Computational Linguistics in the Netherlands Journal 12 (2022) 125-140 Submitted 12/2022; Published 12/2022

# Improving Domain-specific Cross-lingual Embeddings with Automatically Generated Bilingual Dictionaries

**Pranaydeep Singh**[1] pranaydeep.singh@UGent.be
**Ayla Rigouts Terryn**[2] Ayla.RigoutsTerryn@KULeuven.be
**Els Lefever**[1] Els.Lefever@UGent.be

*LT*$^3$ *Language and Translation Technology Team, Department of Translation, Interpreting and Multilingual Communication, Faculty of Arts and Philosophy, Ghent University, Belgium*

*Centre for Computational Linguistics, KU Leuven (KULAK), Belgium*

## Abstract

This paper reports on a set of proof-of-concept experiments performed to evaluate and improve the alignment of monolingual embeddings for a specialised domain, viz. the medical use case of heart failure. The presented approach, which creates domain-specific dictionaries on-the-fly from cross-lingual Wikipedia links, achieves good results for cross-lingual alignment of this specialised vocabulary in three language pairs: English-Dutch, English-French, and Dutch-French. The experimental results show that the setup incorporating a smaller but dedicated domain-specific dictionary outperforms the alignment incorporating a larger but general-domain seed dictionary. A detailed error analysis reveals that many potentially useful (near-)equivalents are found beyond those present in the gold standard, and it inspires strategies for further improvements, such as lemmatisation and improved tokenisation.

## 1. Introduction

The recent introduction of large pre-trained language models has caused a considerable performance improvement for many natural language processing tasks. This is, even more, the case for the transformer architectures, such as BERT (Devlin et al. 2019) and RoBERTa (Liu et al. 2019). However, these language models suffer from the important disadvantage of requiring large training data collections and computation power. As a result, they have mainly been trained and evaluated on resource-rich languages and general domain data.

To overcome this lack of resources, researchers have started to investigate the use of cross-lingual information, where knowledge or data from a well-resourced language, like English, is used to improve the modelling for a specific NLP-task in a low(er)-resourced target language. The idea of cross-lingual embeddings has been introduced by Mikolov et al. (2013), who hypothesised that vector spaces in different languages share a certain similarity and that a projection can be learned from one language to another. Since this seminal work, a lot of research has been proposed to perform cross-lingual alignment (see Section 2). Although these cross-lingual alignments have been shown to work well for a wide range of NLP tasks and languages, research evaluating these alignments for domain-specific terms is scarce. Since most of the applications for cross-lingual alignments lie in specialised domains like medicine, finance, etc, it is important to evaluate the quality of the alignments in a domain-specific setting rather than on general domain data. One of the few studies on this topic is presented by Shakurova et al. (2019), who investigate best practices for constructing the seed dictionaries used for cross-lingual alignment for a specific domain, being the sequence labelling task of Curriculum Vitae parsing.

The aim of the proposed research was twofold. First, we wanted to evaluate the performance of existing alignment methods, and Vecmap (Artetxe et al. 2018a) more in particular, on specialised vocabulary for a given domain, namely the medical use case of heart failure for this research. Sec-

ond, we wanted to research a methodology to improve the performance for domain-specific cross-lingual alignment. To this end, we have investigated an automatic approach to construct smaller domain-specific seed dictionaries from Wikipedia pages and evaluated it on three language pairs, two involving English as the source language (English-Dutch and English-French), and one having Dutch as the source language (Dutch-French).

The remainder of this paper is organised as follows. Section 2 gives a brief overview of relevant related research, whereas Section 3 elaborates on the proposed methodology to use Wikipedia cross-lingual links to create domain-specific dictionaries. Section 4 provides an overview and analysis of the experimental results, and Section 5 ends this paper with concluding remarks and indications for future research.

## 2. Related research

Various methodologies have been proposed to align monolingual word embeddings into a common space, assuming that a perfect mapping can be learned by traversing between vector spaces in different languages. Mikolov et al. (2013) learned a linear mapping from one space to another and optimised the performance by means of a bilingual lexicon. Other approaches also rely on the assumption of a similarly structured embedding space to project monolingual spaces into a shared space, either based on a seed translation dictionary (Faruqui and Dyer 2014), or some other form of cross-lingual supervision based on parallel corpora (Guo et al. 2015, Søgaard et al. 2015, Vulić and Moens 2016).

As large bilingual lexicons are often lacking for low-resourced languages or specific domains, approaches have been proposed that either completely eliminate or drastically reduce the size of the bilingual lexicon. Artetxe et al. (2017) further explored these ideas by using a combination of back translation and denoising. This approach was, however, severely lacking in terms of performance as compared to a method with cross-lingual signals. The advent of adversarial networks brought on some unique ideas which opened up a lot of new research directions: a discriminator is trained to identify whether an embedding originates from a source language or a target language and a mapping is trained to fool the discriminator. The underlying principle is that there is an orthogonal matrix $W$, which can transform embeddings in one language to embeddings in another language. VecMap (Artetxe et al. 2018a) uses this base adversarial learning approach but with a lot of clever additions and tweaks. Unsupervised initialisation of the transformation is done using the gram matrices of the individual languages and Singular Value Decomposition (SVD). Then iterative training is performed to obtain an increased similarity for the pre-initialised dictionary.

