Definition Generation for Word Meaning Modeling: Monolingual, Multilingual, and Cross-Lingual Perspectives

Francesco Periti
KU Leuven - Flanders Make
francesco.periti@kuleuven.be

Roksana Goworek
Queen Mary University of London
r.goworek@qmul.ac.uk

Haim Dubossarsky
Queen Mary University of London
h.dubossarsky@qmul.ac.uk

Nina Tahmasebi
University of Gothenburg
nina.tahmasebi@gu.se

Abstract

The task of Definition Generation has recently gained attention as an interpretable approach to modeling word meaning. Thus far, most research has been conducted in English, with limited work and resources for other languages. In this work, we expand Definition Generation beyond English to a suite of 22 languages and evaluate Llama-based models within a monolingual, multilingual, and cross-lingual setting. Our experiments show that monolingual fine-tuning consistently outperforms pretrained baselines, with the largest gains observed in languages with lower initial performance; and that multilingual fine-tuning does not consistently improve performance on the individual fine-tuning languages. Our cross-lingual evaluation reveals that models fine-tuned on a single language typically lose the ability to generate definitions in other languages, whereas multilingual models exhibit robust generalization even to languages unseen during fine-tuning.

1 Introduction

Recent advances in text generation have opened up new opportunities for modeling word meaning. Moving beyond the traditional reliance on word embeddings (Camacho-Collados and Pilehvar, 2018), the research community is now shifting toward an interpretable approach to modeling word meanings through the Definition Generation task (Periti et al., 2024). The premise underlying this shift is that generative language models can serve as interpreters of meaning, providing roughly equivalent sense definitions for word occurrences that share the same meaning across different contexts.

The task of Definition Generation is as follows:

> Given a target word w and an example usage e, the goal is to generate a natural language definition d that is grammatical, fluent, and faithful to the meaning of the target word w as used in the example usage e.

Until now, as with other Natural Language Processing tasks, Definition Generation has exhibited a strong bias towards English, with only limited research, benchmarks, and models available for other languages (Fedorova et al., 2024b).

In this work, we address this multilingual gap by extending research on word meaning modeling through Definition Generation to a suite of 22 different languages.

We start by evaluating Large Language Models (LLMs) in a **monolingual** setting, fine-tuning and testing them on the same language. In our experiments, we focus on Llama models as a case study, but the analysis method can be applied to other models. Our results show that fine-tuned models consistently outperform their pre-trained baselines, and that languages with lower baseline performance benefit the most from monolingual fine-tuning, regardless of fine-tuning data size.

We then explore how **multilingual** fine-tuning across multiple languages simultaneously can enhance performance over the individual training languages. Our evaluation indicates that fine-tuning models on multiple languages does not consistently improve performance on the individual fine-tuning languages. Notably, Slavic languages tend to be negatively affected by cross-family fine-tuning, whereas Germanic languages benefit from greater language diversity.

Finally, we examine the **cross-lingual transfer** capabilities of fine-tuned models on languages not encountered during fine-tuning, offering valuable insights for low-resource and cross-lingual scenarios. We observe that models fine-tuned on a single language often lose the ability to generate definitions in other languages. In contrast, models fine-tuned on a multiple language consistently generate definitions in the target language, even for languages not seen during fine-tuning.

By expanding Definition Generation to multiple languages, we foster innovation across diverse

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 26004–26024

research communities and unlock concrete applications across several fields. For example, in lexicography, it can automate the drafting of dictionary entries, offering linguists editable definitions that reflect contemporary usage beyond manual corpus analysis (Barrios et al., 2009). In language learning, generated definitions offer learners simple, accessible explanations tailored to their target language, which can potentially be adjusted for proficiency levels (Yuan et al., 2022). In sociolinguistic research, Definition Generation enables the study of regional, social, or diachronic variation in word meanings across different communities or time periods (Giulianelli et al., 2023).

2 Experimental setup

Data. Inspired by prior work on Definition Modeling (Kabiri and Cook, 2020), we adopt Wiktionary as the primary resource for Definition Generation in our work. Wiktionary is a web-based collaborative project led by the Wikimedia Foundation, aimed at creating a free-content dictionary across multiple languages. In particular, we utilize Dbnary (Sérasset, 2015; Sérasset, 2012), a structured linguistic resource derived from Wiktionary, as it provides a pre-filtered and structured dataset that facilitates the collection of training and evaluation data across different languages.

Dbnary provides separate data dumps for each language edition of Wiktionary it extracts from. Each language's data is available as a distinct Turtle file, stored using the OntoLex model (McCrae et al., 2017), and can be downloaded individually. In our work, we then used SPARQL to collect data in a format compatible with existing Definition Generation datasets for all available languages. Specifically, each data entry consists of a target word w, an example usage e, a sense definition d. Appendix C provides an example of the dataset structure.

We filtered out all instances where any of the three components were missing. For each language, we randomly split the dataset as follows: 75% for training (**Train**), 5% for validation (**Dev**), and 20% for evaluation (**Test**). The Test sets was further divided into two subsets: **Seen Test**, contains examples e (distinct from those in training) for target-definition (w-d) pairs that were already present in the training data; **Unseen Test**, contains examples for target words that did not appear during training, ensuring an evaluation setting that assesses the model's ability to generalize to novel words and word meanings (Fedorova et al., 2024b).[1]

Dbnary is updated each time a new Wiktionary dump is made available by the Wikimedia foundation. To ensure reproducibility, we release our Train, Dev, and Test sets for each language on **Hugging Face**. Additionally, we make publicly available the code used for downloading and processing Dbnary in **GitHub** allowing future research to extend our dataset with additional examples and languages.[2]

Models. LLMs are currently at the forefront of text generation, with an increasing number of models being developed and made publicly available. Selecting which models to evaluate for Definition Generation across multiple languages is a key challenge, particularly because the number of evaluations is proportional with both the number of selected models and the number of languages tested.

In our work, we focus on decoder-only transformer models from the open-weight Llama family (Meta, 2024, 2023). In preliminary evaluations, we explored models pre-trained or fine-tuned on individual languages, but they consistently underperformed compared to Llama models (see Appendix G). As a result, we opted to proceed exclusively with Llama models, as recent experiments have revealed that Llama possesses impressive multilingual capabilities (Yuan et al., 2024). In particular, we focus on the chat versions, which have already been optimized to generate responses that adhere to specific instruction prompts.

To adapt Llama for Definition Generation, we performed multiple rounds of fine-tuning, which we describe in the following section. Regardless of the specific **Train** data used, we adopted parameter-efficient fine-tuning, implemented as Low-Rank Adaptation (LoRA) (Hu et al., 2022), to reduce computational and storage costs. Following Periti et al. (2024), we fine-tuned Llama2Chat[3] and Llama3Instruct[4] models.

The fine-tuning process was performed using cross-entropy loss, computed over all tokens for 50 epochs. We employed a batch size of 40, a maximum sequence length of 300 tokens, and sequence *packing* (Kosec et al., 2021) to optimize training efficiency by processing multiple samples simultaneously. To prevent overfitting, we applied early

[1] Like **Unseen Test**, **Dev** contains examples for target words that did not appear in **Train**. [2] Please find our models and datasets on **Hugging Face**, the code on **GitHub**, and the model-generated outputs on **Zenodo**. [3] *Llama-2-7b-chat-hf*
[4] *Meta-Llama-3-8B-Instruct*

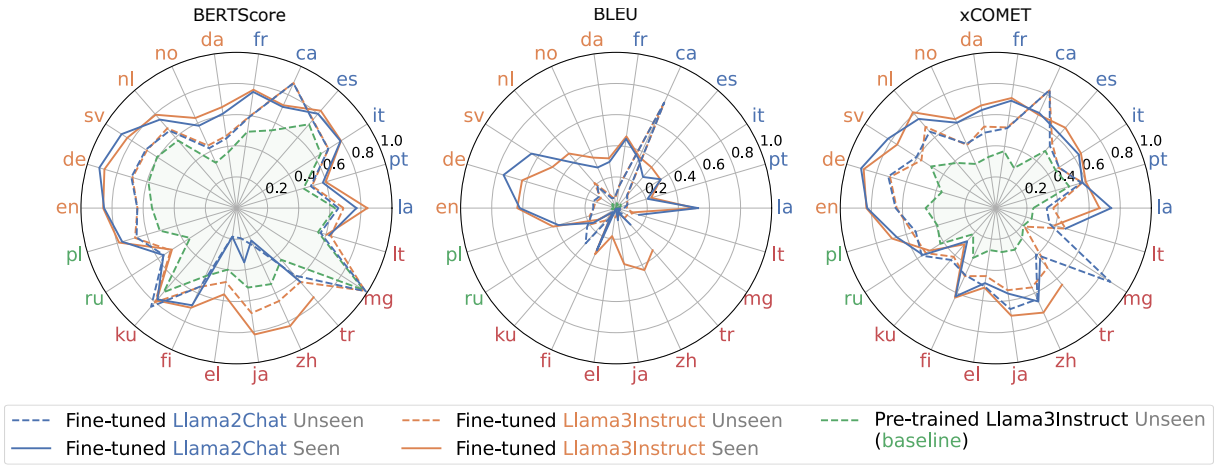

Figure 1: BERTScore, BLEU, and xCOMET scores for the monolingual setting. We report the performance of the fine-tuned Llama2Chat and Llama3Instruct models on both **Seen** and **Unseen Test** sets. For the baseline (pre-trained Llama3Instruct), performance is reported only on the **Unseen Test** set, as no fine-tuning was conducted. For Malagasy, no **Seen Test** set is available.

stopping based on ROUGE-L (Lin, 2004), computed on the Dev set. Instruction prompts for the sets of models and further fine-tuning details are provided in Appendix F.

Evaluation. To assess the performance of our models, we use a range of standard Natural Language Generation metrics. Specifically, we apply both lexical overlap and semantic similarity metrics to evaluate the quality of the generated definitions in comparison to the reference Dbnary definitions in both the **Seen** and **Unseen Tests**.

