Evaluation of the Cross-lingual Embedding Models from the Lexicographic Perspective

Michaela Denisová[1], **Pavel Rychlý**[2]

[12]Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Brno, Czech Republic
E-mail: [1]449884@mail.muni.cz, [2]pary@fi.muni.cz

Abstract

Cross-lingual embedding models (CMs) enable us to transfer lexical knowledge across languages. Therefore, they represent a useful approach for retrieving translation equivalents in lexicography. However, these models have been mainly oriented towards the natural language processing (NLP) field, lacking proper evaluation with error evaluation datasets that were compiled automatically. This causes discrepancies between models hindering the correct interpretation of the results. In this paper, we aim to address these issues and make these models more accessible for lexicography by evaluating them from a lexicographic point of view. We evaluate three benchmark CMs on three diverse language pairs: close, distant, and different script languages. Additionally, we propose key parameters that the evaluation dataset should include to meet lexicographic needs, have reproducible results, accurately reflect the performance, and set appropriate parameters during training. Our code and evaluation datasets are publicly available.[1]

Keywords: cross-lingual embedding models; bilingual lexicon induction task; retrieving translation equivalents; evaluation

1. Introduction

Over the years, cross-lingual embedding models (CMs) have drawn much attraction due to their ability to transfer lexical knowledge across languages. CMs facilitate the alignment of word vector representations of two or more languages into one shared space where similar words obtain similar vectors (Ruder et al., 2019).

These models are appealing for lexicography for multiple reasons. Firstly, the translation equivalents candidates can be extracted from the shared space through the nearest neighbour search. Secondly, unlike parallel data-based methods for finding translation equivalents candidates, they require only comparable data, i.e., comparable corpora. Comparable corpora are often available for low-resource languages or rare language combinations and are balanced in the texts they consist of. Finally, CMs are an active research area increasing the number of papers published constantly and expected to develop and improve continuously.

In the natural language processing field, finding translation equivalents candidates is referred to as bilingual lexicon induction (BLI) task. In the BLI task, the target language words are induced from shared space through the nearest neighbour search for a source

[1] https://github.com/x-mia/Evaluation_of_CWE

language word. Afterwards, they are run against a gold-standard dictionary to measure the quality of the model (Ruder et al., 2019).

The BLI task is a popular way among researchers to evaluate their models (Artetxe et al., 2016; Conneau et al., 2017; Joulin et al., 2018; Glavaš & Vulić, 2020; Parizi & Cook, 2021; Tian et al., 2022: etc.). However, the evaluation is often inconsistent and differs from paper to paper, using various metrics and gold-standard dictionaries from multiple sources (Ren et al., 2020; Karan et al., 2020; Woller et al., 2021; Severini et al., 2022: etc.). This impedes our ability to correctly interpret the results and make models comparable to each other.

Moreover, many currently used gold-standard dictionaries are generated automatically (Conneau et al., 2017; Glavaš et al., 2019; Vulić et al., 2019: etc.). Therefore, they are prone to contain mistakes. For example, the most widely used gold-standard dictionaries, MUSE (Conneau et al., 2017), are criticised for occurring errors and disproportional part-of-speech distribution (Kementchedjhieva et al., 2019; Denisová & Rychlý, 2021).

On top of that, articles dealing with CMs and the BLI task do not consider the utilisation in the lexicography field. They focus on the computational side of the problem and simple word-to-word extraction without reflecting on various aspects of translation and tailoring the evaluation process and gold-standard dictionaries to the lexicographers' needs.

In this paper, we investigate various aspects that influence the training of the CMs. We propose the most suitable parameters for the evaluation dataset based on these aspects while allowing for a lexicography perspective. We show that having a strong evaluation dataset and a clear evaluation process is crucial for setting appropriate training parameters. We assess and discuss the quality of the most common benchmark models on a distant language pair, Estonian-Slovak, a close language pair, Czech-Slovak, and language pair that do not share a script, English-Korean.

Our motivation is to determine important aspects when evaluating CMs on the BLI task and construct a reliable, high-quality evaluation dataset that addresses the above-stated issues. Our contribution is manifold:

1. We set crucial parameters of the evaluation dataset for the BLI task that are reproducible for further research and unifying for the evaluation process.
2. We provide an evaluation of three diverse language pairs on the most cited benchmark CMs in various settings.
3. We link the NLP and lexicography sides of the CMs' evaluation.

This paper is structured as follows. In Section 2, we outline the background information and the benchmark models used for the evaluation. In Section 3, we describe the experimental setup used in training CMs. In Section 4, we list important aspects of the evaluation and provide reasoning for each of them. In Section 5, we offer concluding remarks.

2. Background

In this paper, we train and evaluate three methods: MUSE (Conneau et al., 2017; Lample et al., 2017: two equal articles), VecMap (Artetxe et al., 2016, 2017, 2018a,b), and FastText for multilingual alignment (Joulin et al., 2018). These methods are frequently cited and utilised as benchmarks.

MUSE. This method connects domain-adversarial training with iterative Procrustes alignment. Moreover, it proposes a novel method for matching translation equivalents candidates, cross-domain similarity local scaling (CSLS). MUSE involves supervised and unsupervised training and training that relies on identical strings. Their code, pre-trained multilingual word embeddings and datasets used for training and evaluation are available on their GitHub repository.[2] Evaluation datasets were made automatically for 110 languages.

VecMap. VecMap is a robust self-learning framework with multiple steps and iterative learning depending on the setting. It can be trained in a supervised, semi-supervised, and unsupervised manner or uses identical strings as supervision signals. Similarly to the MUSE framework, it has an open-source GitHub repository.[3]

FastText. The FastText method proposes to optimise the CSLS retrieval criterion used in the MUSE framework. This method provides a supervised setting for training. Their pre-trained aligned models are freely available[4], and their code is published on the GitHub repository.[5]

3. Experimental Setup

CMs require comparable corpora for training. In this case, comparable means non-aligned and similar in size and text genres (Kovář et al., 2016). The comparable corpora are used to train monolingual word embeddings (MEs) incorporated in CM training (Ruder et al., 2019).

In this paper, we experimented with two types of MEs. We used pre-trained FastText MEs (Bojanowski et al., 2017)[6] for Estonian, Slovak, Czech, English, and Korean, which were trained on Wikipedia[7] with dimension 300. The second pre-trained MEs were provided by SketchEngine (Herman, 2021).[8] These embeddings were trained on web corpora using the same method as FastText, with dimensions 100 for Estonian-Slovak and English-Korean, and 300 for Czech-Slovak.

Additionally, the training involves a different level of supervision: supervised, identical-string-relying, or unsupervised (Ruder et al., 2019). In this paper, for MUSE and VecMap, we set supervised and unsupervised settings and mode that relies on identical strings. For FastText, we selected supervised training only.

Methods trained in supervised mode require a word-to-word dataset called a seed lexicon. Word-to-word means one single-word unit to one or multiple single-word units. The size of the seed lexicon usually varies up to 5K word pairs. Exceeding this limit does not influence the resulting quality (Vulić & Korhonen, 2016). In identical-string-relying mode, the seed lexicon consists of identical strings and numerals that occur in MEs of both languages.

[2] https://github.com/facebookresearch/MUSE
[3] https://github.com/artetxem/vecmap
[4] https://fasttext.cc/docs/en/aligned-vectors.html
[5] https://github.com/facebookresearch/fastText/tree/master/alignment/
[6] https://fasttext.cc/
[7] https://www.wikipedia.org/
[8] https://embeddings.sketchengine.eu/

Seed lexicons were from various resources. We used the Estonian-Slovak database that Denisová (2021) constructed for the Estonian-Slovak language combination. We selected 5,000 word-to-word translation equivalents from this database, where the source and target language word occurred in the first 300K words of ME files. As this database's accuracy is only 40%, we manually post-processed selected translation equivalents.

Czech and Slovak are very close languages containing a lot of identical words. Therefore, we matched 5K word-to-word identically spelt translation equivalents which occurred in the first 300K words of ME files. Lastly, we used the MUSE English-Korean training dataset provided by Conneau et al. (2017) as a seed lexicon for the last language pair, English-Korean.

The last crucial parameter in the training setup is the number of word embeddings loaded during training. This parameter influences the vocabulary coverage of the resulting aligned translation equivalents candidates and, therefore, the vocabulary selection for the evaluation dataset. We reason this in Section 4.1.

Among researchers, the standard is to load the first 200K embeddings. In this paper, we experimented with different numbers of loaded embeddings, which we describe in Section 4.1 in further detail.

Additionally, when assessing the models in the evaluation process, we utilise two metrics, i.e., precision and recall. Precision at k (P@k) is the proportion of the number of correct translation equivalents to the number of all extracted translation equivalents' candidates, where k is the amount of extracted target language words for each source language word (Kementchedjhieva et al., 2019). In this paper, the most common is P@*10*, meaning we extract ten target language words for each source language word from the evaluation dataset.

The recall is the proportion of the correct translation equivalents found to the number of all translation equivalents from the evaluation dataset. In this article, we focus mainly on computing recall since this is a more important metric in lexicography. Therefore, the most common number for the induced target language words is ten. However, when assessing the models for language learners, precision is preferred.

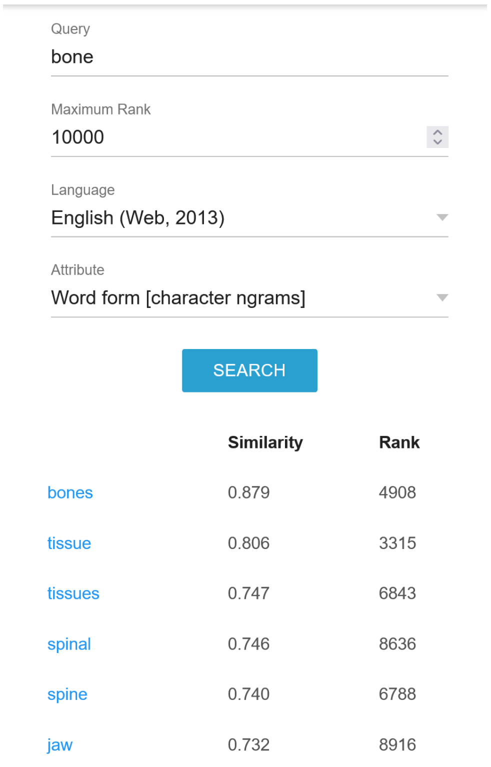
4. Parameters of the Evaluation Dictionary

In this section, we investigate which factors significantly influence the evaluation and training processes. Each subsection discusses different aspects.

4.1 Vocabulary

MEs used in training significantly influence the resulting quality of the aligned cross-lingual spaces (Artetxe et al., 2018c; Vulić et al., 2020). From the vocabulary perspective, the MEs impact the nature of the words that embeddings contain and the size of the resulting dictionary.

These two factors depend on the domain (Søgaard et al., 2018) and the size of the monolingual corpus that MEs have been trained on. Since we do not assess the quality of the MEs, we are restricted to the words they contain. Therefore, we should include only these words in the evaluation dataset to avoid out-of-the-vocabulary (OOV) words, i.e., words that do not occur in the MEs.



Query
bone

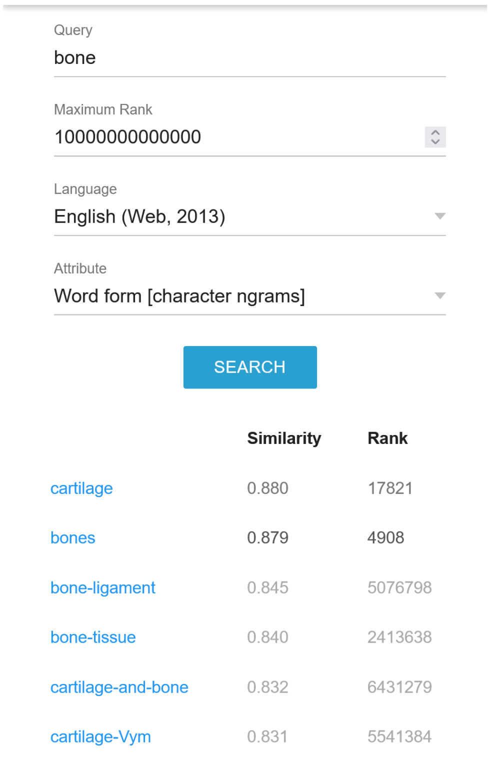
Maximum Rank
10000

Language
English (Web, 2013)

Attribute
Word form [character ngrams]

	Similarity	Rank
bones	0.879	4908
tissue	0.806	3315
tissues	0.747	6843
spinal	0.746	8636
spine	0.740	6788
jaw	0.732	8916

Figure 1: Search for the word *bone* with the SketchEngine tool for monolingual word embeddings with a word rank of 10,000.



Query
bone

Maximum Rank
10000000000000

Language
English (Web, 2013)

Attribute
Word form [character ngrams]

	Similarity	Rank
cartilage	0.880	17821
bones	0.879	4908
bone-ligament	0.845	5076798
bone-tissue	0.840	2413638
cartilage-and-bone	0.832	6431279
cartilage-Vym	0.831	5541384

Figure 2: Search for the word *bone* with the SketchEngine tool for monolingual word embeddings with a word rank of 10,000,000,000,000.

OOV words play a significant role during the evaluation process. They are words that do not occur in the shared cross-lingual word embedding space or the MEs. The reasons behind this are various: they are excluded because they had very low or zero occurrences in the monolingual corpus, the MEs contain only the most frequent words, they consist of multiple words (multi-word expressions), or the CM loaded and aligned only a certain amount of the MEs during the training.

Importantly, the MEs and cross-lingual word embedding models do not handle multi-word expressions and words that do not have a one-word equivalent in the target language (e.g., German word *Grundschule* - *primary school, elementary school*). Therefore, we should not include such words in the evaluation dataset.

However, setting the number of loaded embeddings parameter allows us to increase vocabulary coverage for the evaluation dataset and better reflect the resulting quality of the model. Specifically, we have various numbers of the words in the MEs we utilised for training. FastText contains 329,987 Estonian words, 316,098 Slovak words, 627,841 Czech words, 2,519,370 English words, and 879,129 Korean words. For SketchEngine, the Estonian model has 3,307,785 words, Slovak 1,611,402 words, Czech 3,900,455 words, English 6,658,558 words, and Korean 2,949,340 words.

Naturally, a more extensive corpus produces more words, and more words mean greater coverage. However, our goal is not to have as many words as possible at the expense of the quality of the aligned word embeddings. The disadvantage of training with such huge monolingual embeddings is that it is computationally expensive and time-consuming. Moreover, including less frequent words (words with higher rank) does not necessarily mean better results when extracting translation equivalent candidates based on their cosine similarity. Using the word *bone* as an example, Fig. 1 and 2 show that we get more relevant searches if we limit the word rank to a smaller number.

For demonstration purposes and to be able to compare MEs with each other, we constructed the evaluation datasets for Estonian-Slovak and Czech-Slovak that include words occurring in the MEs and OOV words together. The Estonian-Slovak evaluation dataset was compiled using the Estonian-Slovak dictionary from Denisová (2021), similarly to the seed lexicon's compilation. The evaluation dataset for Czech-Slovak was constructed manually using exclusively words that are different in both languages (e.g., *želva* - *korytnačka*, *turtle*). Notably, the evaluation datasets need to differ from the seed lexicons.

For English-Korean, we used the open-source evaluation dataset MUSE (Conneau et al., 2017), which includes only words occurring in the MEs. While aware of this dataset's drawbacks (Kementchedjhieva et al., 2019; Denisová & Rychlý, 2021), we chose it intentionally to help us demonstrate the crucial parameters of the evaluation dataset. Each evaluation dataset contains 1,500 headwords.

We trained the models using the default (or standard) loaded embedding parameter in this experiment. Afterwards, we adjusted it to be optimal considering computational time, the resulting quality and vocabulary coverage. The recall for default and adjusted training is displayed in Table 1.

Given Table 1, each model improved recall for Estonian-Slovak and Czech-Slovak by 10-20% if we increase the number of loaded embeddings from 50K to 300-400K. Generally, the SketchEngine embeddings for Estonian-Slovak appeared to perform worse than FastText when 50K embeddings were loaded. However, after adjusting the loaded embeddings' parameter, their recall increased, surpassing the models trained with FastText embeddings.

Although the English-Korean evaluation dataset contained words from the first 50K loaded embeddings, the recall for the models trained with FastText embeddings decreased in most cases. It shows that enlarging the number of loaded embeddings in this particular scenario can have a negative impact on recall. As mentioned above, increasing the word rank can include more noise from the MEs and, thus, lower the resulting quality. This is the indicator of the quality of the MEs, not the CMs.

Additionally, the SketchEngine embeddings outperformed FastText embeddings in the majority of cases, except for English-Korean, where FastText embeddings were significantly better. This could be due to the uneven part-of-speech distribution (Kementchedjhieva et al., 2019; Denisová & Rychlý, 2021). Therefore, we constructed a new evaluation dataset for English-Korean to compare the results. We discuss this problem in Section 4.3 in further detail.

Although we changed the parameter, some OOV words from our evaluation dataset remain, except for the English-Korean language, where all selected words for the evaluation were among the first 50K words in the monolingual embeddings.

FastText/ SketchEngine (%)	50K loaded			300-400K loaded		
	ET-SK	CZ-SK	EN-KO	ET-SK	CZ-SK	EN-KO
MUSE-S	19.33 20.00	57.84 70.94	39.97 31.01	27.86 42.40	68.73 78.95	29.98 34.14
MUSE-I	19.26 19.40	57.91 **71.00**	39.00 28.41	25.80 38.93	68.73 79.02	23.17 29.65
MUSE-U	19.80 18.80	58.58 **71.00**	36.46 26.14	24.46 34.80	69.13 79.02	24.58 25.33
VecMap-S	20.73 20.33	58.24 70.67	**50.51** 32.52	34.93 **51.86**	69.73 79.02	49.00 35.44
VecMap-I	21.00 19.20	59.05 **71.00**	**41.59** 28.63	34.73 46.00	71.87 **80.09**	33.98 29.87
VecMap-U	**21.20** 18.86	58.98 70.67	36.35 21.93	33.53 44.80	71.94 **80.09**	29.76 12.42
FastText	20.60 21.06	57.51 70.54	50.40 26.90	31.93 49.33	67.93 78.28	**51.91** 37.60

Table 1: The recall of models before and after changing the parameter for loaded embeddings (Supervised: MUSE-S, VecMap-S, FastText; Identical: MUSE-I, VecMap-I; Unsupervised: MUSE-U, VecMap-U).

There were 331 Estonian OOV words in the Estonian-Slovak language combination trained with FastText (e.g., *aedvili* – *vegetable*, *ellu jäämа* – *to stay alive*, *mürgitama* – *to poison*, etc.) and 40 when trained with SketchEngine (e.g., *enne kui* – *before*, *buteen* – *butene*, *uusik* – *newcomer*, etc.). In Czech-Slovak trained with FastText, there were 11 Czech OOV words (e.g., *cáklý* – *crazy*, *mlsný* – *sweet tooth*, *slušet* – *to suit*, etc.). For SketchEngine, there was 1 OOV word containing a spelling mistake: *onemocněť*, correctly *onemocnět* (*to get ill*). Fig. 3 and 4 show the frequency distribution in the monolingual corpus of the Estonian and Czech OOV words, respectively. The number of occurrences represents the frequency of the OOV words from the monolingual corpus, and word pair rank corresponds to the number of the OOV words.

Finally, we showed that selecting a vocabulary for the evaluation dataset is crucial for setting the number of loaded embeddings during training and mirroring the model's quality more accurately. The evaluation dataset should consist of the words occurring in the loaded word embeddings from MEs. Moreover, the number of loaded MEs should not exceed the highest word rank in the evaluation dataset. And it should omit OOV words and multi-word expressions, as we do not assess the quality of the MEs.

4.2 Inflected Word Forms

Another essential factor to allow for when constructing an evaluation dataset is inflected word forms. MEs are trained on the corpus where words occur in context, not necessarily

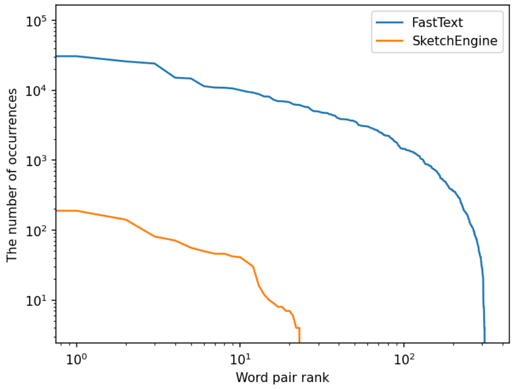


Figure 3: Comparison of the Estonian OOV word pair rank from FastText and SketchEngine MEs and the number of their occurrences from Estonian National Corpus 2017.

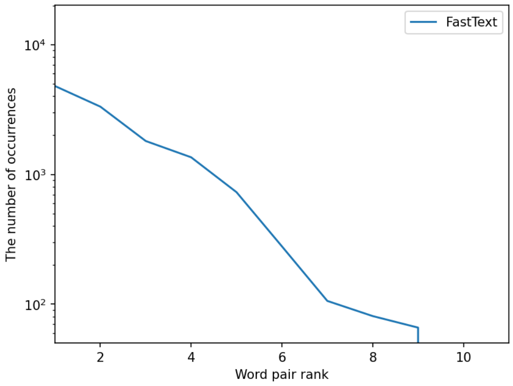


Figure 4: Comparison of the Czech OOV word pair rank from FastText and the number of their occurrences from Czech National Corpus 2017 (SketchEngine had only one OOV with 0 occurrences in the corpus).

in their basic form. Therefore, more common word forms from the text appear in the MEs and subsequently in the CMs.

For example, suppose we extract translation equivalent candidates for the Estonian word *tund* (*hour*) with Slovak as a target language. In that case, we get various word forms such as *hodiny*, *hodinu*, *hodín*, and *hodina*, from which only the last one is the basic word form of this word.

We often seek the word's basic form for the dictionary, although it is not the most common form that appears in the texts. However, if we do not consider morphological variance and include only basic word forms of the source and target language word pair, i.e., *tund* = *hodina*, all other forms would be counted as an error even though their meaning is the same. Moreover, ignoring the morphological variance results in an inaccurate model recall and precision.

In our experiments, we applied Slovak lemmatiser Majka[9] on the extracted translation equivalent candidates for Estonian-Slovak, and Czech-Slovak language combinations to create the basic word form of each word. This caused duplicate translation equivalent candidates. For instance, instead of *hodiny*, *hodinu*, *hodín*, and *hodina*, we had four times the word *hodina*. When counting the recall, we counted this as one translation equivalent candidate. We utilised the same evaluation datasets from Section 4.1.

Afterwards, we compared the recall before and after lemmatisation. The results are displayed in Table 2.

Given Table 2, we can see that the recall for each model increased by approximately 1-7%. Thus, we get more accurate results for the model when allowing for morphological variance either by lemmatising the results or including various word forms in the evaluation dataset.

[9] https://nlp.fi.muni.cz/czech-morphology-analyser/

Non-lemmatised/ Lemmatised (%)	FastText MEs		SketchEngine MEs	
	ET	CZ	ET	CZ
MUSE-S	27.86/ 29.40	68.73/ 70.80	42.40/ 45.00	78.95/ 79.75
MUSE-I	25.80/ 27.86	68.73/ 70.94	38.93/ 42.60	79.02/ 79.89
MUSE-U	24.46/ 28.80	69.13/ 71.20	34.80/ 41.93	79.02/ 79.82
VecMap-S	34.93/ 35.93	69.73/ 71.74	51.86/ 52.93	79.02/ 79.82
VecMap-I	34.73/ 36.06	71.87/ 72.94	46.00/ 50.86	80.09/ 80.96
VecMap-U	33.53/ 35.20	71.94/ 73.01	44.80/ 50.26	80.09/ 80.96
FastText	31.93/ 32.06	67.93 /70.14	49.33/ 50.53	78.28/ 79.09

Table 2: The recall of models before and after lemmatisation.

4.3 Part of Speech (POS)

In this section, we discuss whether including various word pairs from all POS groups is necessary or whether a more relevant POS can adequately mirror the models' performance. Moreover, we show the proportion of the POS in the evaluation datasets we used and the performance of the models on various POS.

The selection of POS of the word pairs is the central topic of the articles that critically examine the evaluation datasets for the BLI task. For instance, the analysis of the POS distribution in the MUSE evaluation datasets (Conneau et al., 2017) conducted by Kementchedjhieva et al. (2019) revealed that these datasets contain a large number of proper nouns. The authors saw this as a problem since proper nouns do not carry any meaning; therefore, they are not suitable for reflecting the models' performance.

Another effort provided by Izbicki (2022) compiled evaluation datasets for 298 languages with as similar POS distribution as possible across the datasets to make results between models and language pairs comparable. However, every POS was represented in their datasets.

We argue that some POS are less relevant to involve in the evaluation dataset than others. Except for proper nouns, such categories as pronouns, conjunctions, articles, and prepositions cannot accurately reflect the models' performance since they play a syntactic role and their meaning changes within the context or is phrase-depending. Moreover, in many cases, they do not correspond to each other across the languages and are either not translated or translated with more than one word. Therefore, these POS do not suit for evaluating word-to-word translations.

In this section, we examined the POS distribution in the evaluation datasets we used. We automatically annotated the evaluation datasets to analyse the POS distribution. For Estonian, we used EstNLTK[10], an open-source tool for processing the Estonian language. We tagged the Czech dataset with Majka and utilised the NLP tool Polyglot[11] for English. Importantly, these tools are designated to tag words in the context that our datasets were

[10] https://estnltk.github.io/

[11] https://polyglot.readthedocs.io/en/latest/

lacking. Moreover, even for human annotators is challenging to determine the POS of the word, especially when the context is missing. Thus, the results might contain discrepancies.

When we look at the POS distribution in our evaluation datasets, the dataset for Estonian-Slovak was disproportional as it contains many nouns, while other POS have significantly smaller representations. This dataset was derived from the Estonian-Slovak dictionary (Denisová, 2021), which consists mainly of nouns. On the other hand, the POS was distributed more evenly in the Czech-Slovak evaluation dataset, which was constructed manually. Fig. 5 and 6 display the graphs of the POS distribution in these datasets.[12]

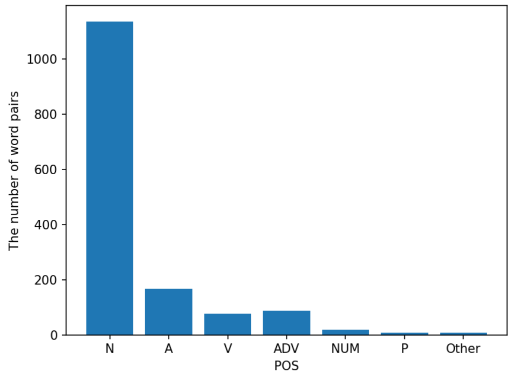


Figure 5: The POS distribution in the Estonian-Slovak evaluation dataset.

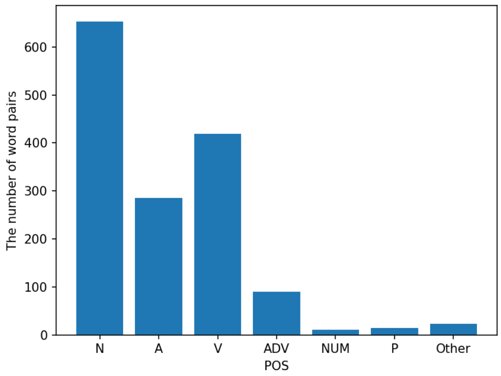


Figure 6: The POS distribution in the Czech-Slovak evaluation dataset.

We utilised the MUSE evaluation dataset (Conneau et al., 2017) to assess the English-Korean language combination. These datasets contain many proper nouns, which is confirmed by the graph in Fig. 7. However, after a manual check, we discovered that some nouns were incorrectly tagged as proper nouns. Moreover, a significant group of words was tagged with the symbol X (in the graph marked as "other") when the tagger could not identify the POS of the current word.

We compiled a new English-Korean dataset with different POS distributions. We sampled word pairs from the English-Korean dictionary. This dictionary was created with the bilingual SketchEngine tool and post-processed manually (Kovář et al., 2016). This assumes the correctness of the translation equivalents (in contrast to the automatically compiled evaluation dataset as MUSE is). We intentionally avoided involving proper nouns, pronouns, articles, conjunctions, and prepositions. Fig. 8 provides the graph of the POS distribution.

In the next step, we computed recall for both evaluation datasets. We set the same conditions for both datasets: no OOV words, around 1,500 headwords in the dataset, and 400K loaded embeddings. Table 3 outlines the results.

Table 3 shows that recall for models trained with FastText MEs dropped drastically. On the other hand, the models trained with SketchEngine MEs did not decrease significantly.

[12] N = nouns; A = adjectives; V = verbs; ADV = adverbs; NUM = numerals; P = pronouns; PN = proper nouns; Other = conjunctions, interjections, prepositions, unknown, etc.

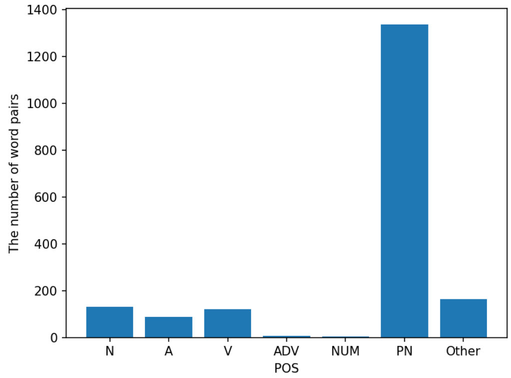


Figure 7: The POS distribution in the English-Korean MUSE evaluation dataset.

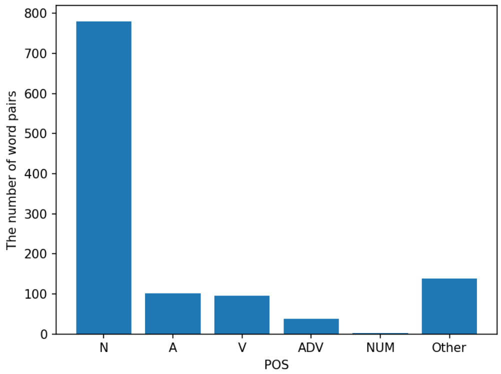


Figure 8: The POS distribution in the English-Korean SketchEngine evaluation dataset.

(%)	MUSE dataset		SketchEngine dataset	
	FastText MEs	SketchEngine MEs	FastText MEs	SketchEngine MEs
MUSE-S	29.98	34.14	20.58	31.83
MUSE-I	23.17	29.65	19.89	24.56
MUSE-U	24.58	25.33	19.37	23.18
VecMap-S	49.00	35.44	**29.41**	36.24
VecMap-I	33.98	29.87	23.44	25.25
VecMap-U	29.76	12.42	22.31	15.22
FastText	**51.91**	**37.60**	28.37	**37.80**

Table 3: The comparison between the results when using the MUSE and SketchEngine evaluation datasets for English-Korean.

However, this changing recall could also result from removing other errors that the MUSE evaluation dataset contains, such as words from different languages, abbreviations, nonsense words, etc. (Denisová & Rychlý, 2021).

The gap between the VecMap trained in a supervised mode evaluated on MUSE and SketchEngine datasets is almost 20%. We investigated some examples of the found and not found word pairs from the evaluation dataset.

A closer look revealed that VecMap likely found a correct equivalent for proper nouns, such as *Abdullah*, *Alexandra*, *Cambodia*, *Cameroon*, *Helsinki*, etc., which made up a significant group in the MUSE dataset, but they had no representation in the SketchEngine dataset.

Furthermore, the VecMap was good at finding equivalents for international words, for example, *algebra*, *alias*, *android*, *idol*, *email*, etc. As in the previous example, these words occurred more frequently in the MUSE than in the SketchEngine dataset.

When looking at the not found words, some were caused by mistakes in the evaluation dataset. For example, in the MUSE dataset were word pairs that consisted of the same word, i.e., *android – android*. In the MUSE dataset occurred words with multiple translation equivalents, from which one was either wrong (*Yemen* translated as *South Yemen*) or VecMap could find only one of them (*fence*).

In the SketchEngine dataset, we observed various words for foods, animals, or numbers, such as *bean*, *tea*, *mudfish*, *rooster*, *two*, *fifty*, etc., for which VecMap could not find an appropriate equivalent but found word that had a similar lexical-semantic relationship (e.g., *tea – coffee*, *fifty – fourteen*, etc.). On the other hand, the MUSE dataset lacks such words. Moreover, the SketchEngine dataset contains more verbs than the MUSE dataset, a problematic group for VecMap to find an equivalent for (see Table 4).

Finally, we computed the recall for the VecMap supervised model for each POS separately and for all language pairs to observe how the results change. Table 4 displays the results.

VecMap-S (%) FT/SE	ET-SK	CZ-SK	EN-KO
Nouns	31.89/ 48.54	76.87/ 86.21	48.46/ **50.00**
Adjectives	48.21/ 63.09	73.07/ 77.97	47.19/ 43.82
Verbs	35.52/ 64.47	63.48/ 67.30	35.00/ 21.66
Adverbs	41.37/ 56.32	70.00/ 81.11	**75.00**/ 37.50
Numerals	61.11/ 77.77	**81.81**/ 81.81	25.00/ 25.00
P/ PN	**62.50**/ 75.00	80.00/ **100**	51.00/ 34.25
Others	25.00/ 37.50	69.56/ **100**	43.55/ 39.26

Table 4: The recall of the supervised VecMap for each POS in each language.

According to Table 4, the results for different POS and MEs varied from each other greatly. As for distant language pairs, they achieved relatively high recall for adjectives, adverbs, and pronouns/ proper nouns. Furthermore, both language combinations gained low results for verbs.

However, they differed in the outcomes for numerals and nouns. Estonian-Slovak achieved the highest recall on numerals, whereas in a language that does not share a script, English-Korean, it was the lowest. The other way around it was by the results for nouns.

The close language pair, Czech-Slovak, was able to find all the equivalents from the evaluation dataset for pronouns, and small POS groups (in the table as *Others*), such as conjunctions, interjections, prepositions, etc. The explanation for this is that these two groups have a small representation in the evaluation dataset and a high word rank in MEs, so it was easier for the model to find an equivalent. High results were also achieved for nouns and numerals. Similarly to the distant language pairs, verbs were the weakest group.

The reasons behind these diverse results are manifold. For instance, recall depends on OOV words, so if nouns are the biggest group in the Estonian-Slovak evaluation dataset, they also contain a high number of OOV words. Thus, their recall is relatively low compared to the nouns in the English-Korean evaluation dataset, with no OOV words. Other factors are: how many senses of the word are included in the evaluation dataset (the discussion is provided in Section 4.4), the word rank of the source language words in the MEs, the quality of the MEs and alignment, and the quality of the tagging tool.

Importantly, Table 4 demonstrates the significant impact of the POS distribution in the evaluation dataset on the resulting quality of the model.

4.4 Senses

Another important component when constructing an evaluation dataset is how many senses of one word to include. For example, the English word *band* as a noun has several meanings, such as *musical group*, *piece of cloth*, *range of values*, etc., or it can be a verb. Therefore, if we want the model to find all meanings in the target language, we should induce the same number of translation equivalents' candidates. However, with more extracted translation equivalents candidates comes much noise in the form of various errors, such as words with different POS, words with other lexical-semantic relationships, shortcuts, etc. (Denisová, 2022).

Thus, the precision decreases when the number of extracted target language words is extended, and we need to find the right amount depending on our goal (higher precision or higher recall). In this section, we investigate how the number of extracted target language words impacts precision and recall.

In our experiments, we measured the precision P@*1*, P@*5*, and P@*10*. Moreover, we computed the recall for each stage to compare the results. All models were trained and evaluated under the same conditions as in Section 4.1. Tables 5 and 6 show the outcomes.

Tables 5 and 6 confirm that as the precision increases, the recall drops; reversely, the higher recall, the lower the precision. This means that the more target language words we induce, the higher recall we achieve. However, the precision of our model declines. Therefore, we should set our aim beforehand, whether to use the resulting induced translation equivalents candidates for lexicography, which requires higher recall, or language acquisition that favours precision.

On top of that, the end user is also essential when selecting the nature of the words. For example, when constructing the Estonian-Slovak or English-Korean dictionary for language students, we should focus on frequently used words or words from the basic vocabulary. However, we should select different words rather than mutual when dealing with a close language pair, such as Czech-Slovak. Especially when assessing identical training mode.

5. Conclusion

In this paper, we have evaluated three benchmark CMs in various settings from different points of view on the BLI task. We have used three language pairs for the demonstration, i.e., a distant language pair, Estonian-Slovak, a close language pair, Czech-Slovak, and language pair that does not share a script, English-Korean. We have discussed various parameters that an evaluation dataset should allow for. We showed that these parameters are crucial for reflecting the model's performance precisely and accurately. Moreover, they are vital for setting the parameters in training, such as the number of loaded MEs.

FT MEs	P@*1* (%)			P@*5* (%)			P@*10* (%)		
Pre./ Rec.	ET-SK	CZ-SK	EN-KO	ET-SK	CZ-SK	EN-KO	ET-SK	CZ-SK	EN-KO
MUSE-S	15.06	62.00	15.06	6.05	17.02	6.08	3.61	8.96	3.88
	14.60	47.96	11.61	24.00	65.33	23.44	27.86	68.73	29.98
MUSE-I	12.10	62.52	12.61	5.21	16.95	4.77	3.34	8.96	3.00
	11.73	48.36	9.72	20.66	65.06	18.42	25.80	68.73	23.17
MUSE-U	10.11	63.38	13.03	5.21	16.95	4.84	3.17	9.01	3.18
	9.80	49.03	10.04	20.66	65.06	18.69	24.46	69.13	24.58
VecMap-S	**21.52**	61.22	31.18	7.95	17.15	10.94	**4.53**	9.09	6.35
	20.86	47.36	24.04	31.35	65.86	42.19	**34.93**	69.73	49.00
VecMap-I	18.70	**65.63**	19.83	7.66	17.73	7.31	4.50	9.37	4.40
	18.13	**50.76**	15.28	30.40	68.06	28.20	34.73	71.78	33.98
VecMap-U	16.29	**65.63**	15.76	7.23	**17.75**	6.36	4.35	**9.38**	3.86
	15.80	**50.76**	12.15	28.66	**68.13**	24.52	33.53	**71.94**	29.76
FastText	17.05	59.93	**31.74**	7.02	16.55	**11.46**	4.14	8.85	**6.73**
	16.63	46.35	**24.44**	27.86	63.52	**44.19**	31.93	67.93	**51.91**

Table 5: The precision (pre.) P@*1*, P@*5*, and P@*10* and recall (rec.) for the models trained with FastText (FT) MEs.

SE MEs	P@*1* (%)			P@*5* (%)			P@*10* (%)		
Pre./ Rec.	ET-SK	CZ-SK	EN-KO	ET-SK	CZ-SK	EN-KO	ET-SK	CZ-SK	EN-KO
MUSE-S	25.51	**72.88**	18.78	7.96	19.82	7.17	4.48	10.24	4.42
	24.73	**56.37**	14.47	37.73	76.41	27.66	42.40	78.95	34.14
MUSE-I	22.55	72.62	12.89	7.24	19.80	6.16	4.11	10.25	3.84
	21.86	56.17	9.94	34.33	76.35	23.77	38.93	79.02	29.65
MUSE-U	19.60	72.71	10.44	6.52	19.82	5.26	3.68	10.25	3.28
	19.00	56.24	8.04	30.93	75.41	20.31	34.80	79.02	25.33
VecMap-S	**32.11**	72.19	21.30	**9.74**	19.68	7.61	**5.48**	10.25	4.59
	31.13	55.84	16.42	**46.20**	75.88	29.33	**51.86**	79.02	35.44
VecMap-I	24.33	72.36	13.66	8.35	**19.91**	6.12	4.86	**10.39**	3.87
	23.60	55.97	10.53	39.60	**76.75**	23.60	46.00	**80.09**	29.87
VecMap-U	24.20	72.45	2.91	8.17	**19.91**	2.27	4.73	**10.39**	1.61
	23.46	56.04	3.78	38.73	**76.75**	8.75	44.80	**80.09**	12.42
FastText	28.74	72.79	**22.21**	9.01	19.67	**8.08**	5.21	10.16	**4.87**
	27.86	56.31	**17.12**	42.73	75.81	**31.17**	49.33	78.28	**37.60**

Table 6: The precision P@*1*, P@*5*, and P@*10* and recall for the models trained with SketchEngine (SE) MEs.

To sum up, the high-quality evaluation dataset for the BLI task should contain words that occur in the MEs used in training and omit OOV words and multi-word expressions. It should take inflected word forms into account or lemmatise the results. Moreover, it should prefer nouns, verbs, adjectives, adverbs, and numerals over pronouns, proper nouns, articles, prepositions, and conjunctions. It should determine the number of the extracted target language words based on the final purpose and the number of senses one headword possesses. Finally, when selecting words and evaluation metrics, we should always consider the language pairs, the end user, and the purpose of the dictionary.

We have provided reproducible criteria applicable for evaluating any model or language pair on the BLI task. These criteria help unify the future evaluation process and make the results comparable and transparent. On top of that, we have made the CMs more approachable for the lexicography field by bringing the lexicography perspective into the evaluation.

Moreover, when observing Table 1, we notice that the results for a close language pair are always better when the unsupervised or identical mode is used. Regardless of the data or MEs utilised during training, the results for the Czech-Slovak language combination are constant and predictable favouring identical or unsupervised mode.

On the other hand, the distant language pairs achieve better results when supervision signals are involved in training. In most cases, the models trained on a distant language pair in a supervised mode surpassed their identical or unsupervised counterparts.

Additionally, the performance of the models trained with SketchEngine MEs exceeded the FastText MEs in many instances. Therefore, high-quality MEs are a key component of the resulting CM.

When looking at the models' recall in Table 1 or Tables 5 and 6, we can conclude that these models cannot be used as a standalone resource in lexicography yet. However, they offer an alternative as supplementary data (e.g., frequently occurring words in the corpus) to parallel-data-based methods for small languages or rare language pairs. Also, they are a good source of lexical-semantically related words in the target language. On top of that, they can be valuable in compiling technical dictionaries, especially when MEs are trained on the domain-specific corpus.

6. Acknowledgements

The research in this paper was supported by the Internal Grant Agency of Masaryk University, project CZ.02.2.69/0.0/0.0/19_073/0016943.

7. References

Artetxe, M., Labaka, G. & Agirre, E. (2016). Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 2289–2294. URL https://aclanthology.org/D16-1250.

Artetxe, M., Labaka, G. & Agirre, E. (2017). Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association* *for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 451–462. URL https://aclanthology.org/P17-1042.

Artetxe, M., Labaka, G. & Agirre, E. (2018a). Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*. pp. 5012–5019. URL https://doi.org/10.1609/aaai.v32i1.11992.

Artetxe, M., Labaka, G. & Agirre, E. (2018b). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 789–798. URL https://aclanthology.org/P18-1073.

Artetxe, M., Labaka, G., Lopez-Gazpio, I. & Agirre, E. (2018c). Uncovering Divergent Linguistic Information in Word Embeddings with Lessons for Intrinsic and Extrinsic Evaluation. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*. Association for Computational Linguistics, pp. 282–291. URL https://aclanthology.org/K18-1028.

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5, pp. 135–146. URL https://aclanthology.org/Q17-1010.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L. & J'egou, H. (2017). Word Translation Without Parallel Data. *ArXiv*, abs/1710.04087. URL https://arxiv.org/abs/1710.04087.

Denisová, M. (2021). Compiling an Estonian-Slovak Dictionary with English as a Binder. In *Proceedings of the eLex 2021 conference*. Lexical Computing CZ, s.r.o., pp. 107–120. URL https://elex.link/elex2021/wp-content/uploads/2021/08/eLex_2021_06_pp107-120.pdf.

Denisová, M. (2022). Parallel, or Comparable? That Is the Question: The Comparison of Parallel and Comparable Data-based Methods for Bilingual Lexicon Induction. In *Proceedings of the Sixteenth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2022*. Tribun EU, pp. 4–13. URL https://nlp.fi.muni.cz/raslan/raslan22.pdf#page=13.

Denisová, M. & Rychlý, P. (2021). When Word Pairs Matter: Analysis of the English-Slovak Evaluation Dataset. In *Recent Advances in Slavonic Natural Language Processing (RASLAN 2021)*. Brno: Tribun EU, pp. 141–149. URL https://nlp.fi.muni.cz/raslan/2021/paper3.pdf.

Glavaš, G., Litschko, R., Ruder, S. & Vulić, I. (2019). How to (Properly) Evaluate Cross-Lingual Word Embeddings: On Strong Baselines, Comparative Analyses, and Some Misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 710–721. URL https://aclanthology.org/P19-1070.

Glavaš, G. & Vulić, I. (2020). Non-Linear Instance-Based Cross-Lingual Mapping for Non-Isomorphic Embedding Spaces. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 7548–7555. URL https://aclanthology.org/2020.acl-main.675.

Herman, O. (2021). Precomputed Word Embeddings for 15+ Languages. *RASLAN 2021 Recent Advances in Slavonic Natural Language Processing*, pp. 41–46. URL https://nlp.fi.muni.cz/raslan/raslan21.pdf#page=49.

Izbicki, M. (2022). Aligning Word Vectors on Low-Resource Languages with Wiktionary. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-*

Resource Languages (LoResMT 2022). Association for Computational Linguistics, pp. 107–117. URL https://aclanthology.org/2022.loresmt-1.14.

Joulin, A., Bojanowski, P., Mikolov, T., Jégou, H. & Grave, E. (2018). Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 2979–2984. URL https://aclanthology.org/D18-1330.

Karan, M., Vulić, I., Korhonen, A. & Glavaš, G. (2020). Classification-Based Self-Learning for Weakly Supervised Bilingual Lexicon Induction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 6915–6922. URL https://aclanthology.org/2020.acl-main.618.

Kementchedjhieva, Y., Hartmann, M. & Søgaard, A. (2019). Lost in Evaluation: Misleading Benchmarks for Bilingual Dictionary Induction. 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)*. Association for Computational Linguistics, pp. 3336–3341. URL https://aclanthology.org/D19-1328.

Kovář, V., Baisa, V. & Jakubíček, M. (2016). Sketch Engine for Bilingual Lexicography. *International Journal of Lexicography*, 29(3), p. 339–352. URL https://doi.org/10.1093/ijl/ecw029.

Lample, G., Denoyer, L. & Ranzato, M. (2017). Unsupervised Machine Translation Using Monolingual Corpora Only. *ArXiv*, abs/1711.00043. URL https://arxiv.org/abs/1711.00043.

Parizi, A.H. & Cook, P. (2021). Evaluating a Joint Training Approach for Learning Cross-lingual Embeddings with Sub-word Information without Parallel Corpora on Lower-resource Languages. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics, pp. 302–307. URL https://aclanthology.org/2021.starsem-1.29.

Ren, S., Liu, S., Zhou, M. & Ma, S. (2020). A Graph-based Coarse-to-fine Method for Unsupervised Bilingual Lexicon Induction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 3476–3485. URL https://aclanthology.org/2020.acl-main.318.

Ruder, S., Vulić, I. & Søgaard, A. (2019). A Survey of Cross-lingual Word Embedding Models. *The Journal of Artificial Intelligence Research*, 65, pp. 569–631. URL https://doi.org/10.1613/jair.1.11640.

Severini, S., Hangya, V., Jalili Sabet, M., Fraser, A. & Schütze, H. (2022). Don't Forget Cheap Training Signals Before Building Unsupervised Bilingual Word Embeddings. In *Proceedings of the BUCC Workshop within LREC 2022*. European Language Resources Association, pp. 15–22. URL https://aclanthology.org/2022.bucc-1.3.

Søgaard, A., Ruder, S. & Vulić, I. (2018). On the Limitations of Unsupervised Bilingual Dictionary Induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 778–788. URL https://aclanthology.org/P18-1072.

Tian, Z., Li, C., Ren, S., Zuo, Z., Wen, Z., Hu, X., Han, X., Huang, H., Deng, D., Zhang, Q. & Xie, X. (2022). RAPO: An Adaptive Ranking Paradigm for Bilingual Lexicon Induction. *ArXiv*, abs/2210.09926. URL https://arxiv.org/abs/2210.09926.

Vulić, I., Glavaš, G., Reichart, R. & Korhonen, A. (2019). Do We Really Need Fully Unsupervised Cross-Lingual Embeddings? 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)*. Association for Computational Linguistics, pp. 4407–4418. URL https://aclanthology.org/D19-1449.

Vulić, I. & Korhonen, A. (2016). On the Role of Seed Lexicons in Learning Bilingual Word Embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 247–257. URL https://aclanthology.org/P16-1024.

Vulić, I., Korhonen, A. & Glavaš, G. (2020). Improving Bilingual Lexicon Induction with Unsupervised Post-Processing of Monolingual Word Vector Spaces. In *Proceedings of the 5th Workshop on Representation Learning for NLP*. Association for Computational Linguistics, pp. 45–54. URL https://aclanthology.org/2020.repl4nlp-1.7.

Woller, L., Hangya, V. & Fraser, A. (2021). Do not neglect related languages: The case of low-resource Occitan cross-lingual word embeddings. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*. Association for Computational Linguistics, pp. 41–50. URL https://aclanthology.org/2021.mrl-1.4.