The more recent contextual embeddings significantly enhanced word and sentence representations and improved upon previous methods of cross-lingual alignment like MUSE (Lample and Conneau 2019) and VecMap (Artetxe et al. 2018a) due to their dynamic nature. Multilingual BERT (mBERT, Devlin et al. (2019)) and XLM (Conneau and Lample 2019) were jointly trained for Masked Language Modelling on 104 languages and significantly outperformed previous approaches for a variety of zero-shot cross-lingual tasks. However, a number of recent works have also discussed the various limitations of large multilingual models, referred to as the curse of multilingualism. Wu and Dredze (2020) talk about poorer representation for lower-resourced languages, due to various issues like smaller vocabulary shares, poor tokenisation strategies for morphologically rich languages, etc. Similar problems are reported by Gerz et al. (2018), who also discuss the fact that fine-grained morphological features are modelled incorrectly due to lower frequencies in the training corpora and that performance in multilingual LMs can vary largely based on the typological features of the languages. Therefore, while these models present an excellent opportunity to unify monolingual learning in a multilingual setting, they aren't as reliable for cross-lingual learning as methods with some supervision of cross-lingual signals like seed dictionaries.

All these methodologies focus on the alignment and evaluation of general language vocabulary, relying on generic seed dictionaries and pretrained cross-lingual embeddings. The evaluation of the

cross-lingual alignment is often performed through the task of Bilingual Lexicon Induction (BLI). Researchers like Laville et al. (2022), however, discuss some issues related to the gold standard data sets (e.g., MUSE lexicon) typically used for this evaluation, such as a limited representation of various Part-of-Speech categories (the evaluation set containing a lot of proper nouns), a high number of word pairs showing a large graphical similarity (many identical word pairs and graphically close ones), and overrepresentation of high-frequency words. As the MUSE lexicon is not only used for evaluation, but also as a seed dictionary for aligning cross-lingual embeddings, similar issues might hamper the training process of the cross-lingual alignments.

Little research has been performed on specialised vocabulary so far. Shakurova et al. (2019) have investigated some best practices to construct seed dictionaries for specific domains. The obtained embeddings are evaluated for Curriculum Vitae parsing, and the experimental results show that the size of the dictionary, the frequency of the dictionary terms in the domain-specific corpora and the source of the data (task-specific or generic domain) do have an impact on the parsing performance. In addition, they show that the bilingual dictionary gets more important in proportion to the smaller size of the training data in the low-resourced language. Recent work has demonstrated good results when combining large pretrained Transformer-based models such as BERT (Devlin et al. 2019) with external linguistic knowledge for the biomedical domain. Such large external resources are, however, not always available for low-resourced languages, but Liu et al. (2021) show that general domain bitext helps to transfer specialised knowledge to languages with little to no in-domain data for the task of biomedical entity linking.

In this research, we want to take one step back and evaluate and improve the performance of cross-lingual alignment, performed with VecMap, on domain-specific words. To this end, we propose a straightforward method using domain-specific dictionaries generated on-the-fly from Wikipedia to improve the alignment performance on downstream tasks for the domain in question.

Wikipedia has proven to be a useful resource for automatic data creation for various NLP tasks, going from using Wikipedia titles in the framework of Neural Machine Translation (Karakanta et al. 2018) or more specialised terminology translation (Molchanov et al. 2021), to using Wikipedia as a multilingual knowledge resource for cross-lingual information retrieval (Nguyen et al. 2009, Sorg and Cimiano 2012) or for mining biomedical synonyms (Jagannatha et al. 2015). More related to the research we propose here is the approach taken by Sharoff (Sharoff 2018, Sharoff 2020), who extracts seed dictionaries for cross-lingual word alignment from the titles of interlinked Wikipedia articles in two languages ("iWiki links"). The titles are word-aligned and the resulting word-level dictionaries are filtered against the respective frequency lists. Jiang et al. (Jiang et al. 2020), however, show that using Wikidata and full Wikipedia pages in different languages appears to be more reliable than using page titles or cross-lingual Wikipedia links, as titles can be ambiguous and cross-lingual links may direct to disambiguation or mismatched pages.

In the next section, we explain how we create domain-specific dictionaries to be incorporated as seed dictionaries for cross-lingual alignment of embeddings. We do not only use the titles, but start from a domain-specific Wikipedia article, and collect all terms in that article for which a separate Wikipedia article exists. Our results (Section 4) show that the alignment quality improves when using this smaller and more focused dictionary instead of the more commonly used seed dictionaries, such as MUSE.

## 3. Alignment of Domain-specific Terms

This section describes the methodology that was applied to align monolingual embeddings for three language pairs, viz. English-Dutch, English-French, and Dutch-French for a use case from the medical domain of heart failure. The approach we present was inspired by two hypotheses: (1) using related words in the seed dictionary for alignment improves alignment for domain-specific words, and (2) re-training embeddings with domain-specific text improves representations and alignments for the respective domain.

### 3.1 Cross-lingual Alignment

As discussed already, the alignment of monolingual embeddings has been studied for a while now, starting with Mikolov et al. (2013) exploiting similar geometries of the embedding spaces to learn a linear mapping between embeddings. The older methodologies, however, required large parallel dictionaries to guide the alignment, which are often not available for domain-specific text. For this research, we opted to use VecMap (Artetxe et al. 2018b), because it has been frequently shown to perform better in cases where the size of the bilingual dictionary is strictly limited. We align monolingual embeddings trained using FastText (Bojanowski et al. 2017) on the Common Crawl corpus and Wikipedia. These models were trained using the Continuous Bag of Words (CBOW) model with position weights, a dimensionality of 300, character n-grams of length 5, a window of size 5, and 10 negative samples.

VecMap focuses on learning the orthogonal matrix using iterative self-learning. The algorithm exploits similar distributions of nearest neighbours for words that have similar meanings, in order to improve a small bilingual dictionary by adding new pairs to it after each iteration. This ends up making the alignments effective with a significantly smaller dictionary.

Given a source language $s$ and a target language $t$, the objective of the classical alignment methods is to learn a transformation

$$E_{s,t} \approx W^{s \to t} E_{s,s} \tag{1}$$

where $E_{s,s}$ represents the embeddings of the source language in their original space, while $E_{s,t}$ represents the embeddings of the source language, in the target language's multi-dimensional space. Inversely,

$$E_{t,s} \approx W^{t \to s} E_{t,t} \tag{2}$$

should also be a possibility. This can now be formulated as an optimisation problem for orthogonal matrix $W$. VecMap achieves this by maximising for similarity over a sparse seed dictionary, which can be initialised with zero supervision or using identical words if a seed dictionary is not available, and iteratively improving the dictionary and re-learning the alignment after each optimisation step. Other approaches, such as MUSE (Lample and Conneau 2019), achieve the same objective by initialising $W$ using an adversarial objective, where $W$ is optimised such that a discriminator model is unable to differentiate between the embeddings originating from $E_{t,t}$ and $WE_{s,s}$. To find the optimal matrix $W$, we use the supervision dictionary $D$ where $D$ is a matrix such that $D_{ij} = 1$ if the $i_{th}$ word in the source language corresponds to the $j_{th}$ word in the target language. The optimisation is therefore formulated by Vecmap as

$$W^* = \underset{W}{argmin} \sum_i \sum_j D_{ij} \left\| E_{i \in s,s} W - E_{j \in t,t} \right\| \tag{3}$$

The baseline analysis and comparisons are done using the monolingual FastText embeddings aligned with VecMap using a generic seed dictionary made available by MUSE (Conneau et al. 2017). These dictionaries were created in an unsupervised manner, exploiting the similarity in the shape of monolingual embedding spaces, and have not been manually verified. They contain around 5000 word pairs for every language pair used. We performed a first analysis of the cross-lingual alignment of domain-specific terms by manually inspecting the nearest neighbours for a sample of terms in the aligned embedding spaces. Table 1 shows the five nearest neighbours for the term *atherosclerosis* both in the English and Dutch embedding space.

Figure 1 shows a visualisation of the multilingual embedding space for a couple of domain-specific terms for the use case of heart failure. From this small sample, it is already clear that the alignments for specialised words are rather poor when using the default VecMap embeddings (and

| nearest neighbours in English space | | nearest neighbours in Dutch space | |
|---|---|---|---|
| **neighbour** | **score** | **neighbour** | **score** |
| atherosclerosis | 1,0000 | trombocytenopenie | 0,7011 |
| atherosclerotic | 0,9137 | vaatziekten | 0,6883 |
| arteriosclerosis | 0,8565 | trombose | 0,6831 |
| hypertension | 0,8292 | longaandoening | 0,6814 |
| hypercholesterolemia | 0,7882 | stofwisselingsziekten | 0,6811 |

Table 1: Five closest nearest neighbours of the English word *atherosclerosis* in the English and Dutch embedding spaces.

seed dictionary). To improve on the alignment for specialised terms, we decided to construct a domain-specific dictionary and to retrain the FastText embeddings using domain-specific data.

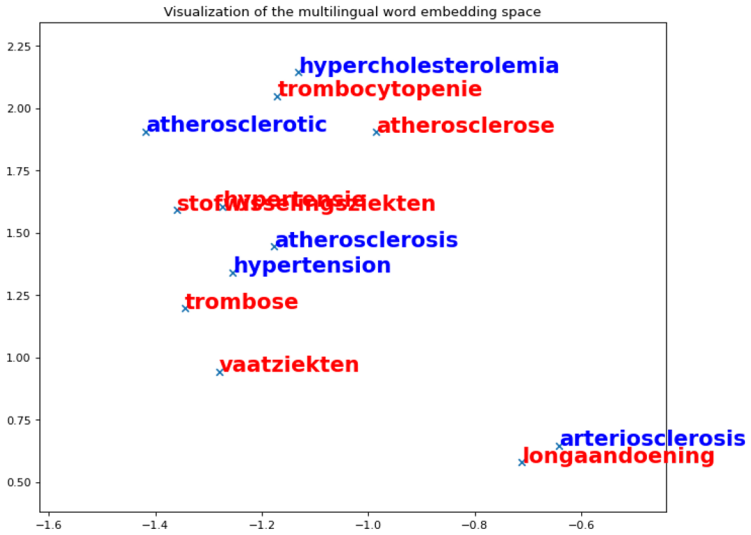



Figure 1: Multilingual word embedding space for a sample of English-Dutch aligned terms for the use case of heart failure.

### 3.2 Creation of a Domain-specific Dictionary

The bilingual seed dictionary is crucial to the process of cross-lingual alignment of the embeddings. As we hypothesise that using a domain-specific dictionary will improve the alignment for specialised vocabulary, we constructed a specialised dictionary (Silvanus) from Wikipedia data for the concerned domain. This was inspired by previous work using the World Wide Web as a free source to create large linguistic resources, as was done for instance by de Groc (2011) to compile large thematic corpora used to automatically generate bilingual terminologies.

The construction of this domain-specific bilingual dictionary is explained in the form of pseudo-code below. In our case, we start from the single Wikipedia entry on *heart failure*, though it is possible to start from multiple articles as well. Each term in that article for which a separate Wikipedia article exists, is collected, e.g., *cardiac arrest* and *beta blockers*. This is considered *depth 0*. Note that not all terms with a hyperlink in a Wikipedia article refer to a separate article. Sometimes they refer to a section within the same article (e.g., *right-sided heart failure*), or to an article with a different title (e.g., the link for *leg swelling* refers to the page on *peripheral edema*). For each term in the collection of terms for which a valid page does exist, this process is repeated, i.e., all terms with

links to other Wikipedia pages are collected. For instance, the term *anemia* is collected from the main page on *heart failure*, and terms like *blood* and *hemoglobin* are in turn collected from the page on *anemia*. This would be depth 1. The process can be repeated, to add terms at depth 2 or beyond. For each instance in this collection of terms for which Wikipedia articles exist, the algorithm will check whether a cross-lingual link exists to the equivalent page in the target language. For example, the English *anemia* page has links to both the Dutch (*bloedarmoede*) and French (*anémie*) pages on the same subject.

**Algorithm 1** Constructing a Domain-specific Bilingual Dictionary (Silvanus)

```
Given a list of Domain Words D
for d in D do
    if d has a valid Wikipedia entry then
        Identify ∀i ∈ I : I is the set of all Wikipedia links on page d
        for i in I do
            Find j where j is the cross-lingual link in the target language L_t
            Add i and j to the domain-specific bilingual dictionary
```

It is not uncommon for Wikipedia articles to start by providing synonyms and abbreviations. For instance, the English page on *heart failure* starts by mentioning a synonym and abbreviations: "Heart failure (HF), also known as congestive heart failure (CHF)". While these are currently not considered, it could be an interesting avenue for future research to take advantage of the additional information to create a more elaborate dataset.

As a proof of concept, we only use a single seed word, i.e., the name of the domain (*Heart Failure*) as a search term. The seed dictionary can obviously be enhanced to make a more comprehensive search, but we explore here the possibility of using a single term as a starting point. Using just a single seed word, we are able to obtain more than 5000 word pairs at depth 1. After filtering multi-word terms (simply removing terms that contain a whitespace character) and terms with encoding issues, we are left with a specialised bilingual dictionary of 2702 word pairs for English-French and 2819 word pairs for English-Dutch for this specific domain. Out of 2702 word pairs for English-French, only 1727 were present in the FastText pre-trained embeddings, whereas for Dutch, out of 2819 word pairs, only 1895 were eventually used for the alignments. Even though we performed our alignment experiments with the heart failure domain, we demonstrate in Table 2 that the bilingual dictionary construction methodology is also viable for other domains, by constructing dictionaries for two additional domains, i.e., Dressage and Wind Power.

| **Domain** | **English-Dutch pairs scraped** | | **English-French pairs scraped** | |
|---|---|---|---|---|
| | **Pre-filtering** | **Post-filtering** | **Pre-filtering** | **Post-filtering** |
| Heart Failure | 6434 | 2819 | 6477 | 2702 |
| Wind Power | 11835 | 3830 | 9150 | 2192 |
| Dressage | 7342 | 2307 | 5745 | 1331 |

Table 2: Bilingual dictionary sizes for English-Dutch and English-French for the domains of Heart Failure, Wind Power, and Dressage. Pre-filtering represents the initial size of the dictionary scraped, while post-filtering refers to the dictionary size after removing multi-word terms and encoding issues.

Using the obtained domain-specific dictionary $D$ and the iterative self-learning of VecMap, we can then construct word alignments for a specific domain that consistently outperform word alignments constructed using generic dictionaries or unsupervised methods, as is shown in Section 4. This confirms our first hypothesis, which states that using related words in the seed dictionary improves the alignment for domain-specific words.

### 3.3 Improving Monolingual Embeddings with Domain-Specific Fine-tuning

Our second hypothesis was that the monolingual embeddings themselves could be improved in their representation of domain-specific terminology by further training them on unlabelled data from the domain. To investigate this, we trained a set of FastText embeddings in English, initialised using the same Common Crawl and Wikipedia embeddings we have used in the rest of the paper. The embeddings were fine-tuned on the PubMed[1] English corpus to enhance the representation of words in the heart failure domain. We continued the training process with the same parameters, i.e., CBOW with position weights, n-grams of length 5, a window of 5, and 10 negative samples, and used these new English embeddings as a replacement for the baseline English embeddings in our alignments. These new embeddings, however, resulted in extremely poor alignments for English-Dutch, obtaining a Mean Reciprocal Rank (MRR) of 0.309, compared to an MRR of 0.567 in our regular setting. We hypothesise that this occurs due to the isomorphism assumption for alignment, which states that two monolingual embedding spaces need to be isomorphic to have an acceptable alignment. Isomorphism is influenced by the training setup, like the data, training time, and parameters. Since the PubMed data would be vastly different from the Common Crawl and Wikipedia corpora, we can assume this makes the spaces non-isomorphic. Because of the bad alignment results, we did not further explore this approach of domain-specific fine-tuning.

## 4. Experimental Results and Analysis

This section introduces the experimental setup and provides a quantitative and qualitative evaluation of the alignment approach proposed for domain-specific terms.

### 4.1 Experimental Setup

We evaluate the proposed methodology using samples from the ACTER dataset (Rigouts Terryn et al. 2020) as a gold standard. ACTER is a manually annotated dataset for term extraction, covering three languages (English, French, and Dutch), and four domains (corruption, dressage, heart failure, and wind energy). For each of the four domains, a comparable corpus (or a parallel corpus in the case of corruption) exists of more or less equal size in all languages ($\pm$50k tokens). Terms (both specialised and more common) and Named Entities were manually annotated in each corpus based on publicly available annotation guidelines[2]. For the comparable corpus on heart failure, the monolingual annotations were supplemented with cross-lingual annotations (Rigouts Terryn et al. 2018). For each annotated term, equivalents were sought among the annotations of the other languages (meaning that only terms and equivalents were added that occur in the specific corpus used for this dataset). Therefore, this dataset does not pretend to offer an exhaustive overview of all relevant terms and their equivalents for the domain, but only attempts to cover those in the specific corpus.

Terms can have multiple alignments due to, among others, synonyms (e.g., the Dutch terms *benauwdheid*, *kortademigheid* and *dyspnoe* are all synonyms and equivalents to the English *breathlessness* and *dyspnea*), abbreviations (e.g., the Dutch *hartfalen* can be aligned to the English *heart failure* or *HF*), or alternative spellings (e.g., the Dutch *bètablokker* can be linked to the English *beta-blocker* or *β-blocker*). Since each term may have multiple correct alignments, we consider only the option that is aligned closest to the source word for the evaluation. We calculate four different scores: P@1 indicating precision for top-1 nearest neighbours (KNNs), P@5 for the top-5 KNNs, P@10 for top-10 KNNs and the Mean Reciprocal Rank formulated as:

1. https://pubmed.ncbi.nlm.nih.gov/
2. https://lt3.ugent.be/publications/acter-terminology-annotation-guidelines/

$$MRR = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{rank_i} \tag{4}$$

We perform the evaluation for three language pairs, being English-Dutch, English-French and Dutch-French. For each language pair we perform the domain-specific alignment using VecMap for five iterations in two setups: (1) using the constructed domain-specific dictionaries and (2) using the large generic dictionary for the respective language pair from the freely available MUSE-generated dictionary sets[3]. We provide the MUSE-dictionary alignments as baseline scores except for the case of Dutch-French, where a MUSE dictionary was not available. We still wanted to present the obtained alignment scores for Dutch-French, however, as we also wanted to explore the possibility of aligning two non-standard monolingual spaces using domain-specific dictionaries, since English is almost always used as a source language in related research.

### 4.2 Experimental results

As is illustrated by Tables 3 and 4, the domain-specific Silvanus dictionaries result in better alignments for both language pairs in question. Even though the dictionaries are almost one-third in size compared to the generic MUSE dictionaries, due to the domain-specific terminology that is predominant in the word pairs, the resulting alignments appear to be better when incorporating the Silvanus dictionaries. This is an important finding, as it indicates that it may be better to have a smaller dictionary with more relevant pairs, than a larger, noisier dictionary for the alignment of the embedding spaces.

| **EN-NL** | **Dictionary Size** | **P1** | **P5** | **P10** | **MRR** |
|---|---|---|---|---|---|
| Baseline | 4959 | 0.447 | 0.629 | 0.688 | 0.534 |
| Silvanus | 1895 | 0.486 | 0.661 | 0.709 | 0.567 |

Table 3: Comparisons between the alignments using the domain-specific Silvanus dictionary versus the generic MUSE dictionaries for the **English-Dutch** language pair.

| **EN-FR** | **Dictionary Size** | **P1** | **P5** | **P10** | **MRR** |
|---|---|---|---|---|---|
| Baseline | 4986 | 0.640 | 0.797 | 0.812 | 0.709 |
| Silvanus | 1727 | 0.668 | 0.797 | 0.814 | 0.726 |

Table 4: Comparisons between the alignments using the domain-specific Silvanus dictionary versus the generic MUSE dictionaries for the **English-French** language pair.

Table 5 shows the results for the Dutch-French cross-lingual alignments. While the results for Dutch-French aren't that promising when incorporating a domain-specific dictionary generated at depth 2, this might be due to the limited size of the Dutch-French Silvanus dictionary (927 term pairs), which is almost half in size compared to English-Dutch and English-French. The size of the dictionary can be expanded by increasing the depth of the search, but a higher depth makes the obtained word pairs less likely to be relevant to the domain in question. Since the size of the initial domain-specific dictionary appeared to be too small for Dutch-French, we experimented with a dictionary collected at depth 3, and the results listed in Table 5 show that this indeed improves the performance considerably. Therefore, even if the words might be slightly less relevant, an increase in the dictionary size still boosts performance.

In the next section, we dive further into the types of mistakes and the potential causes underlying them when using the Silvanus dictionaries for cross-lingual alignments.

3. https://github.com/facebookresearch/MUSE/blob/main/README.md

| NL-FR | Dictionary Size | P1 | P5 | P10 | MRR |
|---|---|---|---|---|---|
| Baseline | - | - | - | - | - |
| Silvanus - Depth 2 | 927 | 0.144 | 0.223 | 0.255 | 0.18 |
| Silvanus - Depth 3 | 2423 | 0.197 | 0.284 | 0.336 | 0.245 |

Table 5: Alignment scores for **Dutch-French** when using the domain-specific Silvanus Dictionary.

### 4.3 Error Analysis

#### 4.3.1 Error Annotation

To gain a better understanding of the results and the types of errors, a detailed error analysis was performed. This more nuanced evaluation gave us a better idea of the usability of the methodology, especially since the gold standard is unavoidably imperfect. Identifying terms and equivalents is, to a certain degree, a subjective task. Moreover, the gold standard is based on cross-lingual annotations of only those terms that occur in the comparable corpus used to create the dataset. As explained in Section 4.1, this corpus consists of original texts (unaligned, not translations) with around 50k tokens per language. Hence, some concepts are only mentioned in one or two of the languages of the corpus, and not all terms for each concept are necessarily used. Consequently, the system may find valid equivalents that are automatically evaluated as incorrect, because they are not part of the gold standard. For the error analysis, a random selection was made of 200 cross-lingual term pairs (100 French-English, 100 Dutch-English) and the ten most highly ranked equivalents per term pair by the system. So, for each term in the source language, we had the gold standard term in the target language and ten ranked, automatically generated potential equivalents. These were manually analysed by a linguist-terminologist, who identified additional equivalents and near-equivalents. Five labels were used:

1. **Additional equivalent** (not in Gold Standard), regardless of number

   e.g., the English equivalent for the Dutch *anticoagulantia* in the Gold Standard is *anticoagulants*. Additional equivalents found were *anticoagulant*, *anti-coagulant*, and *anti-coagulants*.

2. **Near-equivalent with different part-of-speech** (POS)

   e.g., the English equivalent for the Dutch *stabiele* in the Gold Standard is *stable*. Near-equivalents with different POS were *stability*, *stableness*, *stabilizing*, and *stablest*.

3. **Additional equivalent with wrong spelling**

   e.g., the English equivalent for the French *santé* in the Gold Standard is *health*. Wrong spellings that were also found are *helath*, *heatlh*, and *healt*. Note that, in cases with different but very commonly used spellings (e.g., *anticoagulant* and *anti-coagulant*, *haemorrhage* and *hemorrhage*), equivalents would be classified as normal additional equivalents, not as wrong spellings.

4. **Additional equivalent with tokenisation issue**

   e.g., besides the correct equivalent *metabolism*, the system also found *metabolism-*, *metabolism.*, and *metabolism.the*.

5. **Antonyms**

Wrong spellings of terms and tokenisation issues were only annotated for candidates that would otherwise be equivalent. So, if, for instance, a term was misspelled but not in any way equivalent to the source term, the misspelling was not annotated, and not counted in the analysis. In case

multiple labels were applicable, the first one in the list was assigned, e.g., if a near-equivalent with a different POS was also misspelled, it would still be annotated as the former.

Based on this analysis, strict and lenient versions of precision@rank were calculated and compared. For *strict precision*, only the Gold Standard and additional equivalents were considered correct. For *lenient precision*, near-equivalents with a different POS and additional equivalents with spelling mistakes or tokenisation issues were also considered correct. The motivation was that any human using these results would, of course, prefer to get strict equivalents, but would probably also be able to derive a strict equivalent from the other categories relatively easily. For instance, if the noun *ischemia* is suggested instead of the adjective *ischemic*, human users would not have much trouble finding *ischemic* based on the noun. The results of the analysis, with numbers per category and per language pair, can be found in Table 6.

| **Number of instances found per category among top 10 generated suggestions** | **100 NL-EN** | **100 FR-EN** | **Total** |
|---|---|---|---|
| **Additional equivalents (strict)** | 53 | 64 | 117 |
| **Near-equivalents with different POS** | 73 | 86 | 159 |
| **Equivalents with wrong spelling** | 21 | 38 | 59 |
| **Equivalents with tokenisation issues** | 26 | 30 | 56 |
| **Antonyms** | 18 | 16 | 34 |
| **p@1 strict** | .72 | .63 | .68 |
| **p@1 lenient** | .76 | .74 | .75 |
| **p@10 strict** | .87 | .90 | .89 |
| **p@10 lenient** | .87 | .91 | .89 |

Table 6: Summary of the error analysis for a random selection of 100 term pairs in Dutch-English and French-English.

#### 4.3.2 Discussion of Error Analysis

**Additional equivalents and impact on scores:** Several interesting conclusions can be drawn from these numbers. The gold standard only considered those terms and equivalents that were present in the corpus based on which it was constructed, so it is entirely possible that the system found additional equivalents that were simply not present in the gold standard corpus. Moreover, while the gold standard contains multiple potential equivalents for some terms, we only consider the most closely ranked one in our evaluation. Therefore, it was to be expected that some additional equivalents would be found in the error analysis. Nevertheless, the number of additional equivalents, as shown in Table 6, was remarkable, especially since the annotation was relatively strict and did not take into account many near-equivalents. For instance, the Dutch term *terminale* has the Gold Standard English equivalent of *terminal*, but four additional terms were detected that can be considered equivalent in some contexts: *stage-four*, *stage-4*, *dying*, and *incurable*. Since these are only equivalent in some contexts, they were not annotated. Similarly, compounds that contain the correct equivalent, but are not themselves equivalent are regularly found, but not included in the error analysis, e.g., *potassium-rich*, and *potassium-containing* for *potassium*, and *re-diagnosed*, *rediagnosed*, and *misdiagnosed* for *diagnosed*. The presence of multiple potential equivalents, some of which are different from the Gold Standard option, has a considerable impact on the scores. Precision@1 based on just the Gold Standard data for the samples of 100 term pairs NL-EN and FR-EN is only 0.61 and 0.54 respectively, and precision@10 is 0.80 and 0.89 respectively. Compared to the strict precision scores in Table 6, which include additionally annotated equivalents, this is a considerable difference of 1 up to 11 percentage points. The lenient calculation of precision, which includes near-equivalents and equivalents with different POS or tokenisation issues, shows an even

more optimistic view of the results, especially for precision@1, which is 4 (NL-EN) and even 11 (FR-EN) percentage points higher than strict precision.

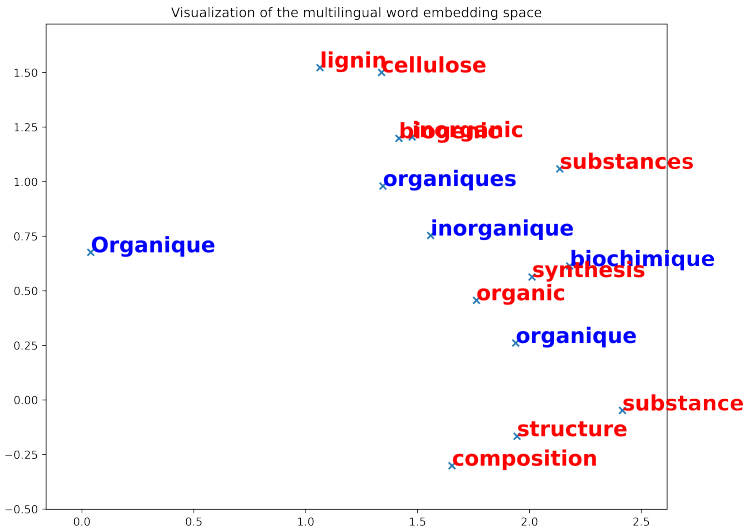
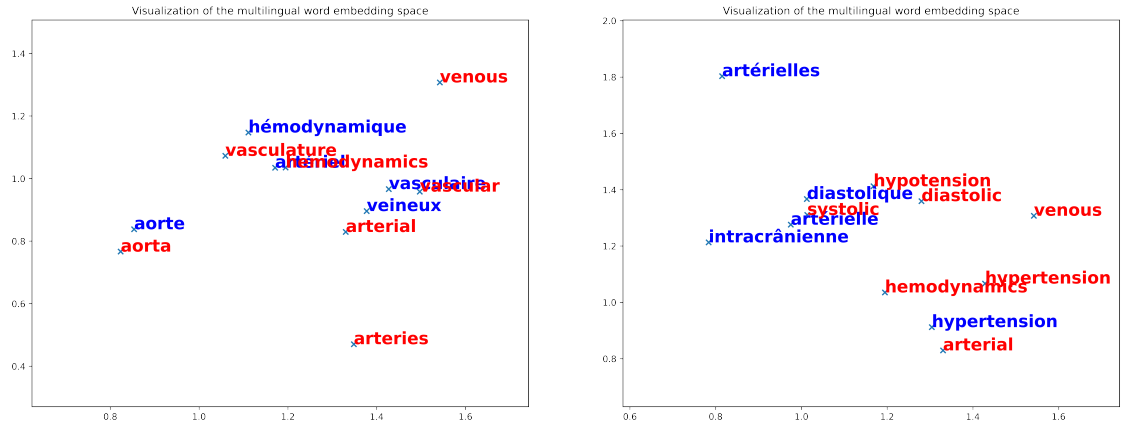



Figure 2: Multilingual word embedding space and nearest neighbours in English and Dutch for the word 'inorganique'.

**Antonyms:** The presence of near-equivalents and compound terms that contain the correct equivalent is expected, and an indication of relevant semantic spaces. However, we also found many antonyms, which is logical (as they will co-occur with the right term a lot), but not intended, since they have the opposite meaning. These antonyms are often indicated by an additional prefix (e.g., *diagnosed* versus *undiagnosed*, *neurological* and *non-neurological*, *organic* and *inorganic*) (see also Figure 2), or a different prefix (e.g., *hypertrophy* and *hypotrophy*). Antonyms are sometimes among highly ranked results, but out of the 34 antonyms found, only 3 occur before the most highly ranked valid equivalent: *non-terminal* comes before *terminal*, *diastolic* precedes *systolic*, and *inorganic* is ranked before *organic* (Figure 2). So, while having antonyms among the results is not ideal, they are not a big issue for the results, and human users would likely be able to recognise many of them relatively easily.

**Impact of language pair:** There are small differences between the language pairs, but given the small sample size, it is difficult to generalise. One of the clearest differences is that strict p@1 is lower for FR-EN than for NL-EN (-0.09), whereas strict p@10 is more similar (-0.04). This difference is much smaller for strict precision scores (-0.02 and -0.03). This seems mostly due to a combination of two factors: there are more equivalents with a different POS or a wrong spelling that are ranked in first position in FR-EN (11) than NL-EN (4); and the valid Gold Standard equivalent has a lower average rank in FR-EN (2.3) than in NL-EN (1.6). Further research is required to test whether this pattern is consistent and how it can be explained and improved.

**Misspellings and tokenisation issues:** The number of terms that were equivalent except for a wrong spelling or tokenisation issue was unexpected. The most remarkable example of wrong spellings was for the English equivalents of the French *physiques*. The first option was the correct version (*physical*), and then there were eight different wrong spellings: *phyiscal*, *phsyical*, *physcial*, *phyisical*, *phsical*, *phisical*, *phyical*, and *physicial* (the tenth prediction was *physiological*). While it is not surprising that wrong spellings can be found in the Wikipedia corpus, it is surprising how many of them occur often enough to be included in the relevant embedding space. These wrong spellings are mostly found with terms that are either relatively common, e.g., *health*, *healthy*, and *diagnosed*, or




(a) (b)

Figure 3: Contrasting embedding spaces showing completely different neighbours of two different forms of the same word, being artériel (a) and artérielle (b).

understandably hard to spell, e.g., *haemorrhage* and *pharmaceutical*. Despite the prevalence of the issue, misspellings were never ranked higher than correct equivalents. These findings do support the idea of using word embeddings to detect spelling variation and common spelling mistakes (Nguyen and Grieve 2020). Tokenisation issues mainly consisted of instances with leftover punctuation, e.g., *brain.*, and *lung-*, or appended words (often with punctuation), e.g., *brain.the*, and *europeand*. Interestingly, these issues were often concentrated around the same terms. Out of twenty-eight terms for which equivalents were found with tokenisation issues, fifteen had at least two different tokenisation issues. For instance, *brain.the*, *brain.this*, *brain.but*, *brain.so*, *brain.and*, *brain.*, and *brain.it*. This issue only occurs for relatively common terms, and a misspelling was never ranked higher than the correct equivalent.

**Missing equivalents:** A few things stand out when looking at those terms for which no good equivalent was found. A first observation is that this happens mostly to specialised terms like *natriuretic*, *trastuzumab*, and *hypertrophic*. There are exceptions, however, like *arterial*, which is not very specialised, yet is not found for either Dutch equivalent (*arterieel* and *arteriële*). It is not found for the French equivalent *artériels* either, though it was found (at ranks 2, 3, and 9 respectively), for the French equivalents *artériel*, *artérielles*, and *artérielle*. This is a phenomenon that was found for other terms that occur in multiple forms as well. For instance, the correct English equivalent *mitral* was only found for one of the three present forms in French (*mitral*, *mitrales*, and *mitrale*). In such cases, the embedding space for the multiple full forms of the same lemma is often quite different, as illustrated by Figure 3 for *artériel* and *artérielle*. This is a strong argument for the use of lemmatisation, so that information on each of the different forms can be combined. On the other hand, sometimes completely different synonymous terms, e.g., *anti-inflammatoir* and *ontstekingsremmende* in Dutch, which both mean *anti-inflammatory*, do have almost identical embeddings. A final potential reason for missing equivalents is when the equivalent is a multi-word term. For instance, the English equivalent for the French *échographie-doppler* is usually written as *Doppler echocardiography*, which means it cannot be found by a system that does not handle multi-word terms.

**In conclusion**, this error analysis showed that the Gold Standard alone cannot capture all relevant information, and it revealed many additional potentially useful (near-)equivalents. While it is impossible to logically explain all errors, the analysis did show a few potential strategies for improvements, such as lemmatisation and improved tokenisation. Handling of multi-word terms is another important and necessary improvement, not only because of missed equivalents seen now,

but also because so many relevant terms consist of multiple words, as illustrated by how many had to be excluded from the current Gold Standard compared to the original ACTER dataset.

## 5. Conclusion and Future Research

In this paper, we propose a fairly intuitive method to create domain-specific dictionaries on-the-fly from Wikipedia pages and cross-lingual links. Our experimental results show a clear improvement in cross-lingual alignment of the embeddings when using a dedicated domain-specific dictionary for the use case of heart failure. Moreover, we demonstrate that it is feasible to collect large domain-specific bilingual dictionaries for other domains as well. A detailed error analysis indicated that the alignments using these domain-specific dictionaries might be further improved by lemmatising before alignments, and building a better and more consistent tokeniser. Even though some of the errors, like alignments with antonyms for instance, may cause issues in a practical setting, many of the errors are semantically related terms that often contain interesting suggestions.

There are many ways to expand upon this work in future research. First, this research presents promising proof-of-concept results, but we would like to evaluate the lexicon creation strategy on other language pairs and domains to test its robustness. Furthermore, since research has shown that a larger dictionary directly correlates with better alignments, it would be interesting to explore dictionaries with higher depths, while filtering words that might be irrelevant using unsupervised clustering.

Second, we will extend our approach to multi-word terms. Multi-word embeddings have been a major sore area for static embeddings in the past, but we plan to explore a simple strategy of adding underscores between the different parts of the multi-word terms.

Third, we would like to experiment with a knowledge distillation approach to distill domain-specific information from multilingual transformers and compare the results for specialised terms with aligned monolingual embeddings.

## References


Artetxe, M., G. Labaka, and E. Agirre (2017), Learning bilingual word embeddings with (almost) no bilingual data, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL, Vancouver, Canada, pp. 451–462.

Artetxe, M., G. Labaka, and E. Agirre (2018a), A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 789–798.

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

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov (2017), Enriching Word Vectors with Subword Information, *Transactions of the Association for Computational Linguistics* **5**, pp. 135–146.

Conneau, A. and G. Lample (2019), Cross-lingual language model pretraining, *in* Wallach, H., H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, Vol. 32, Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/c04c19c2c2474dbf5f7ac4372c5b9af1-Paper.pdf.

Conneau, Alexis, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou (2017), Word translation without parallel data, *arXiv preprint arXiv:1710.04087*.

de Groc, Clément (2011), Babouk: Focused web crawling for corpus compilation and automatic terminology extraction, *in* Boissier, Olivier, Boualem Benatallah, Mike P. Papazoglou, Zbigniew W. Ras, and Mohand-Said Hacid, editors, *Proceedings of the 2011 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2011, Campus Scientifique de la Doua, Lyon, France, August 22-27, 2011*, IEEE Computer Society, pp. 497–498. https://doi.org/10.1109/WI-IAT.2011.253.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019), BERT: Pre-training of deep bidirectional transformers for language understanding, *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)*, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186.

Faruqui, Manaal and Chris Dyer (2014), Improving vector space word representations using multilingual correlation, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, Gothenburg, Sweden, pp. 462–471. https://aclanthology.org/E14-1049.

Gerz, Daniela, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen (2018), On the relation between linguistic typology and (limitations of) multilingual language modeling, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Brussels, Belgium, pp. 316–327. https://aclanthology.org/D18-1029.

Guo, Jiang, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu (2015), Cross-lingual dependency parsing based on distributed representations, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Association for Computational Linguistics, Beijing, China, pp. 1234–1244. https://aclanthology.org/P15-1119.

Jagannatha, Abhyuday, Jinying Chen, and Hong Yu (2015), Mining and ranking biomedical synonym candidates from Wikipedia, *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, Association for Computational Linguistics, Lisbon, Portugal, pp. 142–151. https://aclanthology.org/W15-2619.

Jiang, Chao, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu (2020), