ConLoan: A Contrastive Multilingual Dataset for Evaluating Loanwords

Sina Ahmadi[1,*] **Micha David Hess**[1] **Elena Álvarez Mellado**[2] **Alessia Battisti**[1]
Cui Ding[1] **Anne Göhring**[1] **Yingqiang Gao**[1] **Zifan Jiang**[1] **Andrianos Michail**[1]
Peshmerge Morad[3] **Joel Niklaus**[4] **Maria Christina Panagiotopoulou**[1]
Stefano Perrella[5] **Juri Opitz**[1] **Anastassia Shaitarova**[1] **Rico Sennrich**[1]
[1]Department of Computational Linguistics, University of Zurich
[2]NLP & IR Group, School of Computer Science, UNED [3]University of Münster
[4]University of Bern [5]Sapienza University of Rome
[*]sina.ahmadi@uzh.ch

Abstract

Lexical borrowing, the adoption of words from one language into another, is a ubiquitous linguistic phenomenon influenced by geopolitical, societal, and technological factors. This paper introduces ConLoan–a novel contrastive dataset comprising sentences with and without loanwords across 10 languages. Through systematic evaluation using this dataset, we investigate how state-of-the-art machine translation and language models process loanwords compared to their native alternatives. Our experiments reveal that these systems show systematic preferences for loanwords over native terms and exhibit varying performance across languages. These findings provide valuable insights for developing more linguistically robust NLP systems.

ZurichNLP/ConLoan

1 Introduction

The process of adopting words from one language into another is known as lexical borrowing. Providing significant insights into the social and cultural history of its speakers, the vocabulary of a language is especially prone to borrowing. This phenomenon reflects interactions in specific domains such as technology, politics, religion, or science at particular historical junctures. For instance, the emergence of the widespread usage of commercialized information technology in the 1990s resulted in a flow of related loanwords, a specific type of lexical borrowing, from English into many other languages, as in 'Internet', 'computer' and 'keyboard'. By tracing the spread of loanwords across multiple languages and assessing their level of integration within individual languages, it is possible to historically determine the timing of contact among languages (Grant, 2015). In addition to its importance to historical linguistics, loanwords reflect on the various social and cultural events that the speakers of a language might have faced.

Whether a word is borrowed and adopted in a language by the speakers of a language for need or prestige–the two causalities of borrowing (Carling et al., 2019)–this phenomenon at a larger scope comes with certain implications that have been extensively studied in contact linguistics, such as interference phenomena (Clyne, 2003), changes in linguistic structures (Winford, 2010, p. 175) and language shift (Dorian, 2006). This process can lead to extensive borrowing from the language with a higher social value–the language with a higher "*capacity to be used as a means of communication*" (Sala, 2013, p. 189) in the 'dominated' language as in the Minor Asia Greek influenced by Turkish. As an extreme case, this can also result in creating a mixed language like Michif based on French and Cree (Thomason and Kaufman, 2001, p. 11) or even a potential shift from L1 to L2 due to heavy contact leading to language death (Myers-Scotton, 1992, p. 32). As such, studying this phenomenon is crucial to support language education and preservation, particularly for bilingual and migrant communities where the vocabulary of a minority language is engulfed by a dominant one.

Although there is an extensive literature in contact and historical linguistics (Haspelmath, 2009) along with individual studies on languages such as Arabic (Alhussami, 2020), French (Chesley, 2010) and Austronesian languages (Klamer and Moro, 2023), lexical borrowings have received little attention in computational linguistics and natural language processing (NLP). Furthermore, previous studies focus on loanwords based on word lists and resources such as the World Atlas of Language Structures (WALS) (Haspelmath et al., 2005). To fill the existing gaps in multilingual evaluation of loanwords in context, the current study sheds light on loanwords by creating an annotated dataset containing contrastive sentences in 10 languages, namely Chinese, French, German, Greek, Icelandic, Italian, Northern Kurdish,

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 30070–30090

Portuguese, Russian and Spanish. Using this dataset, we evaluate neural machine translation performance given the contrastive sentence pairs containing loanwords and native equivalents, and analyze how language models' surprisal varies between these alternatives.

2 Background

Lexical borrowing is a linguistic phenomenon that can be broadly categorized into material and structural borrowings (Haspelmath, 2009, p. 38), as depicted in Figure 1. Material borrowings involve the transfer of sound-meaning pairs, i.e. lexemes as in '*best-seller*' in French borrowed from English, while structural borrowings copy syntactic, morphological, or semantic patterns. The latter type can be classified into loan translations, also known as calques, which involve an item-by-item translation of complex lexical units as in English 'loanword' calqued from German '*Lehnwort*' or Kurdish '*da-bezandin*' calqued from English 'download', and meaning extension which occurs when the polysemy pattern of a donor language is copied. For instance, German '*Kopf*' acquiring new meanings based on English 'head' in syntactic phrases.

Loanwords Loanwords are a type of material borrowing, defined as words that entered a language's lexicon as a result of borrowing. They are distinguished from native words, i.e. non-loanwords, in that "we can take [native words] back to the earliest known stages of a language" (Lehmann, 2013, p. 212). The process of lexical borrowing often begins with single-word switches that gradually become conventionalized (Myers-Scotton, 1997) as in '*wesh*' or '*coach*' in nowadays French. In addition to loanwords, there are other types of lexical borrowings such as loanblends which are hybrid borrowings consisting of partly borrowed material and partly native material, e.g., Greek '*σουβλατζής*' (Souvlaki maker) where '*-τζής*' is borrowed from Turkish '*-cı*'.

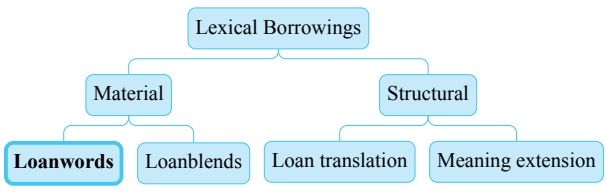

Figure 1: A broad taxonomy of lexical borrowings according to (Haspelmath, 2009). Our focus in this paper is on material loanwords.

It is important to note that the concept of a native word is relative to our knowledge of a language's history. English words like 'disk', 'window', 'bikini' and 'mother', respectively from Greek, Old Norse, Marshallese and Proto-Indo-European, have different etymological backgrounds, some tracing back to other languages or proto-languages. This highlights that while we can identify loanwords, we cannot definitively identify native word.

Prescriptivism & Linguistic Purism The role of loanwords in particular languages has long been a subject of debate. Linguistic purism, a perspective with a radical view on borrowing, considers loanwords as foreign elements that taint the recipient language (Langer and Davies, 2005) and pose a "threat" (Erdem, 2006; Walsh, 2014). This purism often aligns with linguistic prescriptivism, which has historically been associated with ethnonationalist movements, as exemplified by Nazi Germany (Doerr, 2002). A more moderate contemporary instance is the *Académie française*, which aims "*to make [French] pure, eloquent and capable of dealing with the arts and sciences*".[1] However, linguistic purism also manifests in efforts to preserve endangered languages, resisting the influence of dominant languages (Daniel, 2023; Dorian, 1994). This varied application suggests that linguistic purism is not inherently detrimental, but rather a complex phenomenon with diverse motivations and consequences.

3 Related Work

Research on loanwords in computational linguistics has been relatively scarce, with most work focused on related tasks such as parsing (Alex, 2008), automatic speech recognition (Leidig et al., 2014), cognate detection (Rama and List, 2019), code-switching detection (Kent and Claeser, 2019), and transliteration (Ren, 2023). The primary task that has received attention is loanword identification or detection–the task of automatically detecting loanwords within a given text.

Loanword identification has been shown to be challenging and is studied from various aspects in computational linguistics, historical linguistics (Köllner, 2021; Delz, 2013) and corpus linguistics (Chesley and Baayen, 2010). This task

[1] Académie française: https://www.academie-francaise.fr/linstitution/les-missions

has been approached as a classification problem based on various features derived from morphology, phonology, spelling, orthography, and semantics (Ali et al., 2024; Koo, 2015; Mi et al., 2018). Some work has also incorporated features based on optimality theory (Tsvetkov and Dyer, 2016). Recently, neural approaches leveraging monolingual and multilingual word lists have been proposed, utilizing phonological and cognate-based features (Miller et al., 2021; Miller and List, 2023).

A key challenge in this field is the paucity of annotated data, especially for low-resource languages. To address this, some studies have explored data augmentation techniques, such as retrieving loanwords and their original forms using cross-lingual word embeddings (Mi et al., 2020). Efforts have also been made to create annotated datasets, such as WikLoW (Nath et al., 2022), a multilingual dataset of loanwords based on Wiktionary, and corpora of anglicisms in 21,570 newspaper headlines in Spanish (Mellado, 2020). Similarly, unincorporated loanwords are annotated in another corpus (Mellado and Lignos, 2022) containing 370,000 tokens. Consequently, using this resource, detection of loanwords in Spanish was the topic of the ADoBo shared task (Mellado et al., 2021) showing that the task is not trivial and remains an open problem.

Despite these advancements, there are significant gaps in the analysis and study of loanwords in NLP. Most previous work has focused on a limited number of languages, such as Uyghur (Mi et al., 2014), Spanish (Serigos, 2017), Russian (Spektor, 2021), and Turkic languages (Zhang et al., 2021). Additionally, loanwords have been studied in a limited number of tasks where replacing them by native alternative have not received much attention. As such, there is a lack of a multilingual annotated dataset of loanwords in context beyond a word list. Moreover, the impact of loanwords on large language models (LLMs) and machine translation has not been thoroughly investigated. The current study aims to fill these gaps.

4 ConLoan

To address our research questions, we develop ConLoan–a contrastive dataset where loanwords in sentences are replaced by native alternatives, when available. This section details the annotation process and data collection methodology.

4.1 Data Collection

Our primary sources of data are parallel corpora in the selected languages. We refer to the available corpora on OPUS (Tiedemann and Thottingal, 2020) and select one or more corpora that i) contain general-domain sentences rather than a specialized domain as certain domains tend to use more borrowings than others, and ii) have been validated for quality based on previous usage within the research community. In the case of Northern Kurdish, for which no reference translations were available in the corpus, we generate reference translations using Google Translate as the silver standard for evaluation purposes.[2]

To streamline the annotation process and focus on loanword phenomena, we apply the following filters to extract sentences from the parallel corpora:

- Exclude sentences containing named entities in both the source and translation. This is because many named entities, such as locations or organizations, are often borrowed across languages.
- Remove sentences with code-switching. To identify code-switching in languages using non-Latin scripts, we employ GlotScript (Kargaran et al., 2024).
- Include only sentences with at least one occurrence of a loanword. For loanword detection, we primarily rely on Wiktionary, which provides dedicated pages for borrowings in various languages.[3] We select sentences having a token present in the crawled loanword list.

Table A.1 provides detailed information on the specific corpora used and the loanword lists compiled for each language.

4.2 Annotation Setup

The extracted sentences potentially containing loanwords are provided to annotators in spreadsheets with three columns:

Contrastive Sentences This column contains two versions of each sentence:

[2] https://translate.google.com

[3] For example, borrowings in Greek are listed under "Category:Greek borrowed terms": https://en.wiktionary.org/wiki/Category:Greek_borrowed_terms

(A) The original sentence with pre-identified loanwords enclosed in <L></L> tags and highlighted in a distinctive color. For example, the Greek loanword '*ασανσέρ*' ('elevator', borrowed from French '*ascenseur*') in the phrase "... *στο παιδικό ασανσέρ* ..." ("in the children's elevator") is presented as "...*στο παιδικό <L1>ασανσέρ</L1>*..."

(B) The same sentence with tagged loanwords replaced by empty <N></N> tags. Annotators are instructed to fill these tags with native alternatives and apply grammatical changes if need be, e.g., "...*στο παιδικό <N1>ανελκυστήρα</N1>*...".

Both <L> and <N> tags are enumerated to establish a clear correspondence, e.g., <L1> maps to <N1>.

Translations The second column provides translations of the sentences in the first column. Since loanwords and their native alternatives are semantically equivalent, the translation remains the same for both versions. Annotators are required to verify the accuracy of these translations before proceeding with the loanword annotation.

Suggestions The third column offers synonyms to assist in the annotation process. These suggestions are primarily derived from Wordnets (Miller, 1995) available for many of the selected languages through the `wn` package.[4] While these synonyms may not be specifically related to borrowing, they can provide useful alternatives to facilitate the replacement task. Annotators are encouraged to consult additional resources if needed to find appropriate replacements.

Figure A.1 in the appendix provides a visual example of the annotation spreadsheet layout.

4.3 Annotation Task

Given a spreadsheet, the annotator examines each sentence to determine if it contains a loanword. If a loanword is present, the annotator validates the instance by checking the appropriate checkbox. Subsequently, the annotator replaces the pre-identified loanwords with a native alternative. Annotators are also asked to identify and annotate any additional loanwords that have not been previously specified, following the annotation setup. If the annotator is unaware of a native alternative, the loanword is retained in the sentence.

[4] https://github.com/goodmami/wn

To assist in detecting loanwords and finding replacements, the following guidelines are provided:

- **Foreignisms:** These are words with obvious foreign appearance, especially anglicisms, that may feel unfamiliar to native speakers. For example, the verb '*stalké*' in the French sentence "*il a stalké ses voisins*" is borrowed from English 'stalk' and has not been fully integrated into French.

- **Morphology:** If a word is morphologically analyzable in one language but not in another, it likely originates from the first language. For example, German '*Grenze*' (border) is a simple, indivisible form, while its Polish source 'granica' can be decomposed into '*grań*' and '*-ica*', suggesting German borrowed this word from Polish.

- **Phonology:** Words showing signs of phonological integration in one language but not in another are likely borrowed from the latter. For example, English 'facade' retains the French pronunciation pattern from '*façade*', including stress on the final syllable, which is unusual in English.

4.4 Annotation Challenges and Insights

Across the selected languages, annotators encountered several common challenges when identifying and replacing loanwords. A primary issue was distinguishing between fully integrated loanwords and native terms, especially in cases of transliteration or deeply assimilated borrowings. Code-switching and unclear etymologies further complicated the process, particularly for languages like Kurdish and Chinese. Annotators often relied on sentence context, intuition, and resources like dictionaries, etymological references, or online platforms to accurately identify loanwords. In terms of speaker perception, loanwords were generally more accepted in technical, modern, or informal contexts, particularly among younger generations, though resistance persisted in some cultures, especially toward anglicisms. Loanwords from older linguistic influences, like French in German or Italian, were often less recognized as foreignisms.

More information on the annotation task along with some unique characteristics of loanwords in specific languages are provided in Appendix B.

Language	# All	# Validated (%)	# Annotations	# Native replacement (%)	Top Donor
Chinese	1328	639 (48.12)	779	186 (23.88)	English
French	1315	413 (31.41)	442	237 (53.62)	English
German	1861	1162 (62.44)	1282	1095 (85.41)	Latin
Greek	1813	503 (27.74)	515	395 (76.7)	French
Icelandic	4023	2902 (72.14)	3076	1130 (36.74)	Latin
Italian	1780	595 (33.43)	652	589 (90.34)	English
Northern Kurdish	630	401 (63.65)	628	528 (84.08)	English
Portuguese	2509	2336 (93.1)	2568	862 (33.57)	French
Russian	300	298 (99.33)	1051	1039 (98.86)	Latin
Spanish	1311	350 (26.7)	387	287 (74.16)	English
All	16870	9599 (56.9)	11380	6348 (55.78)	English

Table 1: Basic statistics of ConLoan. Number (#) of all sentences potentially containing loanwords based on the compiled loanword list. On average, 56.9% of these sentences were validated by annotators to contain at least one loanword. Of the 11,380 total annotations, 55.78% could be replaced with native alternatives. The most frequent donor language is English across all languages combined.

5 Analysis

5.1 Quantitative Analysis

Table 1 provides basic statistics of ConLoan. Initially provided with 16,870 sentences which potentially contain a loanword based on our loanword lists, the annotators validated 9,599 sentences to truly contain a loanword. In other words, identifying loanwords based on a loanword list lookup resulted in 56.9% accuracy indicating the difficulty of the task without considering the context. Among the 10 languages, Icelandic and Portuguese were provided with the highest number of sentences. Sentences in the same languages along with Russian have been highly validated indicating that the loanword list can be useful to identify loanword in these languages to some extent.

Based on the validated sentences, the annotators then carried out 11,380 replacements among which 6,348 are native replacements, i.e. loanwords not replaced by themselves. That is, almost half of the contrastive sentences in ConLoan are identical. Considering individual languages, loanwords in most of the languages, except Chinese, Icelandic, Portuguese and also French, have been mostly replaced by a native alternative. Although annotators' knowledge directly affects this replacement ratios, it can be also due to the lack of existing widely-known replacements in the language and also, the domain of the corpus from which the sentences are extracted.

Among the annotated instances, we also analyze the distribution of donor languages, with the most frequent donor for each language shown in the last column of Table 1. English emerges as the predominant donor across all languages combined, though individual languages show variations. Latin is the primary donor for German, Icelandic, and Russian, while French contributes most significantly to Greek and Portuguese. It should be noted that this only reflects the frequency of the donor languages in ConLoan rather than in the language as a whole, and may be influenced by both our source corpora and Wiktionary's varying coverage across languages. A detailed breakdown of donor language distributions is provided in Appendix C.

Frequency To assess the relative usage patterns of loanwords and their native alternatives, we analyze their frequencies using normalized token counts collected in `wordfreq` (Speer, 2022)[5], which includes text from Wikipedia. For fusional languages, we exclude inflected forms and employ exact string matching delimited by spaces. Our analysis reveals that native alternatives generally exhibit higher frequency than their loanword counterparts across all languages in our study, except for Chinese. Figure 2 visualizes these frequency distributions through density plots. The pattern of higher native word frequency suggests languages tend to maintain frequently-used native vocabulary while adopting loanwords for specific contexts or semantic niches.

[5]Given that Northern Kurdish was not included in `wordfreq`, we calculated word frequency based on the OSCAR corpus (Ortiz Suárez et al., 2019) over 27M tokens.

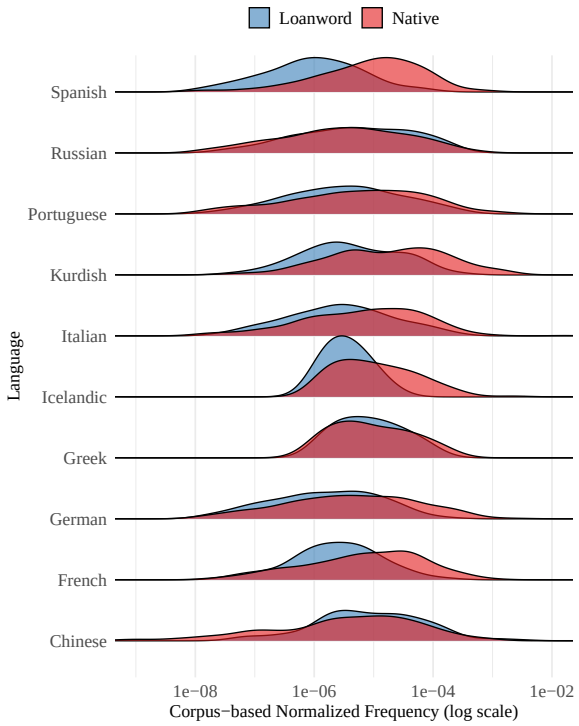

Figure 2: Distribution of normalized frequencies for loanwords and their native alternatives across languages. Patterns highlight that native alternatives generally exhibit higher frequency than loanwords.

5.2 Qualitative Analysis

The following observations and examples offer a qualitative analysis of various linguistic transformations in the annotation task, highlighting the nuanced effects of replacing loanwords with native words. This analysis, mainly based on the annotations in German but also generalizable to the other selected languages, aims to elucidate the impact of such modifications on both the structure and tone of the language.

- **Syntactic Rewording**: The modification of sentences often necessitates adjustments to the sentence structure, varying from subtle tweaks to substantial overhauls. For instance, the transformation of '*restaurieren*' (to restore) to '*wiederherzustellen*' involve a syntactic shift that alters the sentence's rhythm and flow. Similarly, changing "*Isolation des Patienten*" (isolation of the patient) to "*Trennung des Patienten von anderen*" not only simplifies the expression but also shifts its conceptual emphasis.

- **Inelegant Substitutions**: Certain modifications lead to sentences that feel unnatural or clumsy, often due to the replacement of familiar collocations or compounds with less conventional alternatives. For example, substituting '*Einrichtungen*' (facilities) with '*Installationen*' in "*...Einrichtungen installiert...*" results in a redundancy that is rarely used in everyday language. Similarly, replacing '*arzneimittelresistent*' with '*arzneimittelwiderstandsfähig*' and '*Asphalt*' with '*Strassenbelag*' introduces awkwardness making the sentences less fluid and more jarring to native speakers.

- **Academese:** In many instances, substituting loanwords for native terms would result in more formal, academic expressions, or conversely, simplified formulations. For example, replacing '*anstecken*' with '*infizieren*' (to infect) elevates the register of the sentence, aligning it with academic or medical discourse. Likewise, replacing '*auswerten*' and '*schlimm*' respectively by '*evaluieren*' (to evaluate) and '*gravierend*' (serious) illustrates a shift towards more sophisticated or specialized vocabulary, reflecting the tendency of loanwords to lend a more formal tone to the text.

- **Effective Replacements**: The most effective examples of loanword replacement involve using loanwords that enhance the clarity or impact of the text. Substituting '*Team*' for '*Mannschaft*', '*Video game*' for '*Videospiel*' and '*Campingplätze*' (campsites) for '*Zeltplätze*' demonstrates how loanwords can often provide more precise or modern connotations. Such replacements effectively capture the nuances of modern language and usage.

6 Experiments

In addition to analyzing ConLoan, we evaluate how state-of-the-art models handle the distinction between loanwords and their native alternatives. Our experiments focus on contrastive sentence pairs where loanwords were replaced with different native words, excluding cases where loanwords remained unchanged.

6.1 Surprisal

Surprisal is a widely used metric in language modeling that quantifies the unpredictability or information content of a sentence based on a language

model. It reflects how unexpected or unusual a sentence is, with higher surprisal values often indicating either less probable sentences or limitations in the model's ability to anticipate predictable words (Mielke et al., 2019). In the context of our study, we aim to assess the degree to which an LLM generalizes to native sentences compared to counterparts containing loanwords. To that end, we compute sentence-level surprisal of Meta AI's (Touvron et al., 2023) Llama 2.7 model with 7B parameters, Llama 3.1 with 8B parameters (8-bit quantized) and EuroLLM (Martins et al., 2025) with 1.7B parameters by calculating the sum of the negative log-likelihood of the predicted probability distribution as follows:

$$\text{Surprisal}_{\text{sentence}} = -\sum_{i=1}^{t_j} \log p_\theta(x_i \mid x_{<i})$$

where t_j is the number of tokens in the j-th sentence, x_i represents the i-th token in the j-th sentence, and $x_{<i}$ denotes all preceding tokens of x_i.[6] This is unlike perplexity, which is often computed by normalizing the negative log-likelihood by the number of tokens in the corpus and applying exponentiation to obtain a more interpretable scale (Jurafsky and Martin, 2024, p. 40). Our approach avoids normalization by the number of tokens, ensuring that the length of a sentence does not disproportionately influence the comparison, particularly in cases where annotated sentences with native replacements result in longer text.

Figure 3 provides sentence-level surprisal results normalized by the number of sentences per language given the contrastive sentences in ConLoan. Although the surprisal of the model varies to a small extent in the original and annotated sentences, a trend across languages can be seen where the model has a lower surprisal given the loanwords sentences. To assess the significance of surprisal differences, we also conduct a paired t-test revealing significantly higher surprisal scores for sentences with native replacements compared to those with loanwords ($t = -4.029, p < 0.01$). Our analysis confirms the same surprisal pattern in Llama 2.7 with 7 billion parameters and EuroLLM with 1.7B parameters, as shown in Table D.4 in appendix.

We believe that this pattern is due to several factors related to the model's training data and the nature of loanwords. First, many of these loanwords originate from English, which likely serves as a pivot language in the model's multilingual training data. Second, loanwords often appear in specific, well-defined contexts where they have become the conventional choice, making them more predictable for the model in these situations. Third, the widespread use of these loanwords in technical, academic, and professional discourse means they are likely well-represented in the text sources typically used for LLM training. The higher surprisal of native alternatives, despite their greater overall frequency, suggests that these words may appear less natural to the model in contexts where loanwords have become the established norm.

[6] We use natural logarithm unlike Mielke et al. (2019)'s base-2 logarithm.

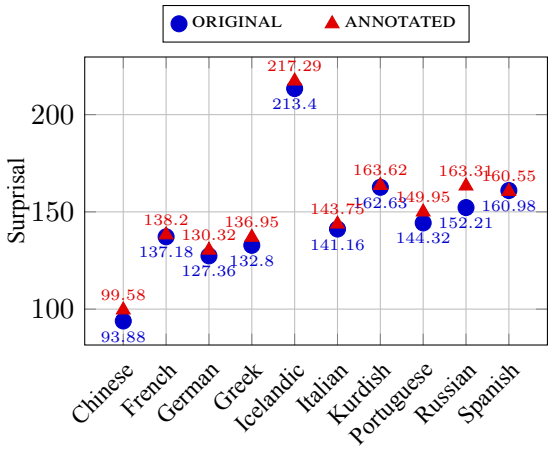

Figure 3: Surprisal (↓) of the Llama 3.1 model (8-bit quantized) for contrastive sentences in ConLoan. Sentences containing annotated native alternatives (ANNOTATED) lead to higher surprisal. This result demonstrates that LLMs find loanwords more predictable than native alternatives, even when the latter are generally more frequent. The same pattern is seen in Llama 2.7 and EuroLLM-1.7B.

6.2 Neural Machine Translation

In this task, we evaluate the performance of neural machine translation (NMT) using these notations:

- $\mathbf{x_{src}}$: Source sentence in the original language containing loanwords.
- $\mathbf{x_{nat}}$: Contrastive annotated sentence in the original language, semantically identical to $\mathbf{x_{src}}$ but containing more native words.
- $\mathbf{t_{src}}$: NMT system output in English when $\mathbf{x_{src}}$ is provided as input.
- $\mathbf{t_{nat}}$: NMT system output in English when $\mathbf{x_{nat}}$ is provided as input.
- $\mathbf{y_{ref}}$: Reference translation of $\mathbf{x_{src}}$ in English from the parallel corpus.

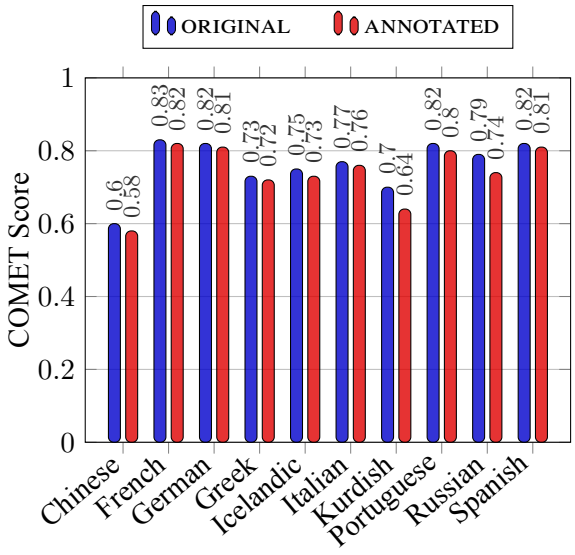

Figure 4: NMT performance in ConLoan using COMET (↑). ORIGINAL refers to sentences containing loanwords, while ANNOTATED denotes sentences with loanwords replaced by native alternatives. Translations of ORIGINAL consistently outperform those of ANNOTATED, , indicating neural MT models are better adapted to loanword usage.

Ideally, a robust NMT system should generate identical outputs $\mathbf{t_{src}}$ and $\mathbf{t_{nat}}$ for both $\mathbf{x_{src}}$ and $\mathbf{x_{nat}}$, regardless of the presence of loanwords or their native alternatives. However, since the source sentences are extracted from parallel corpora, it is expected that an NMT model trained on these corpora would perform better on the original sentences than on our annotated versions.

For our experiment, we employ the No Language Left Behind (NLLB) model (Team et al., 2024) by Meta AI, specifically the `nllb-200-distilled-600M` variant available on HuggingFace. This model, trained on diverse parallel multilingual data, is used for translation into English (X→English). We assess translation quality using COMET (Rei et al., 2020), based on the source and the reference from the parallel corpora along with the hypothesis from NLLB.[7] COMET enables a nuanced evaluation by incorporating human judgment to assess translation quality. We also report NMT performance using BLEU (Papineni et al., 2002)[8], comparing $\mathbf{t_{src}}$ and $\mathbf{t_{nat}}$ to $\mathbf{y_{ref}}$, discussed in more detail in Table D.1.

Our experiments, illustrated in Figure 4, show that NLLB performs less efficiently when translating sentences with native alternatives compared to those with loanwords, showing a consistent drop in COMET scores across all selected languages. These results highlight the sensitivity of NMT systems to loanword usage and replacement. They also align with our surprisal analysis as the consistent pattern across both language modeling and NMT tasks suggests that current neural models may be better tuned to handle loanwords, likely due to their prevalence in multilingual training data and their consistent usage in specific contexts. The notably lower performance on Chinese sentences is likely due to the mixed presence of Cantonese and Mandarin in the parallel corpora. To verify this, we evaluate NLLB with two language indicator tokens: Chinese (Traditional, `zho_Hant`) and Cantonese (`yue_Hant`); both yield almost similar performance.

Reference-free Metrics We further evaluate the impact of loanwords using reference-free MT metrics, which assess translation quality by comparing the translated text directly to the source without requiring reference translations. This approach is particularly suitable as it eliminates potential biases introduced by reference translations containing loanwords. Specifically, we denote $\mathcal{M}(\mathbf{s}, \mathbf{t})$ as the score assigned by metric $\mathcal{M}$ to translation $\mathbf{t}$, given its source $\mathbf{s}$. Using English as the source language and ConLoan languages as targets, we compute the difference between scores assigned to translations with loanwords and native alternatives: $\Delta_{\mathcal{M}} = \mathcal{M}(\mathbf{y_{ref}}, \mathbf{x_{src}}) - \mathcal{M}(\mathbf{y_{ref}}, \mathbf{x_{nat}})$. We experiment with four reference-free metrics:

- CometKiwi (Rei et al., 2022), a metric based on the Info-XLM encoder model (Chi et al., 2021) used as a baseline in WMT (Freitag et al., 2023, 2024);
- CometKiwi-XL (Rei et al., 2023) sharing the same architecture of CometKiwi, but replacing InfoXLM with XLM-R XL (Conneau et al., 2020);
- XCOMET-QE-XL (Guerreiro et al., 2024) based on XLM-R XL and belonging to the XCOMET metric family;
- and MetricX-24-XL (Juraska et al., 2024) based on mT5-XL (Xue et al., 2021) and belonging to the MetricX-24 metric family, which obtained the highest performance at WMT 2024 (Freitag et al., 2024).

All metrics output scores in $[0, 1]$, except MetricX-24-XL which uses $[0, 25]$.

[7] We use `wmt22-comet-da`.

[8] `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.4.2`

Metric	$\Delta_{\mathcal{M}}$ (Loan - Native)
CometKiwi	0.01
CometKiwi-XL	0.03
XCOMET-QE-XL	0.04
MetricX-24-XL	0.27

Table 2: Average difference between metric scores ($\Delta_{\mathcal{M}}$) assigned to translations containing loanwords versus native alternatives using reference-free metrics. Positive values indicate all metrics consistently prefer translations with loanwords over native alternatives.

Table 2 presents the average score differences between translations containing loanwords and their native alternatives across all reference-free metrics. All metrics show positive differences, indicating a consistent preference for translations containing loanwords. This aligns with our previous findings from surprisal and MT experiments, further supporting that current neural models handle loanwords more effectively than their native alternatives. The complete set of metric assessments in Tables D.2 and D.3 show that XCOMET-QE-XL displays the strongest preference for loanwords, favoring them across all target languages, while the other metrics favor native alternatives in certain languages.

7 Conclusion and Discussion

This study sheds light on the complex role of loanwords in context, addressing a significant gap in computational analysis of borrowed lexical items. We introduce ConLoan, a novel contrastive dataset comprising sentences containing loanwords juxtaposed with versions where these loanwords are replaced by native alternatives. Our approach to loanword identification extends beyond historical and etymological considerations, focusing instead on the perceptions of contemporary speakers. This resource holds particular value for investigating the effects of loanwords and their replacement on language technology. Do efforts to replace loanwords with native variants, for example for the purpose of language education and preservation, have inadvertent effects on language modeling and machine translation? Our research demonstrates that NMT systems exhibit varying performance when evaluated on sentences containing loanwords versus their native counterparts. Notably, we observe reduced performance in translating sentences with a higher proportion of native terms. Similarly, our analysis with LLMs, namely Llama 2.7 and 3.1 along with EuroLLM indicates that replacing loanwords with native alternatives leads to higher surprisal on average.

We believe that ConLoan not only serves evaluation purposes but also lays the groundwork for contextual loanword identification and the suggestion of native alternatives. Future research should explore additional dynamics of borrowing, particularly the phenomenon of imposition, where non-native speakers unintentionally retain linguistic features during the borrowing process (Haugen, 1950). Furthermore, investigating other types of corpora, especially those containing oral content where loanwords are more prevalent, represents a promising avenue for future study. Sociolinguistic dynamics of borrowing is a topic that deserves future work of its own as well (Stewart et al., 2021). A key consideration in evaluating ConLoan's scope is that borrowings are a sparse phenomenon; previous literature has pointed out that the presence of borrowings is around 1-2% of most languages (Poplack et al., 1988). That means that, in order to compile annotated datasets that are rich in borrowings, large amounts of text must be available to begin with. With ConLoan we are hoping to create the first borrowing-centered resource of its kind. Finally, we suggest exploring other variants of surprisal and entropy to take other measures such as word frequency into account (Giulianelli et al., 2024; Lin et al., 2025; Ravichander et al., 2025, *inter alia*).

Limitations A primary limitation of this study is the restricted set of languages examined, which constrains the generalizability of our findings to a broader linguistic context. Additionally, low-resource languages, particularly those lacking a standardized variety and spoken in bilingual communities or diaspora settings, present unique challenges that our current work does not fully address. We acknowledge that loanword identification and the suggestion of native replacements are compelling tasks that warrant further investigation in future studies. While we did not conduct our own inter-annotator agreement study, our work builds on literature where similar projects yielded high inter-annotator agreement when native speakers annotate borrowings, with Cohen's kappa > 0.91, suggesting that loanwords are salient and distinct phenomena that produce agreement among native speakers annotations (Mellado and Lignos, 2022).

Ethics Statement Our dataset was built using publicly available material, ensuring adherence to data privacy regulations. While we took all possible measures to remove personally identifiable information and protect the confidentiality and anonymity of individuals, there is a possibility that some sentences might contain sensitive or offensive content. Annotators, apart from the authors of this paper, were appropriately compensated for their work based on the hours contributed. By following these ethical guidelines, we aimed to conduct the study responsibly and thoughtfully.

Acknowledgments

This work was generously supported by the Swiss National Science Foundation (MUTAMUR; no. 213976). Special thanks to Finnur Agust Ingimundarson and Mérilin Sousa Silva for their assistance in the annotation tasks. The authors are grateful to the anonymous reviewers.

References

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The AMARA corpus: Building parallel language resources for the educational domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1856–1862, Reykjavik, Iceland. European Language Resources Association (ELRA).

Sina Ahmadi, Milind Agarwal, and Antonios Anastasopoulos. 2023. PALI: A language identification benchmark for Perso-Arabic scripts. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 78–90, Dubrovnik, Croatia. Association for Computational Linguistics.

Hayder Al-Saedi. 2015. borrowing loanwords from Arabic to Sorani Kurdish. *Misan Journal for Academic Studies*, 14:20–35.

Beatrice Alex. 2008. Comparing corpus-based to web-based lookup techniques for automatic English inclusion detection. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May-1 June 2008, Marrakech, Morocco*, pages 2693–2697. European Language Resources Association (ELRA).

Ahmed Abdullah Alhussami. 2020. *Mutual linguistic borrowing between English and Arabic*. Cambridge Scholars Publishing.

Felermino Dario Mario Ali, Henrique Lopes Cardoso, and Rui Sousa-Silva. 2024. Detecting loanwords in Emakhuwa: An extremely low-resource Bantu language exhibiting significant borrowing from Portuguese. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 4750–4759. ELRA and ICCL.

Marta Bañón, Miquel Esplà-Gomis, Mikel L. Forcada, Cristian García-Romero, Taja Kuzman, Nikola Ljubešić, Rik van Noord, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Peter Rupnik, Vít Suchomel, Antonio Toral, Tobias van der Werff, and Jaume Zaragoza. 2022. MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 303–304, Ghent, Belgium. European Association for Machine Translation.

Starkaður Barkarson and Steinþór Steingrímsson. 2019. Compiling and filtering ParIce: An English-Icelandic parallel corpus. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 140–145, Turku, Finland. Linköping University Electronic Press.

Gerd Carling, Sandra Cronhamn, Robert Farren, Elnur Aliyev, and Johan Frid. 2019. The causality of borrowing: Lexical loans in Eurasian languages. *PloS one*, 14(10):e0223588.

Paula Chesley. 2010. Lexical borrowings in French: Anglicisms as a separate phenomenon. *Journal of French Language Studies*, 20(3):231–251.

Paula Chesley and R. Harald Baayen. 2010. Predicting new words from newer words: Lexical borrowings in French. *Linguistics*, 48(6):1343–1374.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Chinese Academy of Social Sciences. 2022. 现代汉语词典 *(Xiandai Hanyu Cidian) [Modern Chinese Dictionary]*, 7th edition edition. The Commercial Press, Beijing.

Michael G Clyne. 2003. *Dynamics of language contact: English and immigrant languages*. Cambridge University Press.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Rhianwen Daniel. 2023. Standardization and vitality: The role of linguistic purism in preventing extinction. *Language Problems and Language Planning*, 47(2):182–207.

Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. 2012. OpenWordNet-PT: An open Brazilian Wordnet for reasoning. In *Proceedings of COLING 2012: Demonstration Papers*, pages 353–360, Mumbai, India. The COLING 2012 Organizing Committee. Published also as Techreport http://hdl.handle.net/10438/10274.

Dana Delgado, Kevin Walker, Stephanie Strassel, David Graff, and Christopher Caruso. 2024. KASET - Kurmanji and Sorani Kurdish Speech and Transcripts. In *LDC2024S01 Documents*. Philadelphia: Linguistic Data Consortium.

Marisa Delz. 2013. A theoretical approach to automatic loanword detection. *Master's thesis, Eberhard-Karls-Universität Tübingen*.

Karen Doerr. 2002. *Nazi-Deutsch/Nazi German: An English Lexicon of the Language of the Third Reich*. Bloomsbury Publishing USA.

Nancy C Dorian. 1994. Purism vs. compromise in language revitalization and language revival. *Language in society*, 23(4):479–494.

Nancy C Dorian. 2006. Negative borrowing in an indigenouslanguage shift to the dominant national language. *International journal of bilingual education and bilingualism*, 9(5):557–577.

S Defne Erdem. 2006. Is Turkish language facing a threat? *Selçuk Üniversitesi Edebiyat Fakültesi Dergisi*, (15):173–184.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Eleni Galiotou, Giannoula Giannoulopoulou, Maria Grigoriadou, Angela Ralli, Christopher Brewster, Arjiris Arhakis, Evangelos Papakitsos, and Anastasia Pantelidou. 2001. Semantic tests and supporting tools for the Greek Wordnet. In *Proceedings of the NAACL Workshop on WordNet and Other Applications, Carnegie Mellon, Pittsburgh, PA*, pages 183–185.

Mario Giulianelli, Andreas Opedal, and Ryan Cotterell. 2024. Generalized measures of anticipation and responsivity in online language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 11648–11669. Association for Computational Linguistics.

Aitor Gonzalez-Agirre, Egoitz Laparra, German Rigau, et al. 2012. Multilingual Central Repository version 3.0. In *LREC*, pages 2525–2529.

Anthony P. Grant. 2015. Lexical Borrowing. In *The Oxford Handbook of the Word*. Oxford University Press.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Aveen Hasan. 2021. Loanwords and their variations in Kurdish. *Journal of Comparative Studies*, 14:10–27.

Martin Haspelmath. 2009. Lexical borrowing: Concepts and issues. *Loanwords in the world's languages: A comparative handbook*, 35:54.

Martin Haspelmath, Matthew S Dryer, David Gil, and Bernard Comrie. 2005. *The world atlas of language structures*. OUP Oxford.

Einar Haugen. 1950. The analysis of linguistic borrowing. *Language*, 26(2):210–231.

Daniel Jurafsky and James H. Martin. 2024. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Online manuscript released August 20, 2024.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Amir Hossein Kargaran, François Yvon, and Hinrich Schütze. 2024. GlotScript: A resource and tool for

low resource writing system identification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7774–7784, Torino, Italia. ELRA and ICCL.

Samantha Kent and Daniel Claeser. 2019. Incorporating code-switching and borrowing in Dutch-English automatic language detection on Twitter. In *Proceedings of the Future Technologies Conference (FTC) 2018: Volume 1*, pages 418–434. Springer.

Marian Klamer and Francesca R Moro. 2023. Lexical borrowing in Austronesian and Papuan languages: Concepts, methodology and findings. *Traces of contact in the lexicon*, pages 1–21.

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

Marisa Köllner. 2021. *Automatic loanword identification using tree reconciliation*. Ph.D. thesis, Universität Tübingen.

Elias I. Konstantinou. 1992. *Dictionary of Foreign Words in the Greek Language (Λεξικό των ξένων λέξεων στην ελληνική γλώσσα)*. Epikairotita, Athens. University of Crete, Digital Collections–Neoellenistis.

Hahn Koo. 2015. An unsupervised method for identifying loanwords in Korean. *Lang. Resour. Evaluation*, 49(2):355–373.

Andrey Kutuzov and Maria Kunilovskaya. 2014. Russian learner translator corpus: Design, research potential and applications. In *Text, Speech and Dialogue: 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings 17*, pages 315–323. Springer.

Nils Langer and Winifred Davies. 2005. An introduction to linguistic purism. *Linguistic purism in the Germanic languages*, pages 1–17.

Winfred P Lehmann. 2013. *Historical linguistics: An introduction*. Routledge.

Sebastian Leidig, Tim Schlippe, and Tanja Schultz. 2014. Automatic detection of anglicisms for the pronunciation dictionary generation: a case study on our German IT corpus. In *SLTU*, pages 207–214.

Nankai Lin, Peijian Zeng, Weixiong Zheng, Shengyi Jiang, Dong Zhou, and Aimin Yang. 2025. Rethinking vocabulary augmentation: Addressing the challenges of low-resource languages in multilingual models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2919–2934, Abu Dhabi, UAE. Association for Computational Linguistics.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G.C. de Souza, Alexandra Birch, and André F.T. Martins. 2025. EuroLLM: Multilingual language models for Europe. *Procedia Computer Science*, 255:53–62. Proceedings of the Second EuroHPC user day.

Elena Álvarez Mellado. 2020. An annotated corpus of emerging Anglicisms in Spanish newspaper headlines. In *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching, CodeSwitch@LREC 2020, Marseille, France, May, 2020*, pages 1–8. European Language Resources Association.

Elena Álvarez Mellado, Luis Espinosa Anke, Julio Gonzalo Arroyo, Constantine Lignos, and Jordi Porta Zamorano. 2021. Overview of ADoBo 2021: Automatic detection of unassimilated borrowings in the Spanish press. *Proces. del Leng. Natural*, 67:277–285.

Elena Álvarez Mellado and Constantine Lignos. 2022. Detecting unassimilated borrowings in Spanish: An annotated corpus and approaches to modeling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3868–3888. Association for Computational Linguistics.

Chenggang Mi, Lei Xie, and Yanning Zhang. 2020. Loanword identification in low-resource languages with minimal supervision. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 19(3):43:1–43:22.

Chenggang Mi, Yating Yang, Lei Wang, Xiao Li, and Kamali Dalielihan. 2014. Detection of loan words in Uyghur texts. In *Natural Language Processing and Chinese Computing: Third CCF Conference, NLPCC 2014, Shenzhen, China, December 5-9, 2014. Proceedings 3*, pages 103–112. Springer.

Chenggang Mi, Yating Yang, Lei Wang, Xi Zhou, and Tonghai Jiang. 2018. A neural network based model for loanword identification in Uyghur. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

S. J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What kind of language is hard to language-model? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4975–4989. Association for Computational Linguistics.

George A Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

John Miller, Emanuel Pariasca, and Cesar Beltran Castañon. 2021. Neural borrowing detection with monolingual lexical models. In *Proceedings of the Student Research Workshop Associated with RANLP 2021*, pages 109–117, Online. INCOMA Ltd.

John E. Miller and Johann-Mattis List. 2023. Detecting lexical borrowings from dominant languages in multilingual wordlists. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2599–2605, Dubrovnik, Croatia. Association for Computational Linguistics.

Carol Myers-Scotton. 1992. *Codeswitching as a mechanism of deep borrowing, language shift, and language death*, pages 31–58. De Gruyter Mouton, Berlin, Boston.

Carol Myers-Scotton. 1997. *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press.

Abhijnan Nath, Sina Mahdipour Saravani, Ibrahim Khebour, Sheikh Mannan, Zihui Li, and Nikhil Krishnaswamy. 2022. A generalized method for automated multilingual loanword detection. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 4996–5013. International Committee on Computational Linguistics.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

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.

Shana Poplack, David Sankoff, and Christopher Miller. 1988. *The social correlates and linguistic processes of lexical borrowing and assimilation*. Walter de Gruyter, Berlin/New York Berlin, New York.

Taraka Rama and Johann-Mattis List. 2019. An automated framework for fast cognate detection and Bayesian phylogenetic inference in computational historical linguistics. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6225–6235, Florence, Italy. Association for Computational Linguistics.

Abhilasha Ravichander, Jillian Fisher, Taylor Sorensen, Ximing Lu, Maria Antoniak, Bill Yuchen Lin, Niloofar Mireshghallah, Chandra Bhagavatula, and Yejin Choi. 2025. Information-guided identification of training data imprint in (proprietary) large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1962–1978, Albuquerque, New Mexico. Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-Unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yuying Ren. 2023. Back-transliteration of English loanwords in Japanese. In *Proceedings of the Workshop on Computation and Written Language (CAWL 2023)*, pages 43–49, Toronto, Canada. Association for Computational Linguistics.

Eiríkur Rögnvaldsson. 2013. Icewordnet. Technical report, The Árni Magnússon Institute for Icelandic Studies, Reykjavík, Iceland.

Adriana Roventini, Antonietta Alonge, Nicoletta Calzolari, Bernardo Magnini, and Francesca Bertagna. 2000. ItalWordNet: a large semantic database for Italian. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Benoît Sagot and Darja Fišer. 2008. Construction d'un wordnet libre du français à partir de ressources multilingues. In *Actes de la 15ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, pages 171–180, Avignon, France. ATALA.

Marius Sala. 2013. *Contact and borrowing*, page 187–236. Cambridge University Press.

Jacqueline Serigos. 2017. Using distributional semantics in loanword research: A concept-based approach to quantifying semantic specificity of Anglicisms in Spanish. *International Journal of Bilingualism*, 21(5):521–540.

Melanie Siegel and Francis Bond. 2021. OdeNet: Compiling a German wordnet from other resources. In *Proceedings of the 11th Global Wordnet Conference (GWC 2021)*, pages 192–198.

Raivis Skadiņš, Jörg Tiedemann, Roberts Rozis, and Daiga Deksne. 2014. Billions of parallel words for free: Building and using the EU bookshop corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1850–1855, Reykjavik, Iceland. European Language Resources Association (ELRA).

Robyn Speer. 2022. rspeer/wordfreq: v3.0. Online. Software Package.

Yulia Spektor. 2021. Detection and morphological analysis of novel Russian loanwords. *CUNY Academic Works*.

Ian Stewart, Diyi Yang, and Jacob Eisenstein. 2021. *Tuiteamos o pongamos un tuit?* investigating the social constraints of loanword integration in Spanish social media. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 286–297.

NLLB Team et al. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841.

Sarah Grey Thomason and Terrence Kaufman. 2001. *Language contact*. Edinburgh University Press Edinburgh.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Yulia Tsvetkov and Chris Dyer. 2016. Cross-lingual bridges with models of lexical borrowing. *Journal of Artificial Intelligence Research*, 55:63–93.

Olivia Walsh. 2014. '*Les anglicismes polluent la langue française*'. Purist attitudes in France and Quebec. *Journal of French Language Studies*, 24(3):423–449.

Donald Winford. 2010. Contact and borrowing. *The handbook of language contact*, pages 170–187.

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.

Liqin Zhang, Ray Fabri, John Nerbonne, John Nerbonne, EO Aboh, and CB Vigouroux. 2021. Detecting loan words computationally. *Contact Language Library*, 59:269–288.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1. 0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534.

A Annotation Setup

Language	Parallel Corpus	Loanword List	Potential Replacements
Chinese	NLLB[a], WMT19[b] UNPC (Ziemski et al., 2016) QED (Abdelali et al., 2014) NeuLab-TedTalks[c]	Wiktionary	
French	QED (Abdelali et al., 2014) Europarl (Koehn, 2005)	Wiktionary	WOLF (Sagot and Fišer, 2008)
German	QED (Abdelali et al., 2014) Europarl (Koehn, 2005)	Wiktionary	OdeNet (Siegel and Bond, 2021)
Greek	MaCoCu (Bañón et al., 2022) NTREX-128 (Federmann et al., 2022) EUbookshop (Skadiņš et al., 2014) Prime Minister corpus[d]	Wiktionary (Konstantinou, 1992)	Greek Wordnet (Galiotou et al., 2001)
Icelandic	QED (Abdelali et al., 2014) MaCoCu (Bañón et al., 2022) ParIce (Barkarson and Steingrímsson, 2019) Ríkiskaup (Central Public Procurement)	Wiktionary *Íslensk orðsifjabók*[e]	IceWordNet (Rögnvaldsson, 2013)
Italian	QED (Abdelali et al., 2014) Europarl (Koehn, 2005)	Wiktionary	ItalWordNet (Roventini et al., 2000)
Northern Kurdish	KASET (Delgado et al., 2024)	Wiktionary	
Portuguese	QED (Abdelali et al., 2014) Europarl (Koehn, 2005)	Wiktionary	OpenWordNet-PT (de Paiva et al., 2012)
Russian	UNPC (Ziemski et al., 2016) RusLTC (Kutuzov and Kunilovskaya, 2014)	Wiktionary	
Spanish	Europarl (Koehn, 2005)	Wiktionary	Multilingual Central Repository (Gonzalez-Agirre et al., 2012)

[a] https://opus.nlpl.eu/NLLB/corpus/version/NLLB
[b] http://www.statmt.org/wmt19
[c] https://opus.nlpl.eu/NeuLab-TedTalks/corpus/version/NeuLab-TedTalks
[d] https://catalog.elda.org/en-us/repository/browse/ELRA-W0272
[e] https://ordsifjabok.arnastofnun.is

Table A.1: Resources used to create ConLoan. Sentences are extracted from the selected parallel corpora, then filtered based on the loanword lists which are mostly from Wiktionary. The extracted sentences are provided in spreadsheets to annotators, shown in Figure A.1. For each loanword, a synonym is suggested using Wordnet which can possibly be a native alternative.

Source	Target	Replacement Suggestions	Comments?
Όταν έρθει η ώρα να κοιμηθούν, τοποθετήστε τις τσάντες τους στο καρότσι αποσκευών και κυλήστε το *στο παιδικό* **<L1>ασανσέρ</L1>** *που θα* τους μεταφέρει στα δωμάτιά τους.	When it's time to go to sleep, place their bags in the luggage cart and roll it onto the kid-powered elevator to bring them to their rooms.	ανελκυστήρας ασανσέρ	
Όταν έρθει η ώρα να κοιμηθούν, τοποθετήστε τις τσάντες τους στο καρότσι αποσκευών και κυλήστε το *στο παιδικό* **<N1>ανελκυστήρα</N1>** *που θα* τους μεταφέρει στα δωμάτιά τους.	When it's time to go to sleep, place their bags in the luggage cart and roll it onto the kid-powered elevator to bring them to their rooms.		
☑			
Φαίνεται ότι τους διαφεύγει ο *δραματικός σχεδόν* **<L1>συμβολισμός</L1>** της ενέργειας αυτής.	It seems that they fail to grasp the almost dramatic symbolism of this action.	συμβολική αναπαράσταση συμβολισμός	
Φαίνεται ότι τους διαφεύγει ο *δραματικός σχεδόν* **<N1></N1>** της ενέργειας αυτής.	It seems that they fail to grasp the almost dramatic symbolism of this action.		
☐			

Figure A.1: Two instances of contrastive sentences in Greek provided in a spreadsheet for annotation. The annotator's task is to determine if the pre-identified word in the <L></L> tags specified in red is a loanword. If yes, the instance is validated by checking the checkbox and the loanword is manually replaced by a native alternative in the <N></N> tags in blue in the succeeding line. Suggestions are provided to assist in the replacement. In this spreadsheet, the second instance containing *συμβολισμός* (symbolism) not being validated indicates that the annotator has not detected a loanword in the sentence while the first instance is validated. Only validated instances are included in the contrastive dataset. The annotator can leave difficult cases in the comments section for further discussions.

B Lexical Borrowing in the Selected Languages

This section further discusses the annotation task experienced by the annotators of each language. It also briefly provides some of the peculiarities of such words in individual languages when it comes to lexical borrowing.

B.1 Chinese

The most challenging aspect of annotating loanwords was determining their origin, especially since many loanwords in Chinese are transliterations or have become so integrated into the language that their foreign origins are not immediately obvious. Identifying whether a loanword is from English, Japanese, or another language often required careful examination. For example, terms like '咖啡' (coffee) and '可乐' (cola) are transliterations, while others might have been borrowed through Japanese from their original Western sources. Loanwords from Sanskrit, particularly in philosophical contexts, were also noted but less frequent. Identifying the source language and understanding the context in which the word is used added to the complexity of the task.

Sentences were excluded from annotation primarily due to the misalignment in parallel data or presence of named entities or homonyms where the native term resembled the loanword. To identify loanwords, annotators often started by checking the tags but also read the entire sentence to understand the context. Resources such as the Xin Hua Zi Dian (Chinese Academy of Social Sciences, 2022) and various online dictionaries were crucial. In cases where loanwords were not clearly identifiable through tags alone, additional tools like Google and Wikipedia were consulted.

Chinese speakers generally perceive loanwords as neutral or acceptable, given their deep integration into the language. Loanwords from Japanese, in particular, are often so embedded that their origins are not always recognized. For modern concepts and technical terms, borrowings are typically well-accepted.

B.2 French

The annotator highlighted several challenges when working with loanwords in French. The primary difficulty was determining whether a marked word was truly a loanword, ensuring that the replacement was also a French term, and verifying that the substitution preserved the original meaning. Sentences were excluded from annotation when they lacked loanwords or had issues like incorrect translations or mismatched language pairs.

To identify loanwords, the annotator read entire sentences and noted that English was the most common donor language, with occasional words from German, Spanish, and Latin, or Old French via English. Words with Greek etymology or borrowings from sister languages like Italian were not considered loanwords. When replacing loanwords, the annotator relied on intuition, classical French equivalents for anglicisms, e.g., '*courriel*' for 'email', and resources like the *Dictionnaire de l'Académie française*.[9]

Loanwords, particularly anglicisms, were noted to be more frequent in technical conversations, although the annotator was not often in contact with younger French speakers. The perception of borrowings varied, with older generations more likely to view anglicisms as unnecessary and feel that English threatens the French language.

B.3 German

Determining what qualifies as a loanword posed significant challenges in German, with edge cases where the status of a term as a loanword was ambiguous. Also, it was challenging to determine how to replace borrowed acronyms, e.g., 'NATO', and understanding terms that were incomprehensible or where the meaning of the loanword could not be inferred.

For some sentences, comprehension issues or presence of non-relevant content, such as sentences entirely in English, led to their exclusion from annotation. Additionally, sentences with severe grammar mistakes were also excluded. Full sentence review was crucial for accurate annotation. Resources like Wiktionary were extensively used for identifying loanwords and finding replacements. Intuition also played a significant role, especially when evaluating less straightforward cases.

In German, loanwords predominantly originated from French in the past centuries (18th and 19th), while modern loanwords were mostly from English. An example where both phenomena are visible is an instance of the loanword '*riskieren*' (to risk) which stems from the French '*risque*'. It was translated by the annotator with the multi-

[9] https://www.dictionnaire-academie.fr

word phrase '*Gefahr laufen*' as in "*(...) [riskieren wir ständig → laufen wir ständig Gefahr], vor der Welthandelsorganisation verklagt zu werden*". As the annotator noted, the marked prevalence of loanwords from French might be an effect of the domain (parliamentary debates). The pattern showed that French words were often used in academic or upper-class contexts, whereas English and even Arabic loanwords were more common among younger generations and in various informal settings. However, some historical loanwords from Romance languages might not be easily recognized as loanwords today. Perceptions of loanwords vary, with borrowings from Romance languages sometimes seen as posh or snobbish by less educated individuals, while anglicisms may be viewed unfavorably by older generations.

B.4 Greek

Determining whether a term is a loanword or a native Greek word often required thorough investigation into etymologies. Challenges included dealing with unclear or incorrect use of loanwords. Tags were somewhat helpful, but there were instances where they denoted native words instead of loanwords, and thus, the suggested replacements were frequently incorrect.

Loanwords primarily came from French, Italian, and Turkish, often related to inventions or traditions introduced to Greece. In Cyprus, code-switching between Greek and English is very common, both in everyday terms and technical discussions. Proper loanwords have been well-integrated into everyday speech, although there is a cultural divide: philologists may criticize the use of native words for some terms, yet the general population accepts loanwords.

For identifying loanwords, reading the entire sentence and understanding the context was essential. This approach helped to spot both tagged and untagged loanwords. Many modern terms, especially those related to materials and technical fields, had Latin origins, reflecting their widespread use in scientific and technical vocabulary. Greek teaching materials and resources like Wiktionary were crucial for finding appropriate replacements.

Greek speakers are generally accepting of loanwords. While English terms are common among younger people and in technical fields, there is a preference for Greek synonyms when available. The historical incorporation of loanwords, including those from Turkish, has shaped current attitudes, with recent trends showing a conscious effort to use Greek terms when possible, e.g., Greek native word *φιλοξενούμενος* (guest, visitor) instead of *μουσαφίρης* borrowed from Turkish '*misafir*' (itself borrowed from Arabic 'مسافر').

B.5 Icelandic

The most challenging aspect of annotating loanwords in Icelandic was finding native alternatives that were not only accurate but also commonly used. Many of the Icelandic words discovered as replacements were unfamiliar and potentially obscure even to experts, making it more reassuring to find loanwords that had more frequent native equivalents–terms that might be used interchangeably in everyday language.

Sentences were excluded from annotation primarily due to issues with alignment in the parallel data or because native words were homonyms of the loanwords, causing confusion. To identify loanwords, the annotator typically started with the tags and then checked the alignment of the sentences to ensure accuracy. A notable pattern observed was that a significant portion of the data came from EU regulations, which included technical terms related to chemicals, gases, and scientific terms. When replacing loanwords, the annotator relied heavily on two key resources: Snara[10], which provides multiple Icelandic and English dictionaries, and the Icelandic Term Bank[11], covering a broad range of languages and domains.

Loanwords, particularly anglicisms, were more frequently encountered in technical and slang contexts, with younger generations using them more commonly. However, older speakers also used loanwords, especially in technology-related discussions. While there is a strong inclination to create Icelandic equivalents for borrowed terms, their adoption largely depends on their perceived quality and usefulness.

B.6 Italian

The main challenge in annotating loanwords in Italian was determining whether a foreign word had a native equivalent or had become so ingrained in the language that it no longer felt foreign. Words like "click" or "byte" presented difficulties, as native alternatives were either non-existent or extremely

[10] https://snara.is

[11] idord.arnastofnun.is

uncommon, making it hard to decide whether to replace them. In some cases, words were excluded from annotation because they either lacked loanwords or contained foreign terms that were too embedded in Italian to be considered borrowings, such as Latinisms or French-derived words. Sentences with incorrect translations or large chunks of text in English were also skipped.

To identify loanwords, both tagged words and the entire sentences were carefully reviewed. Most loanwords originated from English, particularly in technical or industrial contexts, with a smaller number from French. Borrowings from these languages were common in fields like computer science and technology. For replacements, the annotators frequently relied on Italian dictionaries and the *Treccani encyclopedia*[12] along with Wikipedia.

Loanwords are more common among younger generations and in technical conversations, particularly when referring to products, tools, or actions related to technology and industry. Most of these borrowings are anglicisms. While younger speakers are more comfortable using loanwords, older generations tend to resist them. In fact, in 2022, a law was proposed to limit the use of foreign words in official contexts, aiming to preserve the integrity of the Italian language.[13] Despite this, loanwords are generally perceived positively, though native alternatives are preferred when they are available.

B.7 Kurdish

Lexical borrowing is widely seen among Kurdish speakers. This is chiefly due to the lack of education and support for Kurdish as an official language (Ahmadi et al., 2023). As such, Kurdish languages spoken in everyday life are not immune to extensive usage of borrowed words and terms from dominated languages in the region, namely Arabic, Persian and Turkish, but also English and French (Hasan, 2021; Al-Saedi, 2015). Given that Kurdish is spoken across different countries, the impact of the dominate language of each country is more seen on the variety spoken in the Kurdish regions. As such, Northern Kurdish is widely under the influence of Arabic and Turkish.

Annotators identified two major challenges dealing with the task: i) differentiating between code-switching and loanwords; due to the wide usage of some loanwords without any known native alternatives, it was difficult to draw a line between these two categories and ii) identifying loanwords without a comprehensive linguistic resource. Also, the corpus used for Kurdish–being based on oral utterances–contained many expressions and code-switching into English and sometimes, proverbs and quotations from other languages, making the annotation task particularly interesting but also challenging.

Dealing with loanwords using NLP techniques is an under-explored field of vital importance for Kurdish. Many Kurdish dictionaries lack etymological information. On the other hand, many language enthusiasts who attempt to preserve the language by employing the native alternatives, are often misled by the absence of a word in a dominant language to consider it of native Kurdish origin. For instance, the word for "border" in Kurdish spoken in Iran is 'مەرز' (*merz*) borrowed from Persian '*marz*' while Kurdish speakers in Iraqi Kurdistan use 'سنوور' ('*sinûr*') borrowed from Turkish '*sınır*', itself borrowed from Greek '*σύνορο*'. This lack of knowledge about the etymologies of the words may lead to the latter being considered a native alternative by Kurds in Iran.

B.8 Portuguese

The annotator faced several challenges while identifying and replacing loanwords in Portuguese. A key difficulty was determining when to use a native word, especially for terms like 'bug', where the context–whether referring to an insect or a programming error–determines the choice between a native term ('*inseto*') or a loanword. Another challenge involved native equivalents not fitting perfectly due to subtle differences in meaning, requiring the annotator to retain the loanword in some cases. Sentences were often skipped if they contained encoding issues or non-Portuguese text in the Portuguese column. Annotators found the <L> and <N> tags helpful in identifying loanwords. While suggestions in the spreadsheet were sometimes useful, the annotator primarily relied on Portuguese synonym websites and dictionaries to find the best matches, ensuring that the synonym captured the same meaning as the loanword.

Loanwords, especially anglicisms, are most common in topics related to electronics, modern services, and media, and are predominantly used by younger generations. French loanwords, on the other hand, have been ingrained in the language for a longer time and are used across all

[12] https://www.treccani.it/enciclopedia

[13] https://www.camera.it/leg19/126?tab=&leg=19&idDocumento=734&sede=&tipo=

age groups without much awareness. The annotator noticed that Brazilian Portuguese speakers tend to use more loanwords, particularly anglicisms, likely due to Brazil's proximity to the U.S. In contrast, European Portuguese speakers may struggle with or resist the influx of newer loanwords. Both variants of Portuguese, however, incorporate a significant amount of lexical borrowing, especially among speakers living abroad.

B.9 Russian

The annotator highlighted several key challenges in identifying and handling loanwords in Russian. One of the most difficult tasks involved deciding how to manage terms without direct Russian equivalents, particularly lexical borrowings introduced alongside new concepts. Sentences were often excluded from annotation if they lacked loanwords or if the borrowed term had a homograph in Russian, such as '*лук*' /luk/ which can refer to the borrowed term "look" or the native Russian word for 'onion'. The annotator relied heavily on their knowledge of other languages to identify borrowings, noting that borrowings were more frequent in topics like politics, technology, and modern culture.

Resources like Wikipedia, online dictionaries, and even AI assistants proved valuable in finding suitable replacements for loanwords. It might also be the case that Russian speakers are largely unaware of the extent of foreign influence on their language, as some borrowed terms have become so ingrained that their non-Slavic origins are forgotten. Political and scientific terms, especially from Latin and Greek, often require higher education to be fully understood. Finally, the annotator reflected on how borrowing in Russian has shifted, particularly with the 2023 government ban on foreign words in official contexts when Russian equivalents exist.

B.10 Spanish

The annotator faced several challenges when deciding whether a word should be considered a borrowing in Spanish. They highlighted the difficulty of distinguishing true loanwords from words that, while having foreign origins, have been fully assimilated into the language over time, such as '*batuta*' (baton) borrowed from Italian. Additionally, they grappled with cases of code-switching, quotations, and proper names, which complicated the annotation process. Sentences were often excluded if the word resembled a foreign term but was native to Spanish, e.g., '*he*', '*social*', or if it involved an English inclusion that didn't qualify as a loanword.

To identify loanwords, the annotator read the entire sentence but focused primarily on the tagged words. English was the most frequent donor language, especially in legal and international policy terms found in parliamentary texts from the European Union. Given the multilingual context of these documents, code-switching and non-loanword inclusions were also common. When replacing a loanword, the annotator relied on their native speaker intuition and consulted resources like Fundéu[14], Wikipedia, and the *Diccionario de la Lengua Española*[15], rather than using the suggested synonyms.

In everyday Spanish, loanwords–especially anglicisms–are most prevalent in conversations about technology, specialized journalism, and areas influenced by American and international culture, such as science and social media. While anglicisms are often seen as prestigious, particularly in professional and marketing contexts, they can be perceived as pretentious in casual conversation. Finally, the annotator noted that many sentences contained words that appeared foreign but were actually native or cognates, and these were excluded from annotation.

C Distribution of Donor Languages

We analyze the donor languages of loanwords based on the annotated sentences in ConLoan, regardless of whether they have been replaced by a native alternative or not. Figure C.1 presents this distribution, showing the four most frequent donor languages for each recipient language, with the remaining donor languages aggregated as "Others" for clarity.

To identify the donor language of each loanword, we rely on the etymological resources listed in Table A.1. While Wiktionary serves as our primary source, its collaborative nature means that coverage and accuracy may vary across different languages and entries. It is important to note that our findings reflect only the distribution of donor languages within ConLoan and should not be generalized to represent the overall loanword composition of these languages. For instance, while French appears as the most frequent donor language for

[14] https://www.fundeu.es
[15] https://dle.rae.es

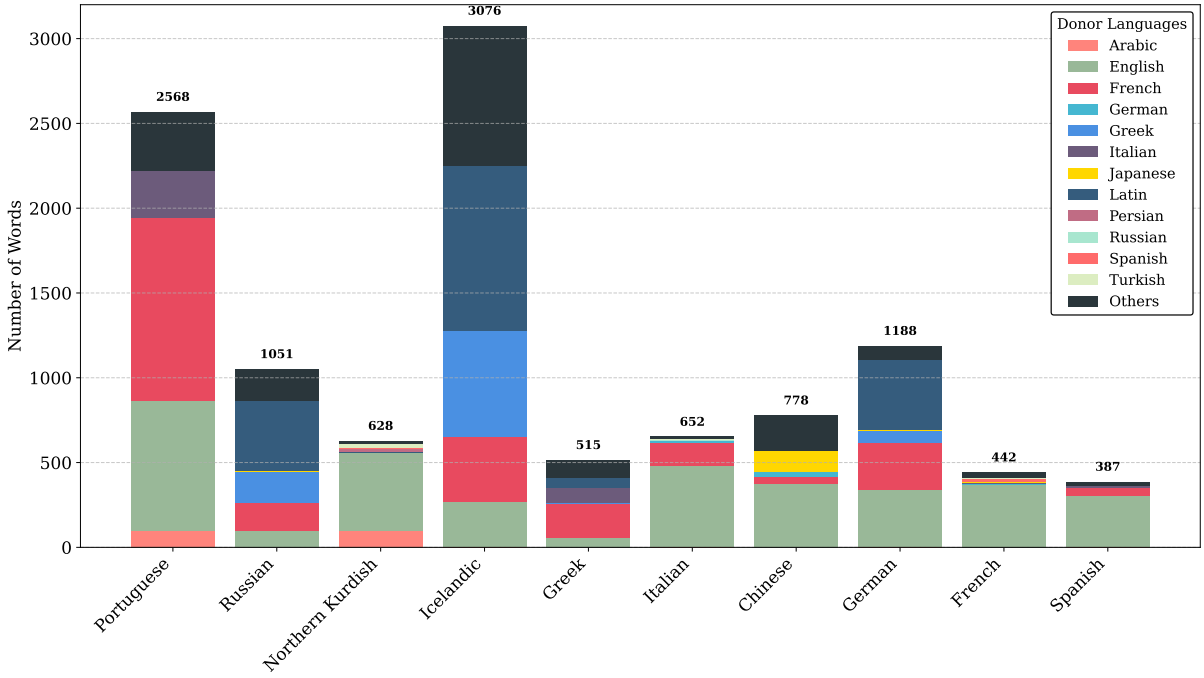

Figure C.1: Distribution of donor languages per recipient language in ConLoan. The stacked bars represent the four most frequent donor languages for each recipient language, with remaining donors combined as "Others". Numbers above bars indicate total loanword annotations per language.

Modern Greek in our dataset, previous research has shown that Italian, Turkish, French, Latin, and English are, in descending order, the most common sources of loanwords in Modern Greek (Konstantinou, 1992). Such differences likely stem from the specific domain and time period of our source corpora, as well as potential biases in our resource coverage.

D Supplementary Experimental Results

In addition to COMET results for NMT performance, we report metric assessments using BLEU in Table D.1 along with reference-free metrics in Tables D.2 and D.3.

There is a clear tendency for the metrics to empirically favor translations containing loanwords (ORIGINAL) compared to the ones with more native replacements (ANNOTATED). XCOMET-QE-XL assigns higher scores to translations containing loanwords in all the target languages, showing the strongest preference for loanwords among the tested metrics. Similarly, CometKiwi, CometKiwi-XL, and MetricX-24-XL exhibit a preference for loanwords but assign higher scores to native alternatives in certain languages. Specifically, all three metrics favor native alternatives in Spanish. This aligns with the surprisal results reported in Section 6.1. Interestingly, unlike the other metrics, MetricX-24-XL also assigns higher scores to translations with native alternatives in French, Italian, and Northern Kurdish. We hypothesize that this behavior might be due to MetricX-24-XL being fine-tuned with synthetic data in addition to data containing human judgments, which might enhance its robustness to modifications that preserve the meaning of translations, such as replacing loanwords with their native alternatives.

Language	BLEU ↑		COMET ↑	
	ORIGINAL	ANNOTATED	ORIGINAL	ANNOTATED
Chinese	10.5	9.7	0.6	0.58
French	34.6	33.3	0.83	0.82
German	30.6	29.2	0.82	0.81
Greek	22.1	20.9	0.73	0.72
Icelandic	25.1	23.4	0.75	0.73
Italian	28.7	26.8	0.77	0.76
Northern Kurdish	27.1	22.9	0.7	0.64
Portuguese	35.3	33.2	0.82	0.8
Russian	25	19.4	0.79	0.74
Spanish	33.1	31.2	0.82	0.81

Table D.1: Evaluation results of NMT based on BLEU and COMET based on sentences with loanwords (ORIGINAL) and annotated ones containing more native alternatives (ANNOTATED). Results show consistent performance drop for sentences with native alternatives, indicating lower translation efficiency for ANNOTATED.

Language	CometKiwi ↑		CometKiwi-XL ↑		XCOMET-QE-XL ↑		MetricX-24-XL ↓	
	ORIGINAL	ANNOTATED	Original	ANNOTATED	ORIGINAL	ANNOTATED	ORIGINAL	ANNOTATED
Chinese	0.51	0.50	0.31	0.28	0.41	0.40	10.38	10.71
French	0.81	0.81	0.65	0.65	0.80	0.78	4.04	3.71
German	0.77	0.76	0.65	0.63	0.87	0.86	2.32	2.46
Greek	0.68	0.67	0.52	0.50	0.63	0.61	7.41	7.91
Icelandic	0.75	0.73	0.68	0.66	0.77	0.73	5.11	5.57
Italian	0.74	0.74	0.61	0.60	0.74	0.72	5.50	5.45
Northern Kurdish	0.36	0.36	0.27	0.25	0.33	0.33	15.39	15.22
Portuguese	0.75	0.73	0.64	0.60	0.78	0.73	5.32	5.98
Russian	0.77	0.71	0.66	0.54	0.79	0.65	4.28	6.14
Spanish	0.80	0.82	0.70	0.71	0.82	0.81	4.47	3.77

Table D.2: Average reference-free metric scores for translations containing loanwords (ORIGINAL) versus their native alternatives (ANNOTATED). For each metric-language pair, ORIGINAL shows the average scores for translations containing loanwords ($\mathcal{M}(\mathbf{y_{ref}}, \mathbf{x_{loan}})$) and ANNOTATED shows the average scores for translations with native alternatives ($\mathcal{M}(\mathbf{y_{ref}}, \mathbf{x_{nat}})$).

Language	CometKiwi	CometKiwi-XL	XCOMET-QE-XL	MetricX-24-XL
Chinese	0.01	0.03	0.02	0.33
French	0.00	0.00	0.02	−0.33
German	0.01	0.02	0.01	0.15
Greek	0.01	0.01	0.03	0.51
Icelandic	0.02	0.02	0.04	0.46
Italian	0.01	0.01	0.02	−0.05
Northern Kurdish	0.00	0.02	0.01	−0.17
Portuguese	0.02	0.04	0.05	0.66
Russian	0.06	0.12	0.14	1.85
Spanish	−0.02	−0.02	0.01	−0.70

Table D.3: Language-specific score differences ($\Delta_{\mathcal{M}}$) showing metric preferences between loanword and native translations. Computed as $\mathcal{M}(\mathbf{y_{ref}}, \mathbf{x_{loan}}) - \mathcal{M}(\mathbf{y_{ref}}, \mathbf{x_{nat}})$, where positive values indicate preference for loanwords.

Language	Llama 2.7 (7B)		Llama 3.1 (8B)		EuroLLM (1.7B)	
	Original	Annotated	Original	Annotated	Original	Annotated
Chinese	97.45	102.92	93.88	99.58	93.46	98.73
French	142.11	143.14	137.18	138.20	133.95	134.51
German	133.79	136.45	127.36	130.32	125.23	128.25
Greek	186.02	187.52	132.80	136.95	123.04	127.31
Icelandic	281.69	285.73	213.40	217.29	356.53	361.26
Italian	151.46	153.79	141.16	143.75	138.15	140.64
Northern Kurdish	189.48	197.31	162.63	163.62	204.12	214.21
Portuguese	154.84	160.09	144.32	149.95	139.15	144.95
Russian	162.00	171.12	152.21	163.31	155.34	162.34
Spanish	169.20	168.78	160.98	160.55	157.54	155.95
Average	**166.80**	170.68	**146.59**	150.35	**162.65**	166.82

Table D.4: Surprisal (↓) of loanwords and annotated native equivalents in ConLoan using Llama 2.7, Llama 3.1 and EuroLLM. Across all models, sentences containing more native alternatives (ANNOTATED) lead to higher surprisal, specified in bold face.