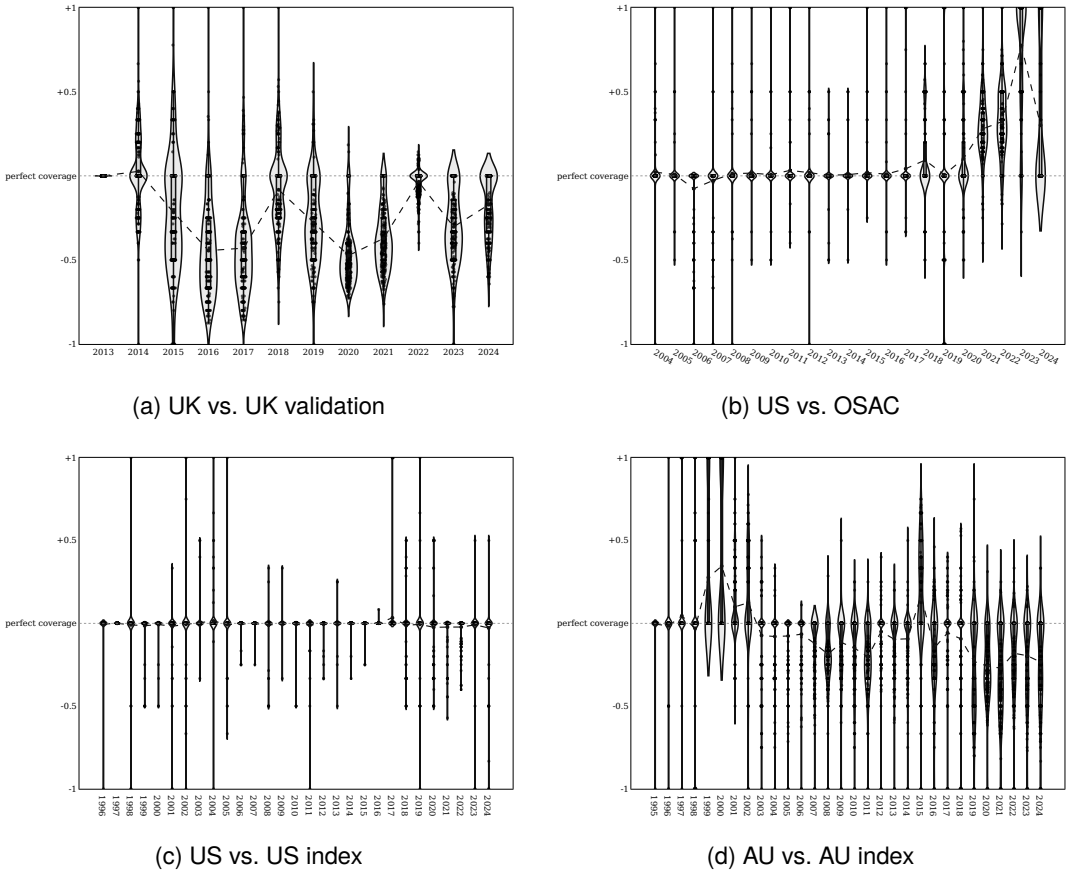
The validation results show that the IAWM, although not without limitations, serves as a reliable source for reconstructing advisory timelines at scale. We obtain a near complete reconstruction for US travel advisories. The agreement with OSAC and with the US index is high and coverage is close to complete. For the UK, comparing country-level counts to the UK National Archives reveals undercoverage in absolute terms, but also very high correlations and no evidence of a subset of destinations being dropped entirely. The UK and Australian results together suggest that the main weakness of our Internet Archive based approach lies in capturing issuers with very frequent updates, especially when snapshots are infrequent. Additionally, not only the total count of updates by country, but also previous work indicates that the US updates less frequently with more major changes per update than the UK and Australia with more updates containing little or only editorial changes.

We expect to get similar results in the future for other issuers with reasonably dense archiving history, such as Canada, Germany, Japan, and France, as for Australia and UK in terms of completeness, while issuers with even fewer snapshots or additional access barriers are likely to be less well covered. We can further deduce insights for our stated goal of building a multilingual global corpus with these results. We identify three main limitations for this goal that are not only technical properties of the Internet Archive but also reflect political economy, socio-technical, and international relations constraints:

1. Some states do not issue advisories at all or have started doing so much later. Others issue advisories with a far smaller scope. Although this is not an archive issue, it affects global analyses and makes some comparisons uneven by design.

2. Language and location visibility matter. English pages have a higher snapshot density on average. We observe this for Canada, where the English version of the index page is better represented than the French version. Similarly, locations such as Europe, the US and Japan seem to be better covered.

3. Access and capture can be shaped by blocking and geo-dependent hurdles. Archived replays return redirects or challenge pages that we cannot parse in some cases. This may occur for domains in specific countries and periods and is a known socio-technical constraint.



(a) UK vs. UK validation

(b) US vs. OSAC

(c) US vs. US index

(d) AU vs. AU index

Figure 4: Yearly coverage of Internet Archive-based advisories relative to (official) reference datasets. Each panel shows, for a given home country and year, the distribution of coverage scores across destination countries (violin plots with per-country points and yearly means).

Our findings support the use of the IAWM for large scale data collection. The IAWM is suitable for multilinguality and different issuer locations when it comes to European countries and Japan, but might lack data due to lower snapshot density for other countries either because of location or language (e.g., Hong Kong) together with often later start (e.g., Indonesia). It is still a good, reproducible, and transparent approach in the absence of viable alternatives.

6. Conclusion

We validated a transparent, reproducible, scalable, and language-independent pipeline to create time-series data for government-issued travel warnings using the Internet Archive's Wayback Machine as a data source. By validating our scraped corpus for the US, the UK, and Australia as major English-speaking countries against various reference sources, we showed that an Internet Archive-based reconstruction can achieve high coverage for some issuers and robust, though not exhaustive coverage for others. The main limitation resulting from our validation analysis for the three selected issuers arises for cases with very frequent updates, where the amount of archived snapshots does not always suffice to capture every version of every country page. At the same time, we did observe differences in terms of data availability of issuing governments of non-English speaking, non-Western-aligned countries for future extensions of the pipeline.

Despite these constraints, the resulting corpus already functions as a comparable resource across issuers and destinations, enabling longitudinal analyses of how states communicate risk and comparative studies of advisory levels and update dynamics. For many applications, it is more important to preserve relative differences and major changes over time than to observe every minor revision, and our corpus achieves this goal. In future work, we plan to extend the approach presented in this paper to include additional countries. Although we aim to create a geographically and linguistically balanced dataset, some of the limitations we outlined above may impede this goal to some extent. In addition, we plan to create additional variables that we want to extract from the advisory text using natural language processing methods, such as more granular

geographic locations and risk tags. We will make the data publicly available once we have a comprehensive and geographically and linguistically diverse corpus and dataset of government-issued travel advisories, which we envision to be valuable for qualitative and quantitative studies in various research areas.

7. Limitations

Our corpus construction relies entirely on the Internet Archive's Wayback Machine as a single archival source. Coverage in this archive is uneven across countries, time periods, and languages, and crawls can miss interim updates or complete pages because of robots.txt policies, technical outages, or crawl scheduling decisions. As a result, the data provide an approximation rather than an exhaustive record of all issued advisories. Our validation also depends on institutional reference data, such as the UK National Archives and the OSAC catalogue, which themselves have temporal gaps and are not guaranteed to offer perfect recall. Finally, we focus on English-language advisories from a small set of issuing countries with comparatively dense archival coverage in this paper, so the generality of our findings to other issuers, languages, and regions remains to be tested. However, we need to establish a baseline against which to validate less well covered issuing countries in the future, for which the countries chosen in this paper are ideal due to the geographic and linguistic reasons with regard to the IAWM outlined above.

8. Acknowledgements

We thank David Bencek, André Bluhm, Marius Hofmann and Daniel Racek, as well as the editors and two anonymous reviewers, for their helpful comments and feedback. We furthermore thank Jake Bickford and Natasha Kitcher from the UK National Archives for their support in obtaining the UK data for validation. The Center for Crisis Early Warning (Kompetenzzentrum Krisenfrüherkennung) is funded by the German Federal Ministry of Defense and the German Federal Foreign Office. The views and opinions expressed in this article are those of the author(s) and do not necessarily reflect the official policy or position of any agency of the German government.

9. References

Yasmin AlNoamany, Ahmed AlSum, Michele C. Weigle, and Michael L. Nelson. 2014. Who and what links to the Internet Archive. *International Journal on Digital Libraries*, 14(3-4):101–115.

Sanjay K. Arora, Yin Li, Jan Youtie, and Philip Shapira. 2016. Using the wayback machine to mine websites in the social sciences: A methodological resource. *Journal of the Association for Information Science and Technology*, 67(8):1904–1915.

Nicholas George Babey. 2019. The Politics of Travel Advisories: Foreign Policy and Error in Canada's Traveller Information Program. *The Journal of Intelligence, Conflict, and Warfare*, 2(1):15–33. Number: 1.

Anat Ben-David and Adam Amram. 2018. The Internet Archive and the socio-technical construction of historical facts. *Internet Histories*, 2(1-2):179–201.

Laura Braun and Christian Oswald. 2025. TRAVELWARN-Crawler: Constructing longitudinal datasets of government-issued travel warnings for political and social science research. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, pages 29–35, Hannover, Germany. HsH Applied Academics.

Yinxiao Chu, Xiaoyu Huang, and Tao Jin. 2021. Political relations and tourism: evidence from China. *Applied Economics*, 53(45):5281–5302. Publisher: Routledge _eprint: https://doi.org/10.1080/00036846.2021.1922591.

Oded Löwenheim. 2007. The Responsibility to Responsibilize: Foreign Offices and the Issuing of Travel Warnings. *International Political Sociology*, 1(3):203–221.

Jamie Murphy, Noor Hazarina Hashim, and Peter O'Connor. 2007. Take Me Back: Validating the Wayback Machine. *Journal of Computer-Mediated Communication*, 13(1):60–75.

Richard Sharpley, Julia Sharpley, and John Adams. 1996. Travel advice or trade embargo? The impacts and implications of official travel advice. *Tourism Management*, 17(1):1–7.

Mike Thelwall and Liwen Vaughan. 2004. A fair history of the Web? Examining country balance in the Internet Archive. *Library & Information Science Research*, 26(2):162–176.

Gerhard Weikum, Nikos Ntarmos, Marc Spaniol, Peter Triantafillou, András Benczúr, Scott Kirkpatrick, Philippe Rigaux, and Mark Williamson. 2011. Longitudinal Analytics on Web Archive Data: It's About Time! pages 109–202.

A. Supplementary Material

Table 3: Country–issuer URLs

Country	Time-start	URL	# all	# 200
Australia	1997	http://www.dfat.gov.au/consular/advice/advices_mnu.html	153	47
	1998	http://www.dfat.gov.au/consular/advice/consadvice_main.html	31	7
	1998	http://www.dfat.gov.au/consular/advice/index.html	224	38
	2004	http://www.smartraveller.gov.au/zw-cgi/view/Advice/	1463	991
	2015	http://smartraveller.gov.au/countries/list.html	47	14
	2016	http://smartraveller.gov.au/Countries/Pages/list.aspx	172	65
	2019	https://www.smartraveller.gov.au/destinations	545	513
		Total	**2635**	**1675**
Canada (en)	1997	http://www.dfait-maeci.gc.ca/graphics/cosmos/CNTRY_E.htm	139	59
	2000	http://voyage.dfait-maeci.gc.ca/destinations/menu_e.htm	145	62
	2003	http://www.voyage.gc.ca/dest/ctry/reportpage-en.asp	1057	772
	2008	http://www.voyage.gc.ca/countries_pays/menu-eng.asp	581	159
	2012	http://travel.gc.ca/travelling/advisories	4699	3850
		Total	**6621**	**4902**
Canada (fr)	1997	http://www.dfait-maeci.gc.ca/graphics/cosmos/cntry_f.htm	42	22
	1999	http://www.dfait-maeci.gc.ca/travelreport/menu_f.htm	93	14
	2003	http://www.voyage.gc.ca/dest/ctry/reportpage-fr.asp	398	267
	2009	http://www.voyage.gc.ca/countries_pays/menu-fra.asp	444	109
	2012	http://voyage.gc.ca/voyager/avertissements	1908	1357
		Total	**2885**	**1769**
China	2005	http://www.fmprc.gov.cn/chn/lsfw/lsxw/fbfgjhcszysx/default.htm	42	36
	2012	http://cs.mfa.gov.cn/lsyj/gbtx/	5	5
	2013	https://cs.mfa.gov.cn/gyls/lsgz/lsyj/	600	466
		Total	**647**	**507**
France	2000	http://www.dfae.diplomatie.fr/voyageurs/etrangers/avis/conseils/minute.asp	38	38
	2005	http://www.diplomatie.gouv.fr/fr/conseils-aux-voyageurs_909/index.html	850	684

Continued on next page

Country	Time-start	URL	# all	# 200
	2006	http://www.diplomatie.gouv.fr/fr/conseils-aux-voyageurs	2167	1906
		Total	**3055**	**2628**
Germany	1998	http://www.auswaertiges-amt.de/5_laende/index.htm	160	65
	2000	http://www.auswaertiges-amt.de/www/de/laenderinfos/reise_warnung_html	281	161
	2006	http://www.auswaertiges-amt.de/diplo/de/Laenderinformationen/01-Reisewarnungen-Liste.html	187	162
	2006	http://www.auswaertiges-amt.de/diplo/de/Laenderinformationen/01-Reisewarnungen.html	63	32
	2010	http://www.auswaertiges-amt.de/DE/Laenderinformationen/01-Reisewarnungen-Liste_node.html	250	136
	2017	https://www.auswaertiges-amt.de/de/ReiseUndSicherheit/10.2.8Reisewarnungen	1992	1296
	2024	https://www.auswaertiges-amt.de/de/reiseundsicherheit/10-2-8reisewarnungen	63	53
		Total	**2996**	**1905**
Hong Kong (ch)	2021	https://www.sb.gov.hk/chi/ota/index.html	37	33
		Total	**37**	**33**
Hong Kong (en)	2009	https://www.sb.gov.hk/eng/ota/	486	256
		Total	**486**	**256**
India	2016	https://www.mea.gov.in/travel-advisories.htm	254	241
		Total	**254**	**241**
Indonesia	2017	https://safetravel.kemlu.go.id/	49	46
		Total	**49**	**46**
Japan	2003	http://www.anzen.mofa.go.jp/	6899	2992
		Total	**6899**	**2992**
Mexico	2003	http://www.sre.gob.mx/delviajero/	255	180
	2024	https://portales.sre.gob.mx/guiadeviaje/	13	13
		Total	**268**	**193**
Russia	2016	http://www.mid.ru/ru/preduprezdenie-dla-rossijskih-grazdan1	125	107
	2023	https://www.mid.ru/ru/useful_information/information/preduprezhdeniya_dlya_rossiyskikh_grazhdan/	2	2
		Total	**127**	**109**
SAR Macao	2023	https://www.dst.gov.mo/zh-hant/tourism-crisis-management/tourism-crisis-management-travel-alert.html	13	13
		Total	**13**	**13**
United Kingdom	1997	http://www.fco.gov.uk:80/reference/travel_advice/countries.html	56	26

Continued on next page

Country	Time-start	URL	# all	# 200
	1998	http://193.114.50.10/travel/	44	15
	2002	http://www.fco.gov.uk/servlet/Front?pagename=OpenMarket/Xcelerate/ShowPage&c=Page&cid=1007029390590	764	574
	2008	http://www.fco.gov.uk/en/travelling-and-living-overseas/travel-advice-by-country/	397	179
	2009	http://www.fco.gov.uk/en/travel-and-living-abroad/travel-advice-by-country/	588	345
	2013	https://www.gov.uk/foreign-travel-advice	11382	9699
		Total	**13231**	**10838**
United States	1996	http://travel.state.gov/travel_warnings.html	4970	1766
	2004	http://travel.state.gov/travel/cis_pa_tw/tw/tw_1764.html	6874	1672
	2004	http://travel.state.gov/travel/warnings_current.html	534	112
	2014	http://travel.state.gov/content/passports/english/alertswarnings.html	4110	463
	2015	https://travel.state.gov/content/passports/en/alertswarnings.html	5129	527
	2017	https://travel.state.gov/content/travel/en/traveladvisories/traveladvisories.html	19198	8647
	2025	https://travel.state.gov/en/international-travel/travel-advisories.html	156	155
		Total	**40971**	**13342**

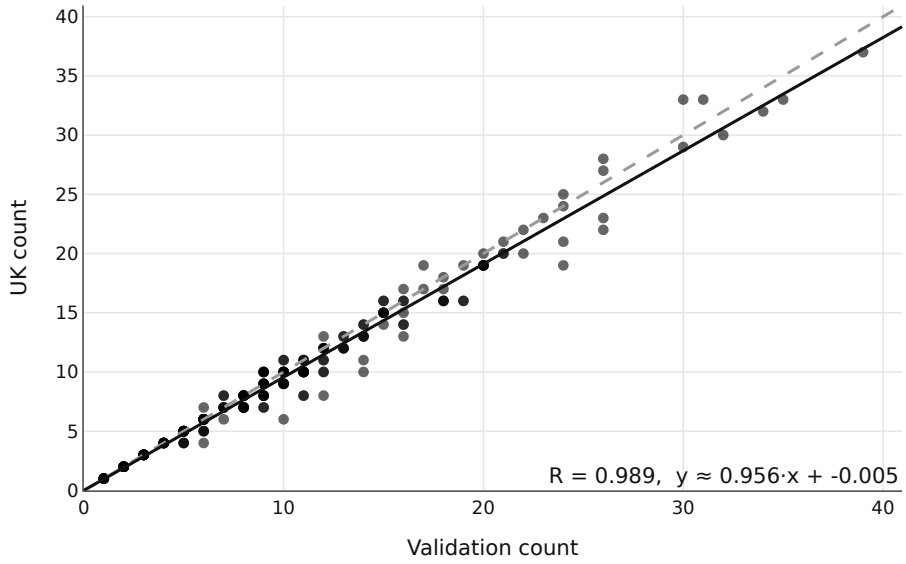


Figure 5: Agreement Internet Archive UK vs. The National Archives UK for the year 2022

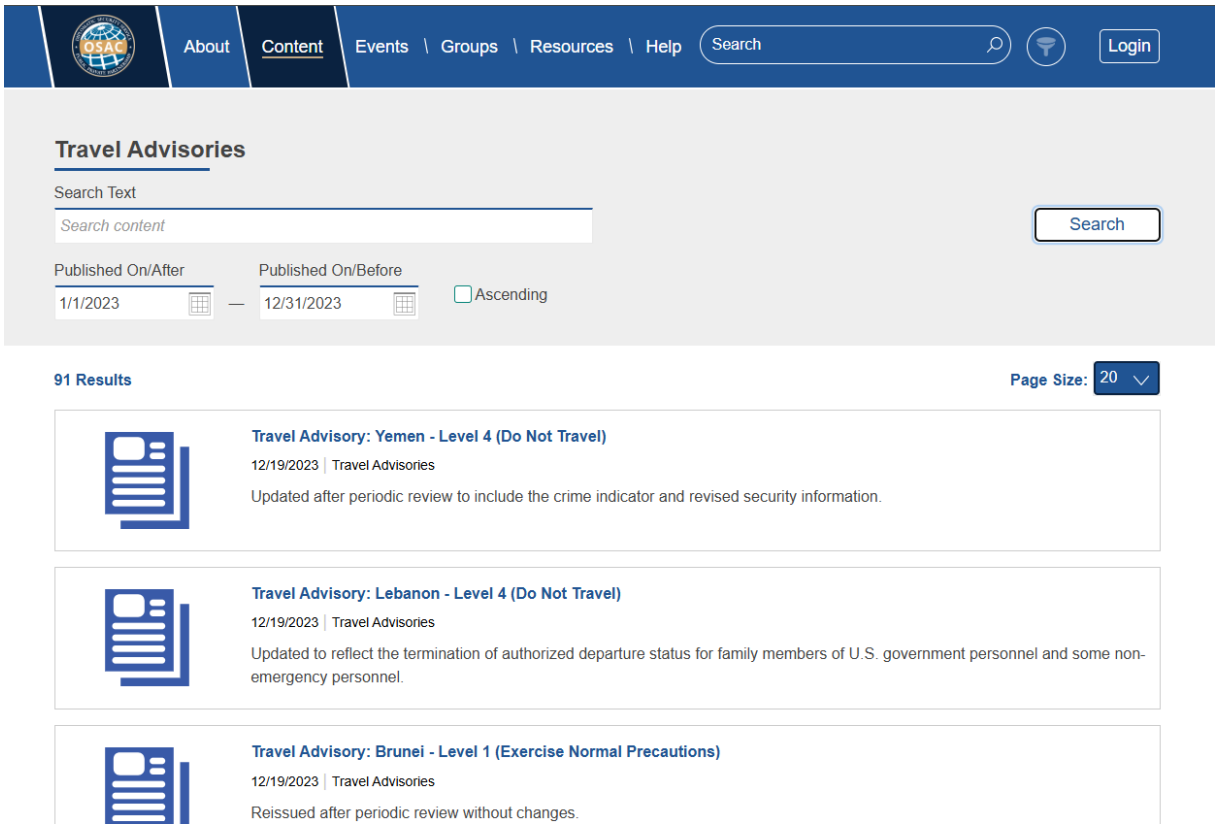


Figure 6: OSAC catalogue with filter on year 2023

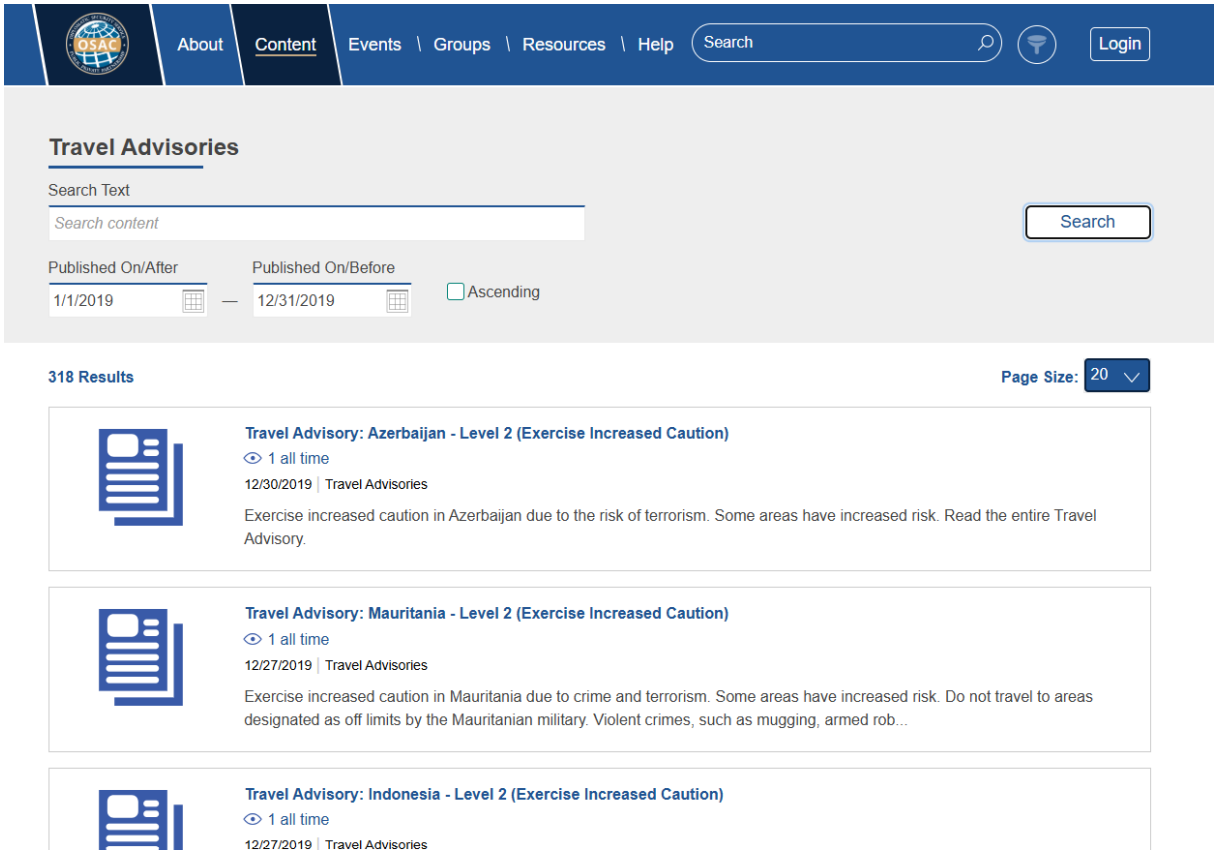
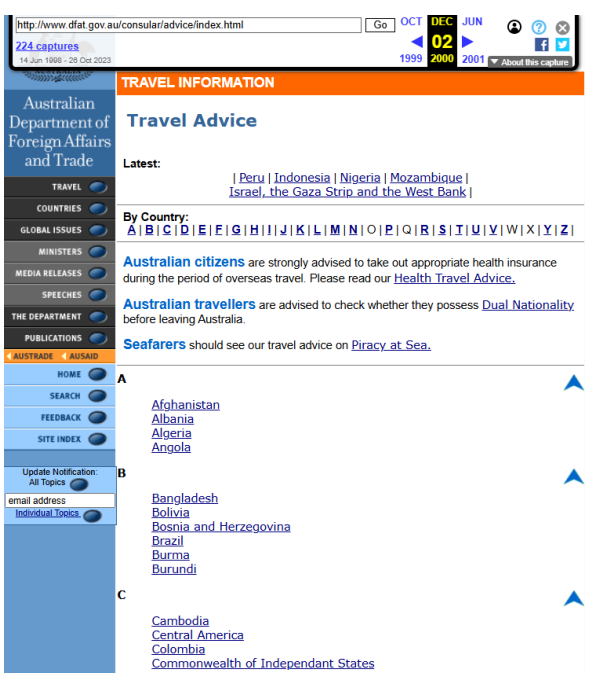
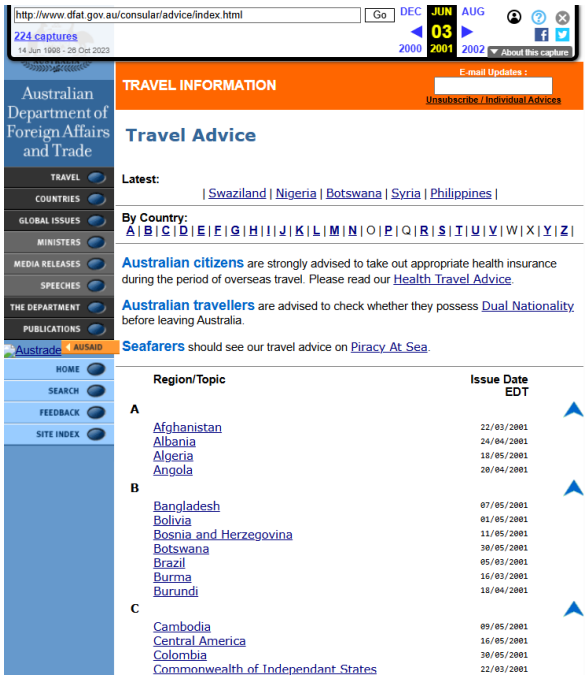


Figure 7: OSAC catalogue with filter on year 2019



(a) Australia advisory index page (without date)



(b) Australia advisory index page (with date)

Figure 8: Australia advisory index pages before and after the design change

Leveraging Comparable Toxicity Lexicons in Prompt Instructions for Multilingual Text Detoxification

Yassir El Attar[1], **Esra Dönmez**[1,2], **Nina Ohlendorf**[1], **Agnieszka Falenska**[1,2]
[1]Institute for Natural Language Processing, University of Stuttgart, Germany
[2]Interchange Forum for Reflecting on Intelligent Systems, University of Stuttgart, Germany
{yassir.el-attar, esra.doenmez, nina.ohlendorf, agnieszka.falenska}@ims.uni-stuttgart.de

Abstract

To mitigate the prevalence of toxic language on digital social media, various NLP approaches have been proposed for automatic text detoxification. However, the potential of toxic expressions lexicons as a comparable cross-lingual resource to guide this process remains largely unexplored. In this work, we investigate how such resources can be effectively used to inform multilingual language models about what should and should not be considered *toxic*. We evaluate four models under two settings—zero-shot prompting and fine-tuning—to assess the impact of incorporating toxic expressions in prompt instruction, including in cross-lingual transfer scenarios. Our results show that both zero-shot prompting and fine-tuning approaches benefit considerably from adding toxic expressions in prompt instructions during training and/or inference. Our findings demonstrate that comparable, lightweight, language-specific toxic expressions lexicons constitute an effective mechanism for injecting explicit information about lexical toxicity into multilingual language models.

Keywords: text detoxification, multilinguality, cross-lingual transfer, comparable corpora, low-resource languages

1. Introduction

Disclaimer: certain figures and examples include potentially offensive content.

Toxic language, following the criteria by Dementieva et al. (2024b), is defined as text containing vulgar or profane content, regardless of whether it directly targets or insults individuals or groups. For instance, a message such as *"I f*cking love this movie!!"* is toxic due to its use of profane language, yet it carries no hateful or offensive intent.

Toxic content is highly prevalent on the internet, especially on social media and in online forums (Vasist et al., 2023; Radfar et al., 2020). It is known to be harmful to people's mental well-being (Waldron, 2012) and specifically affects minority groups (Thomas et al., 2021) and children (Breckheimer, 2001). These potential harms motivate the *text detoxification* task, an automatic mitigation approach defined as a form of text style transfer (Dale et al., 2021) in which the vulgar style of a message is neutralized while its meaning is kept intact. For instance, toxic *"I don't give a sh*t about your opinion!"* could be detoxified into *"I don't care about your opinion!"*: the style changes, but the message stays the same.

The feasibility of the detoxification task increased with the introduction of Large Language Models (LLMs) (Logacheva et al., 2022). However, despite this advancement, detoxification remains a challenging task. This challenge is evidenced by the fact that language models are often pretrained on filtered data in which toxic content has been removed (Mendu et al., 2025). While this filtering is intended to reduce harmful outputs, it may also impair the models' ability to recognize profane or abusive expressions. The most common strategy for addressing this problem is to fine-tune models on large collections of toxic–detoxified text pairs. However, this approach requires substantial amounts of annotated data, which are costly to create. Consequently, such resources are often unavailable for low-resource languages, and multilingual coverage frequently depends on machine-translated corpora (Rykov et al., 2024), potentially introducing additional noise and bias.

A recent line of work has explored a complementary approach: using **comparable, language-specific lexicons of toxic expressions** (e.g., swear words) to inform LM-based detoxification. These corpora consist of lexicons constructed independently across languages around the same conceptual domain, rather than obtained through direct translation (Dementieva et al., 2024b). Importantly, they are considerably less resource-intensive to construct than parallel toxic–detoxified datasets. However, there is currently no consensus on how such corpora should be integrated into detoxification pipelines. Existing approaches include detecting toxic expressions and removing them directly from text (Dementieva et al., 2024b), risking loss of meaning, or masking them for the model to replace (Nuthakki et al., 2025), which can result in unnatural outputs. In a more similar work to ours, Lai-Lopez et al. (2025) proposed tagging toxic expressions

Proceedings of the 19th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC 2026, pages 108–118
11 May 2026.

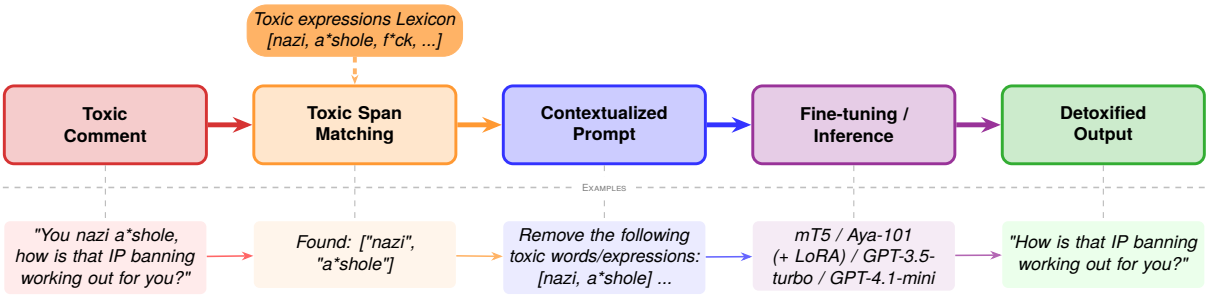


Figure 1: Overview of the toxic-expressions-in-instruction strategy. The upper row illustrates the processing pipeline stages, while the lower row provides a concrete example for each stage. The toxic expressions lexicon feeds into the matching/lookup stage to identify toxic expressions before using them to create our toxic-contextualized prompt instruction.

in inputs via markup (`<toxic>...</toxic>`). Yet, encoding lexical toxicity through such input annotations rather than through explicit instructions leaves open the question of whether Language Models (LMs) can be more effectively guided by directly specifying which expressions are toxic.

In this work, we ask *how to effectively inform language models* about what should and should not be considered *toxic*. We hypothesize that providing toxic expressions from comparable lexicons alongside the input sentence, together with explicit instructions to remove or replace them while preserving meaning, can lead to more controlled and semantically faithful detoxification. This motivates the following research questions:

RQ1: Does providing toxic expressions in model prompt instructions improve detoxification performance across model adaptation settings?

RQ2: To what extent do these improvements generalize in a cross-lingual setup, particularly for low-resource languages not seen during training?

To answer these questions, we evaluate the performance of four multilingual language models with and without toxic expressions in prompt instructions. Figure 1 demonstrates the overall pipeline with an example at each stage. Evaluation covers two model-adaptation settings: zero-shot prompting and fine-tuning. All experiments are evaluated on 15 languages from diverse typological families. In the case of fine-tuned models, evaluation additionally assesses cross-lingual generalization: the models are trained on data from 9 languages and then evaluated both on those languages and on 6 additional unseen languages. We find that across all strategies, model sizes, and language settings, providing toxic expressions in instructions consistently improves detoxification performance (Section 4). Moreover, the proposed method achieves remarkable results, demonstrating that this simple instruction-based knowledge injection is competitive with more data-intensive or architecture-specific approaches. Our results highlight the value of investing in the creation and curation of multilingual toxic expressions lexicons as a comparable resource. These resources can serve as a general mechanism for injecting domain-specific conceptual knowledge into language models, with potential benefits extending well beyond the detoxification task. Our experimental code and scripts are publicly available on GitHub[1].

2. Related Work

Toxic language is commonly conflated in the literature with other related concepts (Fortuna et al., 2020). However, following the definition by Dementieva et al. (2024b), it differs from broader notions such as hate speech, which targets individuals or groups based on characteristics such as race, gender, or religion (Davidson et al., 2017; Basile et al., 2019), or offensive language, which encompasses a wider range of socially unacceptable expressions that may not involve profanity (Fortuna et al., 2020). As Fortuna et al. (2020) highlight, these categories refer to distinct phenomena and require different methodological approaches for their detection and/or removal.

In order to prevent digital violence (Shi et al., 2020) and maintain constructive communication, AI models have been developed to detect (D'Sa et al., 2020; Zampieri et al., 2020), delete (Dementieva et al., 2024b) or block (Cobbe, 2021) toxic language. This detoxification process is a text style transfer (TST) task (Dale et al., 2021): The source style to be changed is the harmful toxic language, which is automatically transformed to the target style, the non-toxic, neutral language counterpart (Mukherjee et al., 2023a). Beyond the style change, the primary objective is to generate text that is fluent and preserves the original text's meaning as much as possible (Dementieva et al., 2021). As a supervised sequence-to-sequence task, this can be per-

[1] https://github.com/YassirELATTAR/multilingual-text-detoxification

	Language	Train	Test	# Toxic expressions
Seen languages	English	19,744 + 400	600	3,390
	Russian	12,206 + 400	600	141,000
	Ukrainian	400	600	7,360
	German	400	600	247
	Arabic	400	600	430
	Spanish	400	600	1,200
	Hindi	400	600	133
	Chinese	400	600	3,840
	Amharic	400	600	245
	Total	**35,550**	**5,400**	**157,845**
Unseen	French	—	600	1,290
	Italian	—	600	815
	Japanese	—	600	328
	Hinglish	—	600	209
	Tatar	—	600	15,600
	Hebrew	—	600	731
	Total	**—**	**3,600**	**18,973**

Table 1: Dataset statistics and toxic expressions lexicon size per language. More details on the sources and the data collection process can be found in Dementieva et al. (2025, 2024a); Logacheva et al. (2022); Dementieva et al. (2024b).

formed using encoder-decoder models trained on parallel data (Logacheva et al., 2022). Although unsupervised approaches exist (Nogueira dos Santos et al., 2018; Floto et al., 2023), supervised methods leveraging parallel corpora have proven particularly effective (Logacheva et al., 2022; Atwell et al., 2022). Subsequent work has focused on fine-tuning sequence-to-sequence models (Zhang et al., 2024), with approaches ranging from mT0 fine-tuning (Dementieva et al., 2024a) and GPT-4 few-shot prompting (Dementieva et al., 2025) to LoRA-based fine-tuning of Gemma-3 (12B) as the current state-of-the-art (Dang and D'Elia, 2025). However, to our knowledge, no prior work has systematically investigated toxic expressions lexicons as comparable corpora across both fine-tuning and prompting paradigms, nor evaluated their cross-lingual transferability to unseen languages, which is the gap the present work addresses.

3. Experimental Setup

We first describe the data resources, including datasets of toxic inputs paired with detoxified target rewrites and comparable multilingual toxic lexicons. Next, we detail experiments covering text detoxification settings, the toxic expressions matching and instruction construction procedures, and the evaluation metrics.

3.1. Data Resources

We make use of two resources: (1) **datasets of toxic comments paired with non-toxic (neutral) rewrites**, with training and test splits in 9 languages (see Table 2 for a few examples and Table 1 for dataset statistics) and additional test sets in six languages, and (2) **a comparable multilingual toxicity lexicon**[2] covering all 15 languages.

Datasets The multilingual dataset (Dementieva et al., 2025) includes training and test splits, with languages grouped into **seen** and **unseen** (as shown in Table 1). It provides 400 training instances per seen language[3] (English, Russian, Ukrainian, German, Arabic, Spanish, Hindi, Chinese, and Amharic) and 600 test instances per language for all seen and unseen languages: French, Italian, Japanese, Hinglish (in Latin alphabet), Tatar, and Hebrew[4].

Additionally, for training, we include English data from Logacheva et al. (2022)[5] ($19,744$ instances), and Russian from Dementieva et al. (2024a)[6] ($12,206$ instances) as shown in details in Table 1[7]. All instances are pairs of toxic comments and their detoxified (neutral) rewrites. After adding high-resource data (English and Russian in our case), the resulting training set becomes imbalanced, reflecting real-world differences in data availability across languages. For a detailed description of the datasets and their collection process, see Dementieva et al. (2025, 2024a,b) and Logacheva et al. (2022). In summary, the input and target pairs were obtained using a collection pipeline (Logacheva et al., 2022) in which human annotators were instructed to manually rewrite each toxic comment into a non-toxic paraphrase, verifying that the target rewrite is (1) non-toxic, (2) fluent, but may contain some minor mistakes depending on the input, and (3) semantically faithful to the original content. Any user names and links were anonymized.

Comparable multilingual toxicity lexicon (Dementieva et al., 2024b) The lexicon is a collection of toxic expressions across 15 languages ($176,818$ instances). It was compiled from multiple sources, existing community-maintained toxic expressions lists for most languages, and manually curated lists for Amharic and Arabic where no such resources existed. Additionally, the Tatar lexicon was created by merging an existing list with Russian toxic expressions translated into Tatar. The lexicons vary considerably in size across languages, rang-

[2] https://huggingface.co/datasets/textdetox/multilingual_toxic_lexicon

[3] https://huggingface.co/datasets/textdetox/multilingual_paradetox

[4] https://huggingface.co/datasets/textdetox/multilingual_paradetox_test

[5] https://huggingface.co/datasets/s-nlp/paradetox

[6] https://huggingface.co/datasets/s-nlp/ru_paradetox

[7] Both of these additional datasets come from the same underlying resource.

Lang.	Toxic comment	Neutral comment
English	lol i'm just f*ckin with ya!	lol i'm just playing with ya!
Spanish	Este país se va a la m*erda *(this country is going to sh*t)*	nada puede salvar a este país *(nothing can save this country)*
German	Weit und breit kein N*ger. *(Not a single n*gro in sight.)*	Weit und breit kein Schwarzer. *(Not a single black person in sight.)*

Table 2: Example pairs of toxic comments and non-toxic rewrites in the dataset from three different languages. The Spanish and German comments are accompanied by their translation.

Instruction-only	Toxic-expressions-in-instruction
Detoxify the sentence: "lol i'm just f*ckin with ya!".	Remove the following toxic words/expressions [f*ckin] from the sentence: "lol i'm just f*ckin with ya!", but keep the meaning and style similar to the original sentence.

Table 3: Prompts used in both zero-shot and fine-tuning settings. Instruction-only (w/o) simply instructs the model to detoxify the sentence, while toxic-expressions-in-instruction (w/) provides toxic expressions (words/expressions) as context in the input instruction.

ing from 133 entries for Hindi to $141,000$ for Russian (Table 1), reflecting the differences in resource availability rather than actual differences in toxic language use.

3.2. Experiments

We design and experiment with two settings (aka. model adaptations): **zero-shot prompting** and **fine-tuning**. In both settings, we employ two instruction strategies to test the effects of toxic expressions in prompt instructions on the model performance: **(a) instruction-only** (left column in Table 3; simply instructs models to detoxify the input based on its knowledge and the input data alone) and **(b) toxic-expressions-in-instruction** (right column in Table 3; explicitly provides the identified toxic expressions to guide the model in locating and handling them). To ensure a realistic scenario, we provide the instructions in the language of the input text (obtained via machine translation[8]).

Toxic expressions in instructions We inject toxic expressions into input instructions by leveraging the multilingual comparable lexicon of toxic expressions across all 15 languages. This step presents notable challenges for low-resource languages such as Tatar and Amharic, as well as for languages with distinct orthographic properties, such as Chinese, Arabic, and Hebrew, where simple string matching is insufficient due to the absence of word boundaries or complex morphology. To address this, we implement a language-aware matching tool that assigns a dedicated lookup function to each language. For whitespace-separated languages like English, Spanish, and German, we rely on word-boundary pattern matching. For Cyrillic languages such as Russian and Ukrainian, we also use stem-based matching to cover inflected forms. For French, we apply rule-based conjugation patterns, so verb forms beyond the infinitive are matched as well. For script-specific languages such as Arabic and Hebrew, we normalize text before searching for matches, whereas Chinese and Japanese rely on direct sub-string matching given the absence of word boundaries.

3.2.1. Zero-Shot Prompting

For zero-shot text detoxification experiments, we evaluate two widely used instruction-following LLMs, GPT-3.5-turbo (OpenAI, 2023) and GPT-4.1-mini (OpenAI, 2025), across all 15 languages using the templates presented in Table 3.

3.2.2. Fine-Tuning

Our framework explores two main sequence-to-sequence multilingual models: mT5 (Xue et al., 2021) and Aya-101 (Üstün et al., 2024). We select these models for three reasons. First, mT5 is the standard encoder-decoder baseline in the text detoxification literature (Dementieva et al., 2024a), ensuring direct comparability with prior work. Second, Aya-101 is one of the few open instruction-tuned multilingual models with broad language coverage, including low-resource languages, for which decoder-only alternatives such as LLaMA or Mistral offer more limited support. Third, comparing a base model (mT5) with an instruction-tuned model (Aya-101) allows us to isolate the effect of instruction tuning on the integration of toxic expressions in prompt instructions. This results in three model configurations: **mT5 (Full):** Full-parameter fine-tuning[9], **Aya-101 (Full):** Full-parameter fine-tuning[10], and **Aya-101 (LoRA):** Parameter-efficient fine-tuning via LoRA[11], enabling us to assess whether lightweight adaptation can match or surpass full fine-tuning in

[8] https://translate.google.com, accessed in November 2025.

[9] mT5 fine-tuned for 3 epochs with a learning rate of 3×10^{-5}, and early stopping with a patience of 4 evaluation steps.

[10] Aya-101 fine-tuned for 5 epochs with a learning rate of 2×10^{-4}, and early stopping with a patience of 5 evaluation steps.

[11] LoRA (Aya-101) fine-tuning using rank $r = 16$, $\alpha = 32$, trained for 5 epochs with a learning rate of 2×10^{-4}.

Model	BLEU	ROUGE	STA	SIM	CHRF	Joint
GPT-3.5-turbo w/o	17.40	19.94	0.59	0.56	0.31	0.15
GPT-3.5-turbo w/	27.24	25.08	0.75	0.72	0.52	0.32
GPT-4.1-mini w/o	25.22	25.54	**0.82**	0.74	0.50	0.33
GPT-4.1-mini w/	**38.96**	**30.93**	0.74	**0.85**	**0.63**	**0.41**

Table 4: Average performance of zero-shot prompting (GPT-3.5-turbo and GPT-4.1-mini) with and without toxic expressions in instructions across all 15 languages: *w/o*: instruction-only (Table 3), *w/*: toxic-expressions-in-instruction (Table 3).

this setting. We tune the models on each instruction strategy, resulting in six fine-tuned models in total.

All models are fine-tuned on the training data described in Section 3.1 and evaluated across all 15 languages (9 seen and 6 unseen), with temperature-based sampling ($\tau = 0.7$) to handle language imbalance.

3.3. Evaluation

We evaluate the detoxification performance by directly adapting three core metrics from Dementieva et al. (2024a,b). These metrics return values between 0 and 1 where higher is better.

Style Transfer Accuracy (STA) (Prabhumoye et al., 2018) is computed using a pretrained multilingual toxicity classifier[12] to score both the input and detoxified output, where a higher STA indicates more successful transfer from toxic to non-toxic style.

Content Similarity (SIM) is calculated using cosine similarity between the embeddings of the original toxic text and the generated detoxified text (Feng et al., 2022). It measures how well the original text's meaning is preserved in the output.

Fluency (CHRF) evaluates the fluency of the generated output. For this, an implementation from the *sacrebleu* library is used (Post, 2018).

Joint score (J) is the average of the product of STA, SIM and CHRF, which has been shown to be highly correlated with human evaluation (Logacheva et al., 2022).

Additionally, in line with other studies on text style transfer (Mukherjee et al., 2023b; Jin et al., 2022), we automatically evaluate model outputs using BLEU score (Papineni et al., 2002) and ROUGE[13] score (Lin and Och, 2004).

[12] https://huggingface.co/textdetox/xlmr-large-toxicity-classifier-v2

[13] ROUGE as the mean of ROUGE-1, ROUGE-2, and ROUGE-L

Model	BLEU	ROUGE	STA	SIM	CHRF	Joint
mT5 w/o	58.85	31.55	0.52	0.83	0.58	0.26
mT5 w/	**61.62**	33.05	0.56	0.90	0.66	0.35
Aya-LoRA w/o	49.83	33.93	0.66	0.89	0.69	0.42
Aya-LoRA w/	50.24	**33.99**	**0.71**	0.91	**0.70**	**0.45**
Aya-Full w/o	49.69	33.71	0.63	0.91	0.69	0.41
Aya-Full w/	51.71	33.96	0.70	**0.92**	0.69	0.44

Table 5: Average performance with and without toxic expressions in instructions across all 15 languages. *Aya-LoRA*: LoRA fine-tuning; *Aya-Full*: full-parameter fine-tuning: *w/o*: instruction-only (Table 3), *w/*: toxic-expressions-in-instruction (Table 3).

4. Results

In this section, we present the results from zero-shot prompting and fine-tuning experiments. We then discuss cross-lingual transfer results for the fine-tuned models across seen and unseen languages.

4.1. Zero-Shot Prompting Results

Table 4 presents the average performance of both GPT models across 15 languages. With instruction-only (w/o), GPT-3.5-turbo achieves a Joint score of 0.15, reflecting limited detoxification ability in a purely zero-shot setting, while GPT-4.1-mini performs considerably better at 0.33, likely benefiting from its more recent and capable pretraining. Adding toxic expressions in the instruction (w/) yields substantial gains over the standard strategy (w/o) in both cases. For GPT-3.5-turbo, the Joint score more than doubles from 0.15 to 0.32, with BLEU improving from 17.40 to 27.24. GPT-4.1-mini similarly benefits from the context, with the Joint score rising from 0.33 to 0.41 and BLEU from 25.22 to 38.96. This consistent pattern suggests that, without explicit guidance, it is not obvious to the models what constitutes toxic language based on their pretraining alone—likely due to the filtered nature of their training data (Mendu et al., 2025). Explicitly providing toxic expressions in instructions serves as a strong guiding signal, reducing ambiguity about what the model should remove or replace. A detailed per-language Joint score breakdown is provided in Table 6. Overall, models perform best on high-resource languages such as English, German, French, and Italian under both instruction settings, with a few exceptions, for example, GPT-3.5-turbo w/o performs poorly on Ukrainian. The toxic-expressions-in-instruction strategy brings the most dramatic improvements for GPT-3.5-turbo on these low-scoring languages, with Ukrainian jumping from 0.03 to 0.51 and Hindi from 0.01 to 0.23. GPT-4.1-mini shows more consistent performance across languages, though it similarly benefits from toxic expression guidance, particularly for

Model	EN	RU	UK	DE	AR	ES	HI	ZH	AM	FR	IT	JA	Hin	TT	HE	Avg.
GPT-3.5 w/o	0.37	0.27	0.03	0.34	0.17	0.19	0.01	0.02	0.02	0.33	0.21	0.08	0.04	0.02	0.02	**0.15**
GPT-3.5 w/	0.38	0.47	0.51	0.41	0.46	0.41	0.23	0.16	0.03	0.45	0.49	0.43	0.08	0.06	0.20	**0.32**
GPT-4.1-mini w/o	0.52	0.46	0.54	0.35	0.25	0.45	0.25	0.11	0.22	0.55	0.55	0.28	0.08	0.13	0.20	**0.33**
GPT-4.1-mini w/	**0.53**	**0.54**	**0.62**	**0.55**	**0.48**	**0.51**	**0.28**	**0.21**	**0.25**	**0.61**	**0.64**	**0.46**	**0.11**	**0.17**	**0.23**	**0.41**

Table 6: Joint score for zero-shot prompting across all 15 languages. *w/o*: instruction-only (Table 3); *w/*: toxic-expressions-in-instruction (Table 3). Language codes: EN: English, RU: Russian, UK: Ukrainian, DE: German, AR: Arabic, ES: Spanish, HI: Hindi, ZH: Chinese, AM: Amharic, FR: French, IT: Italian, JA: Japanese, Hin: Hinglish (Hindi in Latin script), TT: Tatar, HE: Hebrew.

	Seen Languages										Unseen Languages						
Model	EN*	RU*	UK*	DE*	AR*	ES*	HI*	ZH*	AM*	Avg.*	FR†	IT†	JA†	Hin†	TT†	HE†	Avg.†
mT5 w/o	0.40	0.37	0.38	0.41	0.39	0.31	0.16	0.07	0.19	**0.30**	0.31	0.34	0.24	0.10	0.12	0.13	**0.21**
mT5 w/	0.52	0.50	0.57	0.52	0.51	0.44	0.22	0.10	0.25	**0.40**	0.42	0.48	0.25	0.11	0.19	0.16	**0.27**
Aya-LoRA w/o	0.55	0.57	0.65	0.59	0.56	0.48	**0.26**	0.11	0.33	**0.46**	0.61	0.62	0.37	0.13	0.25	0.22	**0.37**
Aya-LoRA w/	**0.57**	**0.59**	**0.68**	**0.63**	**0.58**	**0.50**	0.24	**0.13**	**0.37**	**0.48**	**0.65**	**0.63**	**0.41**	**0.15**	**0.30**	**0.26**	**0.40**
Aya-Full w/o	0.54	0.56	0.65	0.60	0.55	0.45	0.24	0.11	0.32	**0.45**	0.57	0.58	0.35	0.13	0.26	0.22	**0.35**
Aya-Full w/	0.55	0.58	0.66	0.62	**0.58**	0.49	**0.26**	**0.13**	0.32	**0.47**	0.63	0.62	0.39	0.15	**0.30**	**0.26**	**0.39**

Table 7: Joint score for fine-tuned models across all 15 languages. $\star$: seen languages; languages were used for fine-tuning. †: unseen languages; not included during fine-tuning. *w/o*: instruction-only (Table 3); *w/*: toxic-expressions-in-instruction (Table 3). *Aya-LoRA*: LoRA fine-tuning; *Aya-Full*: full-parameter fine-tuning. Language codes: EN: English, RU: Russian, UK: Ukrainian, DE: German, AR: Arabic, ES: Spanish, HI: Hindi, ZH: Chinese, AM: Amharic, FR: French, IT: Italian, JA: Japanese, Hin: Hinglish (Hindi in Latin script), TT: Tatar, HE: Hebrew.

Japanese (0.28 $\rightarrow$ 0.46) and French (0.55 $\rightarrow$ 0.61). Chinese (ZH) and Hinglish (Hin) remain the most challenging languages for both models across both settings.

4.2. Fine-Tuning Results

Table 5 presents the average performance of all three fine-tuned models on 15 languages. mT5 w/o achieves a Joint score of 0.26, while both Aya variants perform considerably better at 0.42 (Aya-LoRA w/o) and 0.41 (Aya-Full w/o), reflecting the benefit of instruction-tuned pretraining. Adding toxic expressions in instructions consistently improves performance across all three models. For mT5, adding toxic expressions in instructions yields a noticeable gain in the Joint score (0.26 $\rightarrow$ 0.35), alongside improvements in BLEU, SIM, and CHRF. A similar trend is observed for both Aya-101 variants, where the toxic-expressions-in-instruction prompt improves the Joint score from 0.42 to 0.45 for Aya-LoRA and from 0.41 to 0.44 for Aya-Full. These findings confirm that providing toxic expressions in instructions consistently benefits model performance across all three model configurations, consistent with findings in the zero-shot experiments. Notably, Aya-LoRA outperforms Aya-Full in the instruction-enhanced setting across most metrics, achieving the highest Joint score (0.45) and the best STA and CHRF scores overall. This suggests that parameter-efficient fine-tuning via LoRA not only reduces computational cost but also yields a more effective model for the detoxification task despite using significantly fewer trainable parameters. A detailed breakdown of Joint score across all 15 languages, including performance on unseen languages, is provided in Table 7 and discussed further in Section 4.3.

4.3. Cross-Lingual Transfer Results

Table 7 presents the Joint score of fine-tuned models across seen ($\star$) and unseen (†) languages, revealing two consistent patterns. First, including toxic expressions in the prompt generally improves performance across nearly all settings. The gains vary by model but remain consistent overall, although Aya-LoRA and Aya-Full are exceptions in that they already achieve strong baseline performance with instruction-only prompts. Second, the fine-tuned models generalize remarkably well to unseen languages, with Aya-LoRA still achieving an average Joint score of 0.37–0.40 on unseen languages compared to 0.46–0.48 on seen ones—a marginal performance drop, yet still indicative of meaningful cross-lingual transfer. Looking at individual languages, French and Italian show the strongest performance among unseen languages (0.65 and 0.63 respectively on Joint score) even matching or exceeding the performance of some seen languages. This benefit is likely due to their typological proximity to Spanish and English, which were seen during training. In contrast, lower-resource and more distant languages such as Tatar, Hebrew, and Hinglish show considerably weaker results, suggesting that cross-lingual transfer is largely driven by language family similarity and script overlap.

Overall, Chinese (ZH) and Hinglish (Hin) remain the most challenging languages across all models

Original (Toxic)	Reference	Model Output	Model	Joint
*he is a f*cking parasite.*	*He is a bad person*	*he is a bad person.*	mT5 w/	0.0002
*you are a bit of a d*ck.*	*You are a bit intense*	*you are a bad person*	mT5 w/	0.0001
planet earth is fucking crazy.	*Planet earth is crazy*	*Planet earth is crazy.*	Aya-Full w/	0.0014
they re both fucked up.	*they are both in trouble*	*They are both messed up.*	Aya-Full w/	0.0003
*even bush knows he f*cked up.*	*Bush is well aware that he has messed up*	*Even Bush knows he messed up.*	Aya-LoRA w/	0.0002
liar, liar pants on fire justin.	*Constantly lying Justin*	*justin is lying.*	Aya-LoRA w/	0.0003
*dmvs are un f*cking believable.*	*dmvs are unbelievable*	*dmvs are unbelievable.*	GPT-4.1-mini w/	0.0006
*what an a*se hole you are!*	*What a bad person you are.*	*what a terrible person you are!*	GPT-4.1-mini w/	0.0002

Table 8: Examples of model outputs with low Joint score that may still be considered acceptable detoxifications, illustrating the limitations of automatic evaluation for this task.

and settings. For Chinese, the difficulty likely stems from its logographic script and limited overlap with other languages, which may also affect the quality of toxic span lookup. For Hinglish, the challenge is compounded by the unusual nature of the language itself, Hindi written in Latin script, which may not be well-represented in the pretraining data of any of the models, and whose toxic expressions may not align well with the entries in the toxic expressions lexicon where spelling may vary.

5. Conclusions and Discussion

In this work, we addressed the task of multilingual text detoxification with the objective of automatically transforming toxic text into a non-toxic rewrite without compromising meaning or fluency, making use of comparable toxicity lexicons. Our experiments were structured around two central questions: whether providing toxic expressions in prompt instructions improves detoxification performance, and whether the benefits of toxic expressions in instructions persist in a cross-lingual setup, including for unseen low-resource languages.

With respect to the first question, our results consistently confirm that providing toxic expressions in instructions yields measurable improvements across all models and settings, both in fine-tuning and zero-shot prompting. This holds for mT5, both variants of Aya-101, and the two GPT models, with the effect being particularly pronounced in zero-shot scenarios where no task-specific fine-tuning is available. These findings suggest that explicitly identifying and supplying toxic expressions reduces ambiguity for the model and serves as a reliable guiding signal for the detoxification process. Notably, we observe that a small set of high-frequency profane terms (e.g., *f*ck* in *English*, *p*tain* in *French*, *m*erda* in *Spanish* dominates the matches, while the long tail of the lexicon consists of entries that occur rarely or not at all.

With respect to the second question, the fine-tuned models demonstrate a notable ability to generalize to unseen languages, with a marginal yet meaningful drop in Joint scores compared to seen languages. Aya-LoRA, in particular, still achieves strong performance across both seen and unseen language settings, suggesting that parameter-efficient fine-tuning on multilingual data provides a robust foundation for cross-lingual transfer, even for low-resource languages such as Tatar. Overall, text detoxification remains challenging, particularly when trying to achieve detoxification while preserving content and maintaining fluency at the same time. Future work should incorporate human evaluation, more diverse prompting strategies, and additional low-resource languages.

More broadly, our results underscore the value of multilingual toxic expressions lexicons as a practical, transferable resource for injecting domain-specific knowledge into language models across languages and paradigms.

6. Limitations

One limitation of our work is the reliance on automatic evaluation metrics, which may not fully capture the quality of detoxification outputs. Given the sensitivity of this task to word choice and the degree of toxicity, small lexical changes can disproportionately affect scores such as BLEU or Joint, even when the output is semantically well-detoxified. As illustrated in Table 8, some outputs that received low automatic scores were nonetheless reasonable detoxified sentences upon inspection, highlighting the importance of human evaluation as a complementary assessment. A promising alternative would be LLM-as-a-judge evaluation, which we leave for future work. Furthermore, our detoxification task includes only explicitly toxic comments,

which themselves may have limited coverage, and does not address comments that convey implicit or inherently toxic messages. An example of the latter would be *"f*ck her right in the p*ssy"*, which carries an inherently toxic meaning that persists even when individual words are replaced. Paraphrasing such comments poses a challenge, as thorough detoxification may require altering or removing the original toxic meaning altogether (Dementieva et al., 2024a; Wiegand et al., 2023). Additionally, the use of machine-translated prompt templates across the 15 languages may introduce translation errors, particularly for low-resource languages such as Tatar and Amharic, which could affect model performance independently of the toxic expressions in instructions. Finally, while we experiment with two multilingual models, we acknowledge that larger model sizes and broader language coverage could yield further improvements (Ruan et al., 2024; Kaplan et al., 2020; Brown et al., 2020), which we leave for future work.

7. Ethical Considerations

The aim of our work is to detoxify toxic comments. However, what counts as toxicity is, to some extent, subjective. Automatic detoxification of user-generated comments in online environments should therefore be approached with caution. First, our models cannot guarantee the complete removal of toxicity with 100% accuracy. Second, automatic detoxification might be considered as a violation of freedom of speech (Dementieva et al., 2023, 2025). In line with the propositions made by Dementieva et al. (2025), our intention is to use the detoxification model for the creation of safer online environments and the reduction of harmful content and digital violence—not by enforcing automatic corrections, but rather by offering user-friendly suggestions for rephrasing potentially toxic messages. Finally, although our work is intended to be used for detoxification purposes, we cannot rule out the possibility of misuse, such as generating toxic text from non-toxic inputs (Bose et al., 2023; Floto et al., 2023).

8. Acknowledgments

We would like to acknowledge the reviewers for their helpful comments. We acknowledge the support of the Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg (MWK, Ministry of Science, Research and the Arts Baden-Württemberg under Az. 33-7533-9 19/54/5) in Künstliche Intelligenz & Gesellschaft: Reflecting Intelligent Systems for Diversity, Demography and Democracy (IRIS3D) and the support by the Interchange Forum for Reflecting on Intelligent Systems (IRIS) at the University of Stuttgart. We thank the Institute for Natural Language Processing (IMS), University of Stuttgart, for providing the computational resources used in this work.

9. Bibliographical References

Katherine Atwell, Sabit Hassan, and Malihe Alikhani. 2022. APPDIA: A discourse-aware transformer-based style transfer model for offensive social media conversations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6063–6074, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Ritwik Bose, Ian Perera, and Bonnie Dorr. 2023. Detoxifying online discourse: A guided response generation approach for reducing toxicity in user-generated text. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, pages 9–14, Toronto, Canada. Association for Computational Linguistics.

Peter J Breckheimer. 2001. A haven for hate: The foreign and domestic implications of protecting internet hate speech under the first amendment. *S. Cal. L. Rev.*, 75:1493.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jennifer Cobbe. 2021. Algorithmic censorship by social platforms: Power and resistance. *Philosophy & Technology*, 34(4):739–766.

David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Text detoxification using large pre-trained neural models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7979–7996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Trung Dang and Ferdinando D'Elia. 2025. Gemdetox at textdetox clef 2025: Enhancing a massively multilingual model for text detoxification on low-resource languages.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *International Conference on Web and Social Media*.

Daryna Dementieva, Nikolay Babakov, and Alexander Panchenko. 2024a. MultiParaDetox: Extending text detoxification with parallel data to new languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 124–140, Mexico City, Mexico. Association for Computational Linguistics.

Daryna Dementieva, Nikolay Babakov, Amit Ronen, Abinew Ali Ayele, Naquee Rizwan, Florian Schneider, Xintong Wang, Seid Muhie Yimam, Daniil Moskovskiy, Elisei Stakovskii, Eran Kaufman, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. 2025. Multilingual and explainable text detoxification with parallel corpora. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7998–8025, Abu Dhabi, UAE. Association for Computational Linguistics.

Dementieva, Daryna and Moskovskiy, Daniil and Babakov, Nikolay and Ayele, Abinew Ali and Rizwan, Naquee and Schneider, Frolian and Wang, Xintong and Yimam, Seid Muhie and Ustalov, Dmitry and Stakovskii, Elisei and Smirnova, Alisa and Elnagar, Ashraf and Mukherjee, Animesh and Panchenko, Alexander. 2024b. *Overview of the Multilingual Text Detoxification Task at PAN 2024*. CEUR-WS.org.

Daryna Dementieva, Daniil Moskovskiy, David Dale, and Alexander Panchenko. 2023. Exploring methods for cross-lingual text style transfer: The case of text detoxification. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 1083–1101. Association for Computational Linguistics.

Daryna Dementieva, Sergey Ustyantsev, David Dale, Olga Kozlova, Nikita Semenov, Alexander Panchenko, and Varvara Logacheva. 2021. Crowdsourcing of parallel corpora: the case of style transfer for detoxification. In *Proceedings of the 2nd Crowd Science Workshop: Trust, Ethics, and Excellence in Crowdsourced Data Management at Scale co-located with 47th International Conference on Very Large Data Bases (VLDB 2021 (https://vldb.org/2021/))*, pages 35–49, Copenhagen, Denmark. CEUR Workshop Proceedings.

Ashwin Geet D'Sa, Irina Illina, and Dominique Fohr. 2020. Towards non-toxic landscapes: Automatic toxic comment detection using DNN. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 21–25, Marseille, France. European Language Resources Association (ELRA).

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Griffin Floto, Mohammad Mahdi Abdollah Pour, Parsa Farinneya, Zhenwei Tang, Ali Pesaranghader, Manasa Bharadwaj, and Scott Sanner. 2023. DiffuDetox: A mixed diffusion model for text detoxification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7566–7574, Toronto, Canada. Association for Computational Linguistics.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6786–6794, Marseille, France. European Language Resources Association.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Jared Kaplan, Sam McCandlish, Thomas Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeff Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *ArXiv*, abs/2001.08361.

Nicole Lai-Lopez, Lusha Wang, Su Yuan, and Liza Zhang. 2025. ylmmcl at multilingual text detoxification 2025: Lexicon-guided detoxification and classifier-gated rewriting.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL '04*, ACL '04, pages 605–es. Association for Computational Linguistics.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. ParaDetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818, Dublin, Ireland. Association for Computational Linguistics.

Sai Krishna Mendu, Harish Yenala, Aditi Gulati, Shanu Kumar, and Parag Agrawal. 2025. Towards safer pretraining: Analyzing and filtering harmful content in webscale datasets for responsible llms. In *International Joint Conference on Artificial Intelligence*. Microsoft.

Sourabrata Mukherjee, Akanksha Bansal, Atul Kr. Ojha, John P. McCrae, and Ondrej Dusek. 2023a. Text detoxification as style transfer in English and Hindi. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 133–144, Goa University, Goa, India. NLP Association of India (NLPAI).

Sourabrata Mukherjee, Vojtěch Hudeček, and Ondřej Dušek. 2023b. Polite chatbot: A text style transfer application. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 87–93, Dubrovnik, Croatia. Association for Computational Linguistics.

Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting offensive language on social media with unsupervised text style transfer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 189–194, Melbourne, Australia. Association for Computational Linguistics.

Gopala Krishna Nuthakki, Lekkala Sai Teja, and Atul Mishra. 2025. Team detox at textdetox clef 2025: Multilingual text detoxification using llm. In *Notebook for the PAN Lab at CLEF 2025*, volume 4038 of *CEUR Workshop Proceedings*, Madrid, Spain. CEUR-WS.org.

OpenAI. 2023. GPT-3.5 Turbo (gpt-3.5-turbo) model documentation. https://developers.openai.com/api/docs/models/gpt-3.5-turbo. Accessed: January 2026.

OpenAI. 2025. GPT-4.1 mini (gpt-4.1-mini) model documentation. https://developers.openai.com/api/docs/models/gpt-4.1-mini. Accessed: January 2026.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*. Association for Computational Linguistics.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

Bahar Radfar, K. Shivaram, and Aron Culotta. 2020. Characterizing variation in toxic language by social context. In *International Conference on Web and Social Media*.

Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. 2024. Observational scaling laws and the predictability of language model performance. In *Advances in Neural Information Processing Systems*, volume 37, pages 15841–15892. Curran Associates, Inc.

Elisei Rykov, Konstantin Zaytsev, Ivan Anisimov, and Alexandr Voronin. 2024. Smurfcat at PAN 2024 textdetox: Alignment of multilingual transformers for text detoxification. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024*, volume 3740 of *CEUR Workshop Proceedings*, pages 2866–2871. CEUR-WS.org.

Zheyuan Ryan Shi, Claire Wang, and Fei Fang. 2020. Artificial intelligence for social good: A survey. *ArXiv*, abs/2001.01818.

Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley,

Deepak Kumar, Damon McCoy, Sarah Meiklejohn, Thomas Ristenpart, and Gianluca Stringhini. 2021. Sok: Hate, harassment, and the changing landscape of online abuse. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 247–267.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

P. N. Vasist, Debashis Chatterjee, and Satish Krishnan. 2023. The polarizing impact of political disinformation and hate speech: A cross-country configural narrative. *Information Systems Frontiers*, pages 1–26. Epub ahead of print.

Jeremy Waldron. 2012. *The Harm in Hate Speech*. Harvard University Press.

Michael Wiegand, Jana Kampfmeier, Elisabeth Eder, and Josef Ruppenhofer. 2023. Euphemistic abuse – a new dataset and classification experiments for implicitly abusive language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16280–16297, Singapore. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

Chiyu Zhang, Honglong Cai, Yuezhang Li, Yuexin Wu, Le Hou, and Muhammad Abdul-Mageed. 2024. Distilling text style transfer with self-explanation from LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 200–211, Mexico City, Mexico. Association for Computational Linguistics.

10. Language Resource References

Dementieva, Daryna and Protasov, Vitaly and Babakov, Nikolay and Rizwan, Naquee and Alimova, Ilseyar and Brune, Caroline and Konovalov, Vasily and Muti, Arianna and Liebeskind, Chaya and Litvak, Marina and Nozza, Debora and Shah Khan, Shehryaar and Takeshita, Sotaro and Vanetik, Natalia and Ayele, Abinew Ali and Schneider, Florian and Wang, Xintog and Yimam, Seid Muhie and Elnagar, Ashraf and Mukherjee, Animesh and Panchenko, Alexander. 2025. *Overview of the Multilingual Text Detoxification Task at PAN 2025*. CEUR-WS.org, CEUR Workshop Proceedings.

Author Index