As for lexical overlap metrics, we use BLEU (Papineni et al., 2002), NIST (Doddington, 2002), SacreBLEU (Post, 2018), ROUGE-L (Lin, 2004), and Exact Match. For semantic metrics, we use BERTScore (Zhang et al., 2020) and xCOMET (Guerreiro et al., 2024).

Our analysis in Figure 5 indicates strong correlations among these metrics, making it redundant to report all of them in the main text. Therefore, we only report BLEU (lexical overlap-based, distinct from ROUGE-L used in fine-tuning), BERTScore (which measures semantic similarity by leveraging contextualized embeddings from pre-trained transformer models), and xCOMET (a learned metric for cross-lingual and semantic similarity, which enables semantic comparisons in both monolingual and cross-lingual evaluations). Full results, including additional metrics, are provided in Table 6.

3 Monolingual Definition Generation

For each language considered, we fine-tuned a different *monolingual* LoRA adapter on the **Train** set of that specific language. We then evaluated each model on the **Seen** and **Unseen Test** of the same language and compared its performance to the pre-trained Llama3Instruct model, considered as the baseline. For the pre-trained Llama2Chat, the generated responses were empty for all the languages considered and thus, instead of prompt engineering, we excluded it as a baseline.

Figure 1 summarizes the evaluation using BLEU, BERTScore, and xCOMET. In the discussion, we primarily focus on BERTScore, offering only general observations on the other metrics and leaving their detailed interpretation to the reader. Additional results for fine-tuned models are presented in Table 6, and the complete set of results is provided as supplementary material.

In general, and consistent with prior work, we obtained medium to high performance (depending on the language) on semantic measures (BERTScore and xCOMET), suggesting good quality in the generated definitions. In contrast, the lexical overlap measure (BLEU) yielded lower scores, indicating that the models often used different wording compared to the reference definitions, which, however, is not necessarily a negative outcome (see an example in Appendix A).

Pre-trained Llama3Instruct. As expected, for Llama3Instruct that serves as our baseline, we observed high performance in BERTScore on Germanic languages such as English (0.543) and German (0.587), as well as lower performance on some low-resource languages like Greek (0.399). Unexpectedly, we observed performance drops between

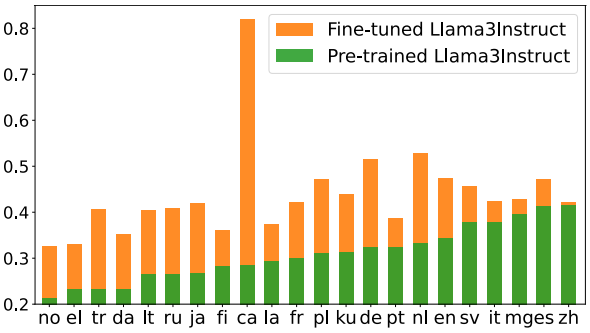

Figure 2: Avg. performance (BERTScore, BLEU, xCOMET) of the fine-tuned and pre-trained (baseline) Llama3Instruct model, sorted by baseline performance. Improvements from fine-tuning over **Unseen Tests** tends to be smaller for languages where the baseline performs better.

linguistically similar languages – for instance, on Norwegian (0.320) and Danish (0.364), compared to Swedish (0.607); and on Portuguese (0.449), compared to Spanish (0.712). The gap between the latter pair is notably reduced when assessed using the xCOMET metric (0.492 vs. 0.486), suggesting that a more comprehensive evaluation benefits from incorporating multiple measures.

We also observed unexpectedly high performance for Kurdish (0.714) and Malagasy (0.972). However, for Malagasy, we attribute this to the low quality of the data compared to other languages, with most definitions being similar to each other, even for different targets.

BLEU scores are generally much lower across all languages, with values close to zero. We attribute this, in part, to the pre-trained model's verbosity when generating responses. For instance, outputs across languages often begin with phrases such as *The word {TARGET} in this context refers to ...*, rather than providing concise definitions.

In contrast, xCOMET scores aligned more closely with our expectations. For this metric, we observed higher scores for widely spoken languages such as English (0.451), and lower for low-resource languages such as Latin (0.241), as well as for languages that use a different writing system, such as Japanese (0.285).

Fine-tuned Llama2Chat and Llama3Instruct. Fine-tuned models consistently outperform the pre-trained baseline across all languages, demonstrating a good ability to generalize to unseen words (**Unseen Test**) and unseen contexts (**Seen Test**). We attribute this, in part, to the fine-tuned model's capacity to reduce verbosity by generating responses in a dictionary-like style, and in part to the improved accuracy of the definitions, which helps minimize the hallucinations occasionally observed in the pre-trained model.

By comparing the performance of the fine-tuned Llama2Chat and Llama3Instruct models, we observe that their performances are very similar across all languages, with Llama3Instruct performing slightly better on average. However, for certain languages, such as Greek, Chinese and Japanese, we also observe a substantial performance gap in BERTScore and BLEU between the fine-tuned Llama2Chat and the Llama3Instruct models (both fine-tuned Llama3Instruct and baseline model), suggesting that Llama3Instruct has been exposed to significantly more data in these languages.[5] Interestingly, for few languages such as Swedish, German, and Russian, we observed slightly but consistently higher performance with fine-tuned Llama2Chat than fine-tuned Llama3Instruct.

For all fine-tuned models, we observe a performance drop on the **Unseen Test** sets compared to the **Seen Test** sets for all languages, Catalan being the only exception. We hypothesize that, similar to Malagasy, the Catalan result is influenced by the quality of definitions in Catalan, as we identified the presence of some trivial examples for *verb* entries (see Appendix B).

Overall, our results indicate that fine-tuning for definition generation is highly effective, even for low-resource languages, with greater average improvements observed in languages where the baseline average performance was lower (see Figure 2). We did not find a clear correlation between **Train** set size and performance.

We now focus on the performance of our fine-tuned Llama3Instruct models. The highest values for BERTScore on the **Seen Test** set are for German (0.884), English (0.857), Latin (0.841), Swedish (0.834), and Chinese (0.832). However, for the **Unseen Test** set, a different set of languages has the highest BERTScore, namely Kurdish (0.800), Spanish (0.750), and Italian (0.702) followed by German (0.696) and Latin (0.688).

For BLEU, we observed strong performance for English (0.635) and German (0.631), followed by Latin (0.505), Swedish (0.487), and French (0.465) on the **Seen Test** set. However, performance on the

[5] Meta (2024) (Llama3Instruct) report categorizing documents into 176 languages, whereas Meta (2023) (Llama2Chat) mention only 27 categorized languages, with 8% of the pre–training data consisting of documents in unknown languages.

Unseen Test set remains considerably lower across languages, but remain higher than the baseline.

Considering xCOMET, we observed the highest scores for German (0.885 on Seen, 0.708 on Unseen) and English (0.832 / 0.637), followed closely by Dutch (0.814 / 0.690) and Swedish (0.754 / 0.568), indicating consistently strong performance across both test settings for these languages.

4 Multilingual Definition Generation

In this setting, we considered different combinations of languages. First, we grouped the languages into three language families: Romance (R), Germanic (G), and Slavic (S). For each family, we selected one set of languages to be used in this multilingual setting, and a different set for the cross-lingual setting (see next section). To mitigate potential biases due to differences in **Train** set sizes across languages, we selected languages for the multilingual setting such that it was possible to randomly sample 13k training examples for each language. We then evaluated the performance of individual languages on their respective **Unseen Test** sets. For the R and G families, we fixed sets of three languages each, while for the S family, we included only two languages, as no additional S languages were available in our dataset.

In addition to individual language families, we examined all possible combinations aimed at incorporating typologically diverse languages: R+G, R+S, G+S, and R+G+S. We also included a combination labeled All, which features languages from other families, such as Greek, Kurdish, and Japanese, in order to further enhance typological diversity. All selected languages and their corresponding combinations are summarized in Appendix D.

For each combination, we fine-tuned a distinct *multilingual* LoRA adapter on the included languages. Specifically, to save computations, we fine-tuned Llama3Instruct exclusively as it achieved higher average performance compared to Llama2Chat in the monolingual setting.

By comparing the performance of the different fine-tuned models, we examine whether including languages from the same or different families enhance or degrade performance, and whether incorporating a larger number of languages leads to higher improvements. To quantify the multilingual learning advantage (MLA), we compute the following aggregated quality measure:

		Multilingual		Crosslingual
Code	**Family**	**Train**	**Test**	**Test**
R	Romance	it es fr	it es fr	la pt ca
G	Germanic	sv de en	sv de en	da no nl
S	Slavic	pl ru	pl ru	-
R+G	Romance + Germanic	it es fr sv de en	it es fr sv de en	la pt ca da no nl
R+S	Romance + Slavic	it es fr pl ru	it es fr pl ru	la pt ca
G+S	Germanic + Slavic	sv de en pl ru	sv de en pl ru	da no nl
R+G+S	Romance + Germanic + Slavic	it es fr sv de en pl ru	it es fr sv de en pl ru	la pt ca da no nl
All	Romance + Germanic + Slavic + Others	it es fr sv de en pl ru el fi ku tr ja	it es fr sv de en pl ru el fi ku tr ja	la pt ca da no nl - zh mg lt

Table 1: Language combinations used for the multilingual and cross-lingual experiments. For each combination, we report the language family, the languages used for training and testing in the multilingual setup, and the languages used for cross-lingual evaluation. In the cross-lingual setting, models are additionally tested on all languages (even those from different families) that were not used during fine-tuning.

$$\text{MLA} = \frac{\Delta\text{BLEU} + \Delta\text{BERTScore} + \Delta\text{xCOMET}}{3}$$

where Δ represents the relative change (i.e., $\frac{x_f - x_p}{x_p}$) in performance between the fine-tuned LoRA model (i.e., x_f) and the corresponding pre-trained model (i.e., x_p). Relative change provides a normalized, scale-invariant measure of improvement or degradation, allowing fairer comparisons across settings with different baselines. Thus, it is important to note that the MLA represents the average of three distinct percentages, and even relatively low MLA values can capture meaningful differences in performance.

A positive MLA was observed for all languages, with varying magnitude depending on the baseline performance. Languages with higher baseline scores, such as Italian (0.378) in Figure 2, showed smaller gains in Figure 3 (i.e., 0.245) when fine-tuned on their language family. In contrast, languages with lower baseline scores, such as Russian (0.264), exhibited substantially larger gains (i.e., 3.778). To account for this behavior and better isolate the effect of multilinguality, we set, for each tested language x, the performance obtained with the model fine-tuned on its respective language family (i.e., Δ_x) as the new baseline (where we potentially expect higher gains; Chronopoulou et al.,

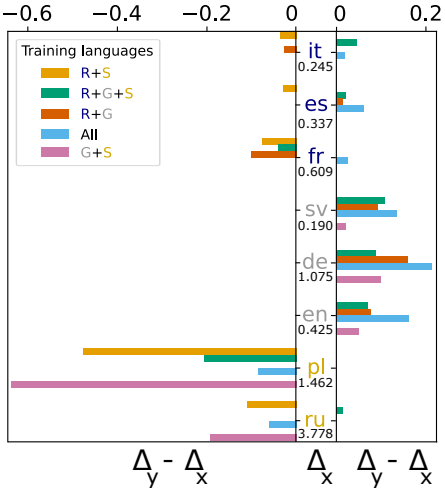

Figure 3: Δ MLA of multilingual models for languages included in their fine-tuning. For each language x, we report below its label the absolute Δ MLA observed when the multilingual model is trained only on its language family (i.e., Δ_x). The bar plot shows the improvement or degradation (i.e., Δ_y-Δ_x) when the model is trained on other language combinations y.

2023; de Vries et al., 2022; Lauscher et al., 2020) and analyze the MLA across different multilingual settings relative to it (i.e., Δ_y). Figure 3 shows differences Δ_y-Δ_x between family-based fine-tuning and alternative multilingual configurations.

In our evaluation, we observed that multilingual fine-tuning can yield either improvements or degradations depending on the target language and the languages included during training. This finding challenges the common assumption that the incorporation of more languages during training, even within the same language family, consistently leads to better performance on the training languages (Conneau et al., 2020). Recent studies align with this observation, suggesting that while adding related languages in multilingual fine-tuning may boost performance for some target languages, incorporating additional languages beyond an optimally selected subset may cause negative interference, reducing overall performance (Nigatu et al., 2023; Dhamecha et al., 2021).

Findings after fine-tuning. For Italian and Spanish, performance remains relatively stable across multilingual settings, with only minor performance degradation. This indicates that the addition of other languages does not significantly degrade performance nor offers any substantial benefit. Specifically, for Italian, adding languages from other families does not result in higher MLA. In contrast, Spanish exhibits the highest MLA, indicating a greater benefit from multilingual training.

French demonstrates clear improvements in the R-only and All configurations, while performance decreases when Germanic languages are introduced (R+G). This suggests that typological similarity (as in R-only) and greater language diversity (as in All) can be beneficial, whereas competition from high-resource Germanic languages (e.g., English and German) may hinder performance.

For the Germanic languages, training exclusively on other Germanic languages results in the lowest performance. Swedish, in particular, performs best when Romance languages are added (in the R+G and R+G+S configurations), while the inclusion of Slavic languages leads to lower results—both when added alone and in combination with Romance. This suggests that the improvements gained from Romance languages may stem from the introduction of complementary syntactic and semantic variety, which helps balance the influence of dominant languages like English and German.

German also improves when Romance languages are added (R+G), suggesting that certain cross-family pairings are more synergistic than same-family groupings. English achieves its highest scores in the All configuration, but the addition of other language families all improve the performance. As a high-resource language with significant representation in pre-trained models, English appears to be more robust when trained alongside a wider variety of languages.

Polish experiences a notable performance drop when trained with Germanic languages (G+S). This suggests that the influence of dominant Germanic languages, such as English and German, may overshadow Polish-specific features during training.

Unlike Polish, Russian does not appear to be negatively affected by the inclusion of Germanic or Romance languages (G+S, R+S, G+R+S); instead, it maintains stable performance across all multilingual configurations. We hypothesize that this may be due, at least in part, to its robust representation in the pre-training data.

Monoligual vs multilingual. A fair comparison between monolingual and multilingual settings would require retraining the monolingual models multiple times using controlled and equivalent training sizes. While this would be desirable, it is beyond the scope of this evaluation due to the significant computational and resource costs involved. However, for Italian and Spanish, we have a comparable amount of language-specific data (13k exam-

ples) in both *monolingual* and *multilingual* settings and we can thus make a comparison.

For Italian (monolingual average performance: 0.422), we observe a significant drop in performance across all multilingual configurations, ranging from 0.358 (G) to 0.408 (R+G+S). This suggests that the inclusion of languages from both related and unrelated families may introduce interference rather than providing a beneficial signal.

In contrast, Spanish maintains relatively stable performance across multilingual settings, from 0.405 (G) to 0.471 (All), with only minor degradations compared to its monolingual average (0.471). This indicates that the addition of other languages does not significantly harm performance but also provides limited benefit.

5 Cross-lingual Definition Generation

In this section, we investigate whether models fine-tuned on individual languages (see Section 3) or multiple languages (see Section 4) can transfer their learned knowledge to other, unseen languages for the task of Definition Generation.

As before, we compute BERTScore, BLEU, and xCOMET values. Specifically, to assess cross-lingual transfer from one trained language to an unseen language, we use the relative change in performance to measure improvements or degradations relative to the pre-trained baseline on the **Unseen Test** sets of the target language. Our results are presented in Figure 6 (Appendix H).

In the following, we first summarize our results and findings for single-language training (1-to-1 transfer) and then for multiple-language training (many-to-1 transfer). In general, we believe that, in this setting, cross-lingual transfer primarily involves task-specific knowledge rather than linguistic features. This is because Llama3Instruct, as a large multilingual language model, already encodes substantial information across a wide range of languages. As such, we expect this behavior to resemble few-shot learning rather than traditional cross-lingual transfer.

1-to-1 transfer When analyzing the performance of models fine-tuned on a single language, we observed signs of *language attrition* (Gallo et al., 2021). Language attrition refers to the decline of active language abilities (e.g., generation) despite the retention of passive skills (e.g., comprehension). In our context, this manifests as the Llama3Instruct model losing its ability to generate coherent definitions in previously known languages after extensive fine-tuning on a different one (see Figure 7).

Since this can lead to artificially low BERTScore and BLEU scores (both of which rely on same-language comparisons and may penalize outputs generated in a language different from the reference) we focus our evaluation on xCOMET that is designed for multilingual evaluation and remains reliable even when the output and reference are in different languages. We use the relative change in xCOMET to measure improvements or degradations compared to the pre-trained baseline on the target language's **Unseen Test** set.

The observed relative change in xCOMET scores confirms our hypothesis: even when models are fine-tuned on only a few hundred examples, they still achieve notable improvements on unseen languages. E.g., models fine-tuned on low-resource languages such as Latin and Lithuanian demonstrate improved performance on several other languages, despite their limited training data. These findings are consistent with the results reported by Yuan et al. (2024), who observed similar trends for both full fine-tuning and LoRA-based fine-tuning.

Firstly, we note that training on almost any donors language improves the performance on the *other* languages, namely Greek (to a smaller extent), Japanese, Turkish, Chinese, Malagasy and Lithuanian, further supporting the transfer of task-specific knowledge rather than cross-lingual knowledge. A few languages stand out: the fine-tuned model on French degrades performance on all other recipient languages. Swedish in turn is not benefited from any language except to a very small extent from fine-tuning on Lithuanian (0.086). In contrast, German as a recipient benefits from almost all other languages (excluding Malagasy). Up to 70% improvement occurs when training on Portuguese, Danish, Kurdish and Lithuanian. English in turn, though both a Germanic and high-resource language like German, does not benefit equally and often degrades in performance.

By analyzing the xCOMET relative change, we observed that models fine-tuned on larger datasets in a single language (e.g., English, German, French, Russian) are more susceptible to forgetting other languages. In contrast, fine-tuning on smaller datasets (e.g., Japanese, Portuguese, and Norwegian) can lead to improved performance on unseen languages compared to the pre-trained baseline.

In general, models fine-tuned on a single language exhibit stronger transfer to Germanic lan-

guages, even when fine-tuned on a language from a different family. In contrast, English, which is associated with the largest amount of data in the baseline pre-training, experiences the most severe forgetting when fine-tuned on another language.

Fine-tuning on Malagasy leads to forgetting across all Germanic, Romance, and Slavic languages. However, as noted in the previous section, we found lower-quality data for this language and report these results only for completeness.

Many-to-1 transfer In contrast to 1-to-1 transfer, language attrition appears to be mitigated—or even entirely avoided—in multilingual fine-tuning settings, where the presence of multiple languages helps preserve the model's generation capabilities across them (see Figure 7).

Models fine-tuned on multiple languages consistently yield positive transfer across all unseen languages, with the sole exception of models trained exclusively on Germanic languages and evaluated on Romance targets such as Italian, Spanish, and Catalan. In these cases, cross-family interference appears to hinder transfer, leading to slight performance degradation.

Moreover, consistent with the multilingual setting, increasing the number of training languages does not necessarily result in better performance. For example, the best improvement for Japanese are obtained when the model is fine-tuned on a combination of G+S languages, rather than on the full R+G+S set.

We also note that German receives the highest transfer benefits across several fine-tuning configurations, likely due to its strong representation in the pre-trained model.

6 Related work

Word embeddings are the current standard for modeling word meaning, but suffer from fundamental limitations that motivate the exploration of alternative approaches. When *type* embeddings are considered – assigning a single vector to each word – the modeling suffers from *meaning conflation deficiency* (Pilehvar, 2019), merging distinct senses into a single representation and obscuring lexical ambiguity. Alternatively, when *token* embeddings are considered – assigning a single vector to each word occurrence – addressing polysemy requires clustering techniques to distinguish word senses. However, clustering tends to capture contextual variation rather than discrete semantic categories (Kutuzov et al., 2022), resulting in clusters that are often noisy and difficult to interpret reliably.

Definition Generation directly addresses these challenges by replacing opaque vector-based representations with explicit, human-readable definitions. In the literature, the task is also referred to as *Definition Modeling* (DM). However, these two terms often refer to different formulations, with the latter typically referring to the original version of the task, which was proposed as a means of interpreting the vector space of word embeddings (Noraset et al., 2017). DM was initially defined as generating a natural language definition given a target word embedding, with early works primarily focusing on the interpretation of *type* embeddings (Washio et al., 2019; Bosc and Vincent, 2018).

The task addressed in this work instead follows the sequence-to-sequence formulation introduced by Ishiwatari et al. (2019); Gadetsky et al. (2018), which aims to generate a sense definition given a target word in context. Thus far, evaluations of different approaches and models have primarily been conducted for English (Periti et al., 2024; Giulianelli et al., 2023; Bevilacqua et al., 2020), with training and evaluation data extracted from English WordNet and Wikipedia (Ishiwatari et al., 2019), Oxford English Dictionary (Gadetsky et al., 2018), Wiktionary (Mickus et al., 2022), and Urban Dictionary (Ni and Wang, 2017).

Languages other than English, however, have received relatively limited attention in this context. Some studies have focused on Chinese (Zheng et al., 2021), with particular emphasis on complexity-controllable definition generation (Yang et al., 2024; Yuan et al., 2022; Kong et al., 2022). Also relevant to our work is the research by Zhang et al. (2023), which investigates the trans-lingual generation of definitions in a target language for words in another language – for example, generating English definitions for Chinese words. Recently, Definition Generation has gained popularity for semantic change detection (Periti and Montanelli, 2024; Periti et al., 2024; Fedorova et al., 2024a; Giulianelli et al., 2023), with new research focusing on Russian, Finnish, and German (Fedorova et al., 2024b). Additional languages that have been explored include Wolastoqey (Bear and Cook, 2021), Portuguese (Dimas Furtado et al., 2024), Spanish (Rodríguez-Betancourt

and Casasola-Murillo, 2023), Polish (Wojtasik et al., 2023), and French (Mickus et al., 2020). These works have typically concentrated on individual languages rather than systematically investigating multiple languages or cross-lingual transfer (Kong et al., 2020). Furthermore, a variety of resources have been used to construct datasets for the two settings of Definition Generation, making it difficult to evaluate and compare performance across different studies and languages.[6]

Generation approaches span a variety of model architectures, reflecting both the evolving nature of the task and broader progress in NLP. Early work relies on RNN-based encoder–decoder models (Noraset et al., 2017), often conditioning the encoder on additional contextual information. For example, Ni and Wang (2017) use a dual-LSTM encoder that incorporates both word-level context and character-level representations of the target word. With the advent of Transformers, later work predominantly focuses on fine-tuning pre-trained models such as MASS (Kong et al., 2022), BART (Bevilacqua et al., 2020), M2M (Zhang et al., 2023), Flan-T5 (Giulianelli et al., 2023), and recently Llama (Periti et al., 2024). These studies typically concatenate each usage example with a prompt to describe the task, introduce lexical constraints, or control definition complexity. In our work, we also focus on Llama models; however, we expand Definition Generation across 22 languages, offering monolingual, multilingual, and cross-lingual perspectives.

7 Conclusion

We presented the first large-scale study of Definition Generation across 22 languages, introducing a new benchmark based on Dbnary data and a systematic evaluation of Llama-based models fine-tuned via LoRA. Our evaluation spans monolingual, multilingual, and cross-lingual settings. In the monolingual setting, fine-tuning leads to substantial improvements over pre-trained baselines – particularly for languages with initially low performance – by producing concise, dictionary-style definitions and reducing hallucinations. In the multilingual setting, we find that incorporating related or typologically diverse languages can benefit certain target languages (e.g., German and English), but may cause negative interference for others (notably Slavic languages when paired with Germanic ones). Cross-lingual experiments reveal the presence of language attrition following single-language fine-tuning, whereas multilingual fine-tuning largely preserves, and often improves, Definition Generation for unseen languages. In future work, we plan to include human evaluations to complement metric-based assessments and further refine definition quality. By releasing our data splits, code, and fine-tuned adapters, we aim to facilitate research on interpretable, multilingual modeling of word meaning beyond English.

[6] A closely related work was published after the completion of our study (Marrese-Taylor et al., 2025). Future work should consider comparing our findings with their recent contribution.

Limitations

By considering this study, the reader should keep in mind a few limitations that might influence the interpretation of our findings:

- **Data quality variability:** Our analysis relies exclusively on data sourced from Wiktionary and Dbnary, both of which are products of collaborative efforts that combine manual contributions and automated processes. While these platforms are valuable, the automated processes may result in inconsistencies and varying quality across some entries. We expect a medium-to-high quality of data, depending on the language, especially considering that Wiktionary has been previously used for word meaning modeling tasks such as Word-in-Context (Raganato et al., 2020), Word Sense Disambiguation (Segonne et al., 2019), and Definition Generation (Kabiri and Cook, 2020). However, we anticipate higher quality for more popular languages compared to low-resource languages, which could introduce variability in the strength of our analysis. Despite this, we believe that the analysis remains meaningful. Future studies should consider incorporating manual annotations to both quantitatively assess data quality and further enhance it, thereby mitigating these issues.

- **Biases in Evaluation Metrics:** We did not conduct human evaluation of the generated definitions across languages. Instead, we relied on widely used automatic evaluation metrics for Natural Language Generation, including both metrics based on lexical-overlap-based and semantic similarity. However, when dealing with languages other than English,

such metrics may introduce biases, particularly due to variations in language representation quality and the pre-training data of the underlying models.

In particular, for BERTScore, one possible option was to use the same multilingual model across all languages. While this would have ensured a consistent evaluation setting, multilingual models often underperform on low-resource languages due to their limited presence in the training corpus. To address this, we opted to use monolingual BERT models pre-trained specifically for each language, except for Kurdish and Malagasy (see Table 6). For Kurdish and Malagasy, we used *XLM-RoBERTa*, as it has been trained on these languages and no suitable language-specific models were available. In the case of Kurdish, we also experimented with the *KuBERT-Central-Kurdish-BERT-Model* (Awlla et al., 2025); however, we encountered compatibility issues when using this model with the BERTScore implementation. While this introduces potential variability due to model differences, we believe it offers a fairer and more accurate evaluation for each language.

We also used xCOMET, a learned metric based on *XCOMET-XL* (Guerreiro et al., 2024) (a fine-tuned version of XLM-R (Conneau et al., 2020)) which enables cross-lingual comparison. However, like other multilingual pre-trained models, it may reflect biases rooted in the uneven distribution of training data across languages. In this study, we do not explicitly account for these potential biases in model-based metrics.

- **Scope of Parameter Evaluation:** Given the vast parameter space of LLMs, conducting a comprehensive evaluation of all possible configurations is impractical. This study focuses primarily on language as the main variable, deliberately excluding other influential parameters such as temperature settings and decoding strategies.

- **Single Language Model:** Our experiments are confined to the Llama model, chosen for its strong performance across multiple benchmarks. Incorporating additional LLMs could have provided a more comprehensive understanding of model behaviors and performance across diverse architectures. However, expanding the scope to include multiple models would have substantially increased the computational and analytical demands of the study, rendering it unfeasible within our resource constraints.

- **Single prompt template:** Throughout our experiments, we employed a single prompt template, translated into various languages. While different prompts could yield variations in performance, prior research on Definition Generation with LLMs has tested multiple prompts and found that the highest results were obtained with the most reasonable prompts, which also showed similar performance (Giulianelli et al., 2023). Therefore, we are confident that our findings remain valid despite this limitation. Importantly, since we applied the same prompt across all baseline and fine-tuned models, any limitations related to prompt selection affect all models equally, ensuring the fairness of our comparative analysis between pre-trained and fine-tuned ones.

Acknowledgments

This work has in part been funded by the research program *Change is Key!* supported by Riksbankens Jubileumsfond (under reference number M21-0021). The computational resources were provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

References

Jordi Armengol-Estapé, Casimiro Pio Carrino, Carlos Rodriguez-Penagos, Ona de Gibert Bonet, Carme Armentano-Oller, Aitor Gonzalez-Agirre, Maite Melero, and Marta Villegas. 2021. Are multilingual models the best choice for moderately under-resourced languages? A comprehensive assessment for Catalan. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4933–4946, Online. Association for Computational Linguistics.

K.M. Awlla, H. Veisi, and A.A. Abdullah. 2025. Sentiment analysis in low-resource contexts: BERT's impact on Central Kurdish. *Language Resources & Evaluation*, 35(1).

David Bamman and Patrick J. Burns. 2020. Latin bert: A contextual language model for classical philology. *Preprint*, arXiv:2009.10053.

María A. Barrios, Guadalupe Aguado de Cea, and José Ángel Ramos. 2009. Enriching a Lexicographic Tool with Domain Definitions: Problems and Solutions. In *Proceedings of the 1st Workshop on Definition Extraction*, pages 14–20, Borovets, Bulgaria. Association for Computational Linguistics.

Diego Bear and Paul Cook. 2021. Cross-Lingual Wolastoqey-English Definition Modelling. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 138–146, Held Online. INCOMA Ltd.

Michele Bevilacqua, Marco Maru, and Roberto Navigli. 2020. Generationary or "How We Went beyond Word Sense Inventories and Learned to Gloss". In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7207–7221, Online. Association for Computational Linguistics.

Tom Bosc and Pascal Vincent. 2018. Auto-Encoding Dictionary Definitions into Consistent Word Embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1522–1532, Brussels, Belgium. Association for Computational Linguistics.

Jose Camacho-Collados and Mohammad Taher Pilehvar. 2018. From Word to Sense Embeddings: a Survey on Vector Representations of Meaning. *J. Artif. Int. Res.*, 63(1):743–788.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. 2023. Language-Family Adapters for Low-Resource Multilingual Neural Machine Translation. In *Proceedings of the Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023)*, pages 59–72, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

Michal Mimino Danilak. 2021. langdetect: Language detection library ported from google's language-detection. `https://pypi.org/project/langdetect/`. Version 1.0.9.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT Model. *Preprint*, arXiv:1912.09582.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2022. Make the Best of Cross-lingual Transfer: Evidence from POS Tagging with over 100 Languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7676–7685, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tejas Dhamecha, Rudra Murthy, Samarth Bharadwaj, Karthik Sankaranarayanan, and Pushpak Bhattacharyya. 2021. Role of Language Relatedness in Multilingual Fine-tuning of Language Models: A Case Study in Indo-Aryan Languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8584–8595, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anna Beatriz Dimas Furtado, Tharindu Ranasinghe, Frederic Blain, and Ruslan Mitkov. 2024. DORE: A Dataset for Portuguese Definition Generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5315–5322, Torino, Italia. ELRA and ICCL.

George Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using n-gram Co-occurrence Statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, page 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Mariia Fedorova, Andrey Kutuzov, and Yves Scherrer. 2024a. Definition Generation for Lexical Semantic Change Detection. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5712–5724, Bangkok, Thailand. Association for Computational Linguistics.

Mariia Fedorova, Timothee Mickus, Niko Partanen, Janine Siewert, Elena Spaziani, and Andrey Kutuzov. 2024b. AXOLOTL'24 Shared Task on Multilingual Explainable Semantic Change Modeling. In *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, pages 72–91, Bangkok, Thailand. Association for Computational Linguistics.

Artyom Gadetsky, Ilya Yakubovskiy, and Dmitry Vetrov. 2018. Conditional Generators of Words Definitions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 266–271, Melbourne, Australia. Association for Computational Linguistics.

Federico Gallo, Beatriz Bermudez-Margaretto, Yury Shtyrov, Jubin Abutalebi, Hamutal Kreiner, Tamara Chitaya, Anna Petrova, and Andriy Myachykov. 2021. First language attrition: What it is, what it isn't, and what it can be. *Frontiers in Human Neuroscience*, 15:686388.

Mario Giulianelli, Iris Luden, Raquel Fernandez, and Andrey Kutuzov. 2023. Interpretable Word Sense Representations via Definition Generation: The Case of Semantic Change Analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3130–3148, Toronto, Canada. Association for Computational Linguistics.

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.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Shonosuke Ishiwatari, Hiroaki Hayashi, Naoki Yoshinaga, Graham Neubig, Shoetsu Sato, Masashi Toyoda, and Masaru Kitsuregawa. 2019. Learning to Describe Unknown Phrases with Local and Global Contexts. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3467–3476, Minneapolis, Minnesota. Association for Computational Linguistics.

Arman Kabiri and Paul Cook. 2020. Evaluating a multi-sense definition generation model for multiple languages. In *Text, Speech, and Dialogue*, pages 153–161, Cham. Springer International Publishing.

Cunliang Kong, Yun Chen, Hengyuan Zhang, Liner Yang, and Erhong Yang. 2022. Multitasking Framework for Unsupervised Simple Definition Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5934–5943, Dublin, Ireland. Association for Computational Linguistics.

Cunliang Kong, Liner Yang, Tianzuo Zhang, Qinan Fan, Zhenghao Liu, Yun Chen, and Erhong Yang. 2020. Toward Cross-Lingual Definition Generation for Language Learners. *Preprint*, arXiv:2010.05533.

Matej Kosec, Sheng Fu, and Mario Michael Krell. 2021. Packing: Towards 2x NLP BERT Acceleration.

John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. GREEK-BERT: The Greeks Visiting Sesame Street. In *11th Hellenic Conference on Artificial Intelligence*, SETN 2020, page 110–117, New York, NY, USA. Association for Computing Machinery.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. *Preprint*, arXiv:1905.07213.

Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2022. Contextualized Embeddings for Semantic Change Detection: Lessons Learned. *Northern European Journal of Language Technology*, 8.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Edison Marrese-Taylor, Erica K. Shimomoto, Alfredo Solano, and Enrique Reid. 2025. Multilingual definition modeling. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25888–25906, Vienna, Austria. Association for Computational Linguistics.

John P. McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The ontolex-lemon model: Development and applications. In *Electronic Lexicography in the 21st Century. Proceedings of eLex 2017 Conference*, pages 19–21. Lexical Computing CZ s.r.o.

Meta. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *Preprint*, arXiv:2307.09288.

Meta. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Timothee Mickus, Mathieu Constant, and Denis Paperno. 2020. Génération automatique de définitions pour le français (Definition Modeling in French). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, pages 66–80, Nancy, France. ATALA et AFCP.

Timothee Mickus, Kees Van Deemter, Mathieu Constant, and Denis Paperno. 2022. Semeval-2022 Task 1: CODWOE – Comparing Dictionaries and Word Embeddings. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1–14, Seattle, United States. Association for Computational Linguistics.

Shuyo Nakatani. 2010. Language detection library for java.

Ke Ni and William Yang Wang. 2017. Learning to Explain Non-Standard English Words and Phrases. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2:*

Short Papers), pages 413–417, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Hellina Nigatu, Atnafu Tonja, and Jugal Kalita. 2023. The Less the Merrier? Investigating Language Representation in Multilingual Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12572–12589, Singapore. Association for Computational Linguistics.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition Modeling: Learning to Define Word Embeddings in Natural Language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Riccardo Orlando, Luca Moroni, Pere-Lluís Huguet Cabot, Edoardo Barba, Simone Conia, Sergio Orlandini, Giuseppe Fiameni, and Roberto Navigli. 2024. Minerva LLMs: The First Family of Large Language Models Trained from Scratch on Italian Data. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, Pisa, Italy. CEUR Workshop Proceedings.

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.

Francesco Periti, David Alfter, and Nina Tahmasebi. 2024. Automatically Generated Definitions and their utility for Modeling Word Meaning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14008–14026, Miami, Florida, USA. Association for Computational Linguistics.

Francesco Periti and Stefano Montanelli. 2024. Lexical semantic change through large language models: a survey. *ACM Comput. Surv.*, 56(11).

Mohammad Taher Pilehvar. 2019. On the Importance of Distinguishing Word Meaning Representations: A Case Study on Reverse Dictionary Mapping. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2151–2156, Minneapolis, Minnesota. 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*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A Multilingual Benchmark for Evaluating Semantic Contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, Online. Association for Computational Linguistics.

Esteban Rodríguez-Betancourt and Edgar Casasola-Murillo. 2023. Exploring the Limits of Large Language Models for Word Definition Generation: A Comparative Analysis. In *2023 XLIX Latin American Computer Conference (CLEI)*, pages 1–7.

Vincent Segonne, Marie Candito, and Benoît Crabbé. 2019. Using Wiktionary as a resource for WSD : the case of French verbs. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 259–270, Gothenburg, Sweden. Association for Computational Linguistics.

Gilles Sérasset. 2012. Dbnary: Wiktionary as a LMF based Multilingual RDF network. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2466–2472, Istanbul, Turkey. European Language Resources Association (ELRA).

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Gilles Sérasset. 2015. Dbnary: Wiktionary as a lemon-based multilingual lexical resource in rdf. *Semantic Web*, 6(4):355–361.

Bram Vanroy. 2024. GEITje 7B Ultra: A Conversational Model for Dutch. *Preprint*, arXiv:2412.04092.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. *Preprint*, arXiv:1912.07076.

Koki Washio, Satoshi Sekine, and Tsuneaki Kato. 2019. Bridging the Defined and the Defining: Exploiting Implicit Lexical Semantic Relations in Definition Modeling. In *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 3521–3527, Hong Kong, China. Association for Computational Linguistics.

Konrad Wojtasik, Arkadiusz Janz, and Maciej Piasecki. 2023. Wordnet for Definition Augmentation with Encoder-Decoder Architecture. In *Proceedings of the 12th Global Wordnet Conference*, pages 50–59, University of the Basque Country, Donostia - San Sebastian, Basque Country. Global Wordnet Association.

Liner Yang, Jiaxin Yuan, Cunliang Kong, Jingsi Yu, Ruining Chong, Zhenghao Liu, and Erhong Yang. 2024.

Tailored Definitions With Easy Reach: Complexity-Controllable Definition Generation. *IEEE Transactions on Big Data*, pages 1–12.

Fei Yuan, Shuai Yuan, Zhiyong Wu, and Lei Li. 2024. How Vocabulary Sharing Facilitates Multilingualism in LLaMA? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12111–12130, Bangkok, Thailand. Association for Computational Linguistics.

Jiaxin Yuan, Cunliang Kong, Chenhui Xie, Liner Yang, and Erhong Yang. 2022. COMPILING: A Benchmark Dataset for Chinese Complexity Controllable Definition Generation. In *Proceedings of the 21st Chinese National Conference on Computational Linguistics*, pages 921–931, Nanchang, China. Chinese Information Processing Society of China.

Hengyuan Zhang, Dawei Li, Yanran Li, Chenming Shang, Chufan Shi, and Yong Jiang. 2023. Assisting Language Learners: Automated Trans-Lingual Definition Generation via Contrastive Prompt Learning. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 260–274, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. *Preprint*, arXiv:1904.09675.

Hua Zheng, Damai Dai, Lei Li, Tianyu Liu, Zhifang Sui, Baobao Chang, and Yang Liu. 2021. Decompose, Fuse and Generate: A Formation-Informed Method for Chinese Definition Generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5524–5531, Online. Association for Computational Linguistics.

A Example of paraphrastic definition with low BLEU score

The following Turkish example illustrates how low BLEU scores may result from paraphrasing rather than poor output quality.

target	kümes (poultry house)
example	Ne kümes de tavuk bırakır mı, ne ahırda hayvan, ne de ağılda koyun. – H. R. Gürpınar (There are neither chickens in the coop, nor animals in the barn, nor sheep in the pasture. – H. R. Gürpınar)
definition	(Mimarlık) Tavuk, hindi gibi evcil hayvanların, barınmasına yarayan kapalı yer. ((Architecture) a closed space used for housing domestic animals such as chickens, turkeys, etc.)
output	Hayvanların barındığı çit veya duvarla çevrili yer, arkaç. (A place surrounded by a fence or wall where animals are kept, a pen.)

In this example, the corresponding metric values for the definition generated by the fine-tuned Llama3Instruct model are as follows: 0.000 for BLEU, but 0.477 for xCOMET and 0.608 for BERTScore.

B Abnormal improvement on Catalan

For all fine-tuned models, we observe a performance drop on the **Unseen Test** sets compared to the **Seen Test** sets for all languages, Catalan being the only exception. We hypothesize that, similar to Malagasy, the Catalan result is influenced by the quality of definitions in Catalan, as we identified the presence of some trivial examples, such as the following:

target	saberiva (knew)
example	Forma algueresa per [ell/ella/vostè] sabria. (Algherese form for [he/she/you formal] *sabria*.)
definition	Tercera persona del singular (ell, ella, vostè) del condicional del verb saber. (Third person singular [he, she, you formal] of the conditional tense of the verb *saber*.)

In such cases, the model outputs nearly the same text as the example. We observed this undesirable pattern only for some *verb* entries.

C Dataset overview

Table 2 provides an example of the dataset structure while Figure 4 offers an overview of the amount of data considered for different languages in our work. As expected, more examples are available for some languages (e.g., English) than for others (e.g., Latin).

	Target w	**Example** e	**Definition** d
EN	tea	Would you like some **tea**?	The drink made by infusing dried leaves or buds in hot water.
IT	pastasciutta	Mangiarsi un bel piatto di **pastasciutta** è uno dei piaceri della vita.	(Gastronomia) pasta alimentare solitamente e principalmente composta di grano tipica della cucina italiana preparata attraverso la bollitura, la scolatura e il condimento.
FR	fromage	Par définition, les **fromages** sont une forme de conservation des deux constituants insolubles du lait, la caséine et la matière grasse. [...]	Aliment moulé, obtenu à partir de la coagulation du lait suivie ou non de fermentation.
TR	çay	Beş **çayına** davetliydik.	Çayla birlikte ufak tefek şeyler ikram edilen toplantı.

Table 2: Example instances from our dataset for English (EN), Italian (IT), French (FR), and Turkish (TR). In our dataset, the examples e may include the target word in both its lemma form and different inflected forms.

D Language combinations for multilingual and cross-lingual experiments

Table 3 summarizes all selected languages and their combinations used in the multilingual and cross-lingual experiments. For each combination, we report the language family, the languages used for training and testing in the multilingual setup, and the languages used for cross-lingual evaluation. In the cross-lingual setting, models are additionally tested on all languages (even those from different families) that were not used during fine-tuning.

		Multilingual		Crosslingual
Code	**Family**	**Train**	**Test**	**Test**
R	Romance	it es fr	it es fr	la pt ca
G	Germanic	sv de en	sv de en	da no nl
S	Slavic	pl ru	pl ru	-
R+G	Romance + Germanic	it es fr sv de en	it es fr sv de en	la pt ca da no nl
R+S	Romance + Slavic	it es fr pl ru	it es fr pl ru	la pt ca
G+S	Germanic + Slavic	sv de en pl ru	sv de en pl ru	da no nl
R+G+S	Romance + Germanic + Slavic	it es fr sv de en pl ru	it es fr sv de en pl ru	la pt ca da no nl
All	Romance + Germanic + Slavic + Others	it es fr sv de en pl ru el fi ku tr ja	it es fr sv de en pl ru el fi ku tr ja	la pt ca da no nl - zh mg lt

Table 3: Language combinations used for the multilingual and cross-lingual experiments.

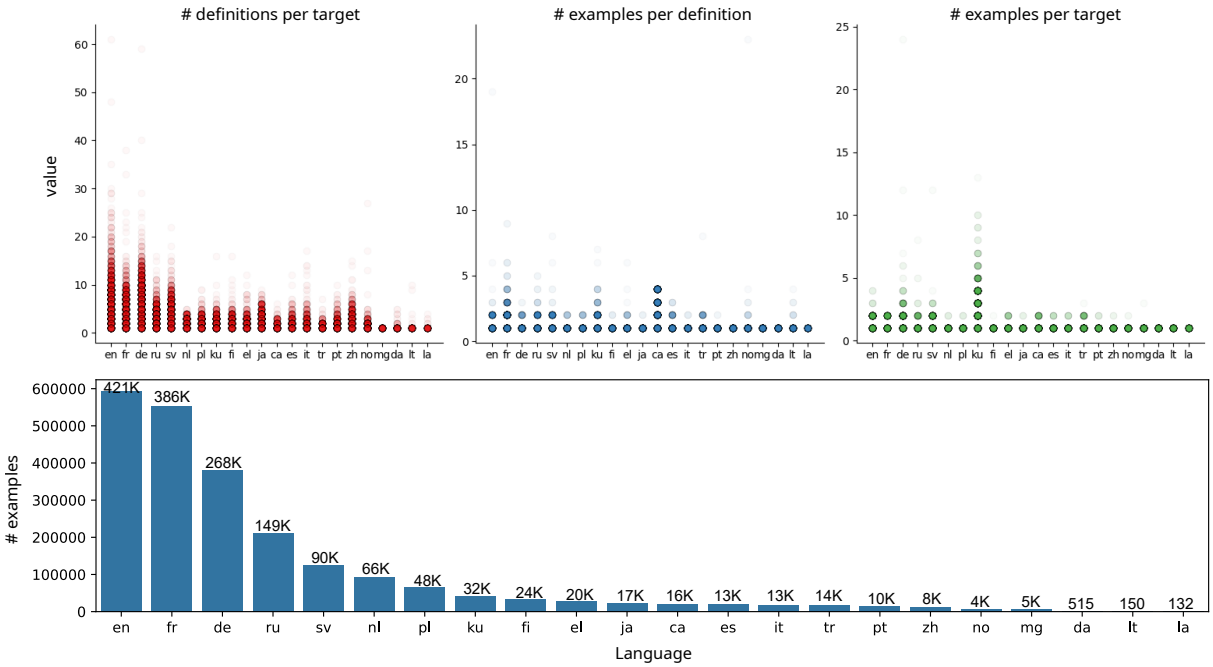

Figure 4: **Top row**: Number of definitions per target (left), examples per definition (middle), and examples per target (right). **Bottom**: Total number of examples available for each language. For each language, the number of training examples is also reported at the top of each bar.

E Correlation of evaluation metrics

Pearson correlation coefficients were calculated over the average performance of fine-tuned and pre-trained models across **Unseen Test** sets in different languages. These can be seen in Figure 5.

F Prompts and fine-tuning parameters

In our experiment, we fine-tune Llama2Chat and Llama3Instruct using the parameters reported in Table 4. Additional parameters are available online in our GitHub repository and Hugging Face checkpoint.

During fine-tuning and inference, we prompt the models using the chat template format they were originally trained on. This format includes specific markers to denote the start and end of system, user, and assistant messages (the assistant message appears only during fine-tuning). For English prompts, we adopt the format proposed by Periti et al. (2024), as shown below:

> **SYSTEM**: `You are a lexicographer familiar with providing concise definitions of word meanings.`
> **USER**: `Please provide a concise definition for the meaning of the word "{TARGET}"in the following sentence: "{EXAMPLE}".`
> **Assistant**: `{DEFINITION}`.

However, we translated these prompts into all considered languages using GPT-4o (OpenAI, 2024).

Fine-tuning details	
pre-trained LLMs	*Llama-2-7b-chat-hf* *Meta-Llama-3-8B-Instruct*
GPUs	A100fat (80GB)
PEFT	LoRA
LoRA dropout	0.1
Weight decay	0.01
Learning rate	1e-4
Lora rank	256
Lora alpha	512
Warmup ratio	0.15
Max train epochs	50
Early stopping patience	5
Early stopping threshold	0.001
Gradient accumulation steps	1
Max seq. length	300
Batch size	40
Optimizer	paged_adamw_8bit
LoRA target modules	q_proj, v_proj, k_proj, o_proj, gate_proj, up_proj, down_proj

Table 4: Settings and parameters for fine-tuning Llama2Chat and Llama3Instruct.

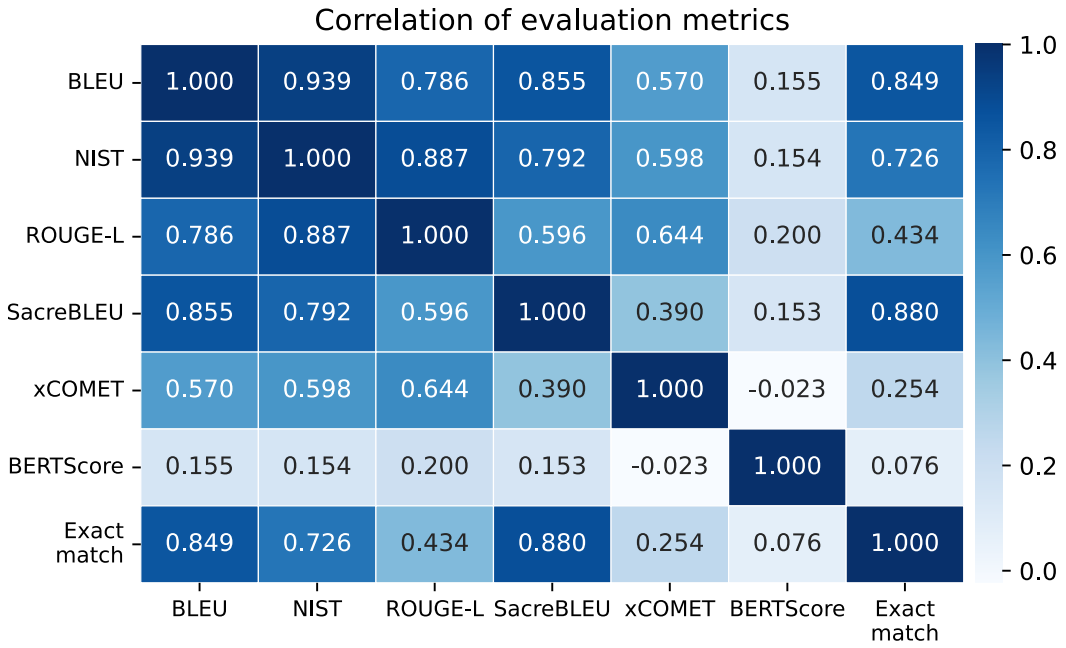

Figure 5: Correlation matrix of evaluation metrics for Definition Generation. Pearson correlation coefficients were computed over the average performance of fine-tuned and pre-trained models across **Unseen Test** sets in different languages.

G Preliminary comparison of language-specific vs. Llama models

Table 5 presents a comparative evaluation of language-specific models versus Llama models in the monolingual setting for Swedish, Dutch, and Italian, detailing performance metrics on both seen and unseen test sets.

H Impact of Fine-tuning in the cross-lingual setting

Figure 6 illustrates the Multilingual Advantage (MLA) across various source and target language pairs, highlighting the dynamics of transfer and forgetting in cross-lingual definition generation.

I Language attrition

To measure language attrition-the tendency of a model to produce outputs in the fine-tuning language instead of the input language-we use the langdetect Python library (Nakatani, 2010; Danilak, 2021) to automatically identify the language of model outputs. For each model, we randomly sample 1,000 outputs per input language (or use all outputs if fewer are available) and apply langdetect to each output to determine its language.

We observe that monolingual models frequently generate outputs in the fine-tuning language, regardless of the input, whereas multilingual models are more robust and generally produce outputs in the target language.

To analyze this behavior across models, we construct confusion matrices where each row corresponds to the input language and each column represents the language predicted by langdetect for the generated output. To summarize across models, we aggregate confusion matrices within each model type (monolingual or multilingual) by taking the element-wise maximum—this highlights the most severe cases of language mismatches observed in each condition.

To assess the reliability of langdetect for our purposes, we evaluated its performance on 1,000 sampled word usage examples per language. The tool achieved an overall F1 score of 0.57. Excluding Latin and Kurdish—languages not supported by langdetect—improves this to 0.62. Notably, the lower scores are largely due to misclassifications in just five languages: Japanese, Greek, Russian, Malagasy, and Chinese. When these are excluded, the F1 score rises substantially to 0.91 across the remaining 15 languages, demonstrating that langdetect is a suitable tool for approximately measuring the language attrition in the majority of our evaluation set.

Language	Model	Training	BLEU	NIST	ROUGE-L	SacreBLEU	xCOMET	BERTScore	Ex. Match
Swedish	*Meta-Llama-3-8B-Instruct*	pre-trained	-	-	-	-	-	-	-
			0.030	0.160	0.096	2.105	0.495	0.607	0.000
		fine-tuned	**0.487**	**1.525**	**0.593**	**45.267**	**0.754**	**0.834**	**0.313**
			0.116	**0.354**	**0.215**	**9.805**	0.568	**0.685**	**0.037**
	Llama-3-8B-instruct Llama fine-tuned on Swedish	pre-trained	-	-	-	-	-	-	-
			0.026	0.150	0.086	1.661	0.365	0.603	0.000
		fine-tuned	0.432	1.388	0.554	40.206	0.743	0.817	0.251
			0.106	0.345	0.209	8.729	**0.578**	0.680	0.029
Dutch	*Meta-Llama-3-8B-Instruct*	pre-trained	-	-	-	-	-	-	-
			0.036	0.222	0.109	1.973	0.384	0.574	0.000
		fine-tuned	**0.464**	**1.527**	**0.560**	**43.866**	**0.814**	**0.793**	**0.313**
			0.214	**0.700**	**0.311**	**19.635**	**0.691**	**0.678**	**0.117**
	GEITje-7B-ultra Mistral fine-tuned on Dutch (Vanroy, 2024)	pre-trained	-	-	-	-	-	-	-
			0.016	0.140	0.066	0.790	0.309	0.558	0.000
		fine-tuned	0.371	1.270	0.492	34.604	0.782	0.759	0.183
			0.180	0.590	0.291	16.499	0.658	0.664	0.075
Italian	*Meta-Llama-3-8B-Instruct*	pre-trained	-	-	-	-	-	-	-
			0.035	0.231	0.095	1.969	0.455	0.644	0.000
		fine-tuned	**0.339**	**1.337**	**0.448**	**33.994**	**0.651**	**0.799**	**0.158**
			0.085	**0.354**	**0.168**	**8.830**	**0.480**	**0.703**	**0.015**
	Minerva-7B-instruct-v1.0 Pre-trained on Italian (Orlando et al., 2024)	pre-trained	-	-	-	-	-	-	-
			0.036	0.204	0.081	2.183	0.345	0.644	0.000
		fine-tuned	0.040	0.245	0.080	3.262	0.204	0.612	0.000
			0.021	0.148	0.054	1.368	0.200	0.604	0.000

Table 5: Preliminary evaluation in the monolingual setting for Swedish, Dutch, and Italian. For each language, we report the performance obtained for each evaluation metric on the **Seen Test** set (top) and the **Unseen Test** set (bottom). For each language, we highlight in bold the best results for each metric on both the **Seen** and **Unseen Test** sets. Since the language-specific models consistently underperformed compared to the pre-trained and fine-tuned Llama models, we decided to reduce the computational cost of our study by focusing exclusively on Llama models for our study.

J Performance of fine-tuned models in the monolingual setting

We report in Table 6 the performance of the fine-tuned Llama3Instruct models in the monolingual setting across all metrics and languages.

The highest values for BERTScore on the **Seen Test** set are for German (0.884), English (0.857), Latin (0.841), Swedish (0.834), and Chinese (0.832). However, for the **Unseen Test** set, a different set of languages has the highest BERTScore, namely Kurdish (0.800), Spanish (0.750), and Italian (0.702) followed by German (0.696) and Latin (0.688).

For BLEU, we observed strong performance for English (0.635) and German (0.631), followed by Latin (0.505), Swedish (0.487), and French (0.465) on the **Seen Test** set. However, performance on the **Unseen Test** set remains considerably lower across languages, but remain higher than the baseline.

Considering xCOMET, we observed the highest scores for German (0.885 on Seen, 0.708 on Unseen) and English (0.832 / 0.637), followed closely by Dutch (0.814 / 0.690) and Swedish (0.754 / 0.568), indicating consistently strong performance across both test settings for these languages.

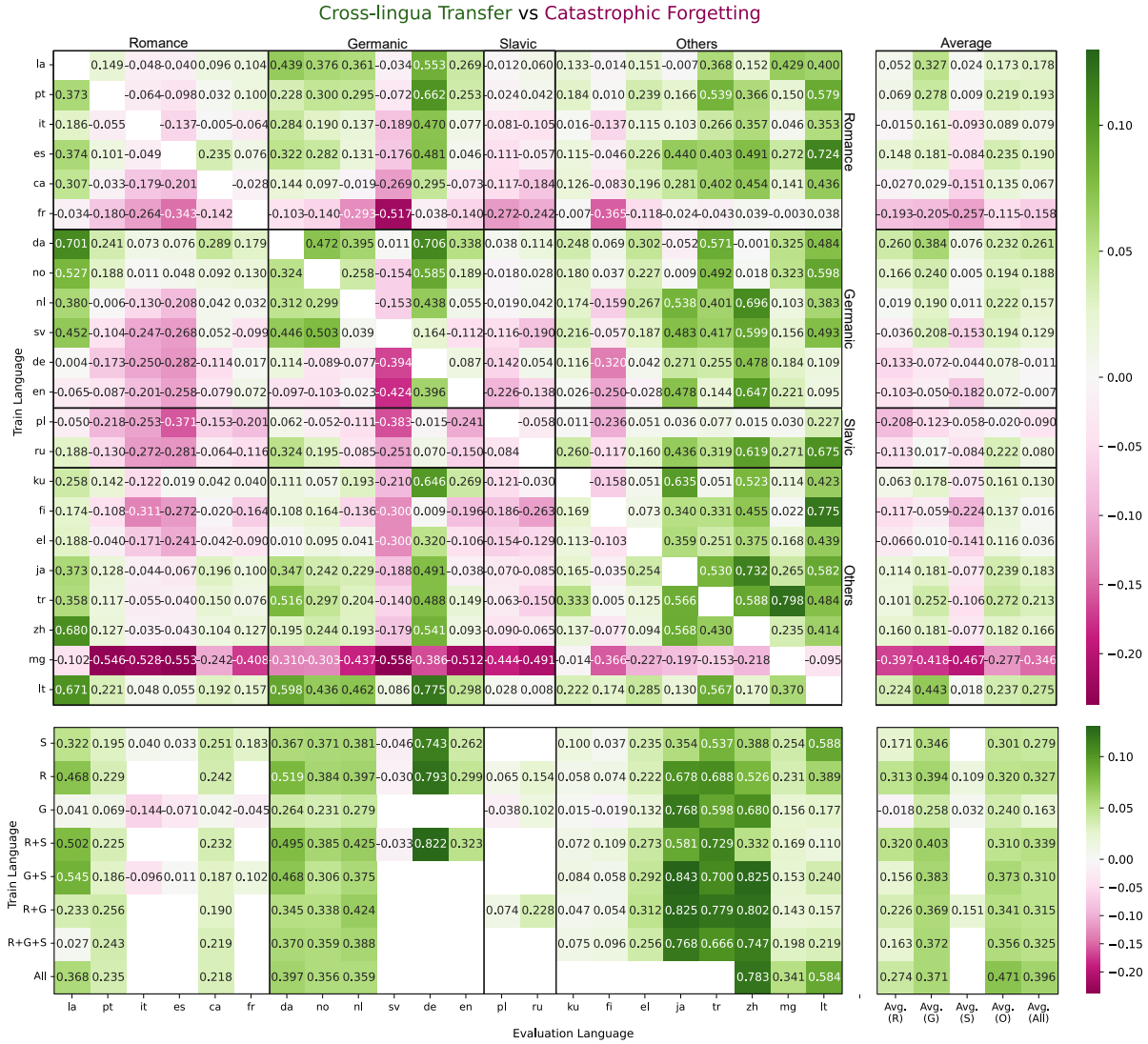

Figure 6: Transfer vs. Forgetting in Cross-lingual Definition Generation. This matrix displays the Multilingual Advantage (MLA) for models fine-tuned on various source languages (rows) and evaluated on unseen target languages (columns). Positive MLA values (green) indicate successful transfer (i.e., performance gains over the pre-trained baseline), whereas negative MLA values (pink) indicate forgetting (i.e., performance deteriorations). The y-axis also reports the number of training examples for models fine-tuned on each source language. Note that training sizes were not controlled in the monolingual setting but were fixed in the multilingual fine-tuning experiments. White cells indicate language pairs where the model was both trained and evaluated on the same language; these overlap with monolingual and multilingual settings and are thus excluded from cross-lingual comparisons.

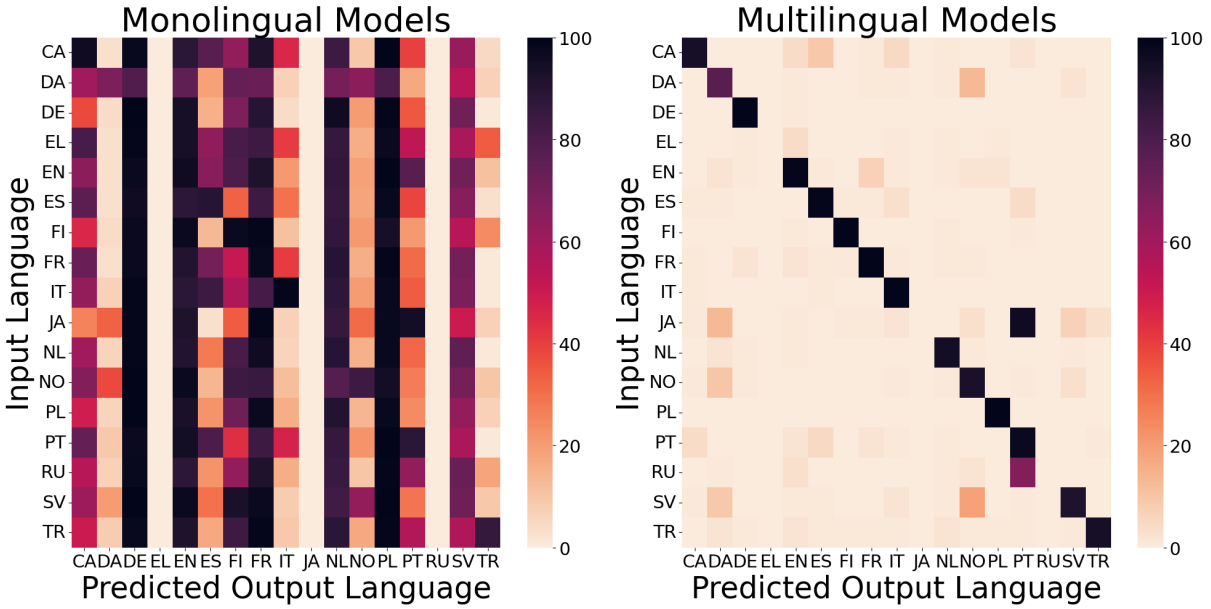

Figure 7: Confusion matrices showing detected output languages for monolingual (left) and multilingual (right) models. Rows indicate the true input language, and columns show the output language as identified by langdetect. Each cell displays the maximum number of outputs (across all models in that category) where the input language was mapped to the detected output language. This highlights the strongest language mismatches observed. Latin and Kurdish are excluded as they are not supported by langdetect.

ISO	Language	Family	BERT	BLEU	NIST	ROUGE-L	SacreBLEU	xCOMET	BERTScore	Ex. Match
en	English	Germanic	*bert-base-uncased* (Devlin et al., 2019)	0.636	2.028	0.703	60.437	0.832	0.857	0.475
				0.144	0.501	0.277	11.793	0.637	0.837	0.030
fr	French	Romance	*bert-base-french-europeana-cased*	0.465	1.473	0.567	45.022	0.714	0.768	0.313
				0.149	0.476	0.253	13.584	0.523	0.595	0.049
de	German	Germanic	*bert-base-german-cased*	0.631	2.063	0.706	61.606	0.885	0.885	0.487
				0.139	0.428	0.229	12.203	0.709	0.697	0.046
ru	Russian	Slavic	*rubert-base-cased* (Kuratov and Arkhipov, 2019)	0.138	0.367	0.003	13.841	0.505	0.493	0.065
				0.191	0.289	0.001	9.573	0.535	0.502	0.149
sv	Swedish	Germanic	*bert-base-swedish-cased*	0.487	1.525	0.593	45.267	0.754	0.834	0.313
				0.116	0.354	0.215	9.805	0.568	0.685	0.037
nl	Dutch	Germanic	*bert-base-dutch-cased* (de Vries et al., 2019)	0.464	1.527	0.560	43.866	0.814	0.793	0.313
				0.214	0.700	0.311	19.635	0.691	0.678	0.117
pl	Polish	Slavic	*bert-base-polish-uncased-v1*	0.422	1.205	0.534	38.788	0.695	0.787	0.254
				0.177	0.496	0.283	17.927	0.559	0.680	0.065
ku	Kurdish	Iranian	*xlm-roberta-base* (Conneau et al., 2020)	0.039	0.147	0.107	3.716	0.307	0.000	0.123
				0.105	0.142	0.198	3.413	0.413	0.000	0.023
fi	Finnish	Uralic	*bert-base-finnish-cased-v1* (Virtanen et al., 2019)	0.326	0.814	0.403	29.492	0.632	0.703	0.219
				0.048	0.097	0.103	5.831	0.480	0.556	0.009
el	Greek	Hellenic	*bert-base-greek-uncased-v1* (Koutsikakis et al., 2020)	0.182	0.661	0.007	17.355	0.516	0.559	0.334
				0.068	0.252	0.001	6.313	0.441	0.479	0.205
ja	Japanese	Japonic	*bert-base-japanese*	0.363	0.316	0.029	24.861	0.698	0.820	0.274
				0.043	0.047	0.009	6.696	0.533	0.681	0.007
ca	Catalan	Romance	*roberta-base-ca* (Armengol-Estapé et al., 2021)	0.360	0.972	0.451	33.779	0.660	0.726	0.234
				0.745	2.621	0.765	74.392	0.831	0.883	0.713
es	Spanish	Romance	*bert-base-spanish-wwm-cased* (Cañete et al., 2020)	0.347	1.219	0.445	32.294	0.683	0.827	0.200
				0.112	0.430	0.199	8.893	0.553	0.750	0.027
it	Italian	Romance	*bert-base-italian-uncased*	0.339	1.337	0.448	33.994	0.651	0.799	0.158
				0.085	0.354	0.168	8.830	0.480	0.703	0.015
tr	Turkish	Turkic	*bert-base-turkish-cased*	0.438	0.223	0.010	19.643	0.738	0.832	0.354
				0.046	0.036	0.006	7.826	0.560	0.657	0.012
pt	Portuguese	Romance	*bert-base-portuguese-cased* (Souza et al., 2020)	0.235	0.696	0.327	21.196	0.603	0.598	0.128
				0.075	0.206	0.154	7.291	0.580	0.507	0.018
zh	Chinese	Sino-Tibetan	*bert-base-chinese*	0.438	0.223	0.010	19.643	0.738	0.832	0.354
				0.046	0.036	0.006	7.826	0.560	0.657	0.012
no	Norwegian	Germanic	*nb-bert-base*	0.356	1.006	0.432	35.334	0.628	0.636	0.256
				0.090	0.240	0.150	9.305	0.445	0.441	0.035
mg	Malagasy	Austronesian	*xlm-roberta-base* (Conneau et al., 2020)	-	-	-	-	-	-	-
				0.084	0.215	0.175	8.208	0.220	0.578	0.000
da	Danish	Germanic	*danish-bert-botxo*	0.323	1.046	0.443	31.414	0.666	0.655	0.110
				0.058	0.174	0.131	5.584	0.528	0.468	0.000
lt	Lithuanian	Baltic	*litlat-bert*	0.100	0.088	0.082	6.201	0.394	0.605	0.077
				0.114	0.226	0.156	14.641	0.467	0.632	0.040
la	Latin	Romance	*bamman-burns-latin-bert* (Bamman and Burns, 2020)	0.506	1.439	0.575	50.337	0.667	0.841	0.391
				0.067	0.214	0.079	9.601	0.368	0.689	0.038

Table 6: Definition generation performance in the monolingual setting. For each language, we report its ISO code, language family, the BERT model used for BERTScore, and the performance obtained for each evaluation metric on the **Seen Test** set (top) and the **Unseen Test** set (bottom). For xCOMET, the same model was used across all languages: *XCOMET-XL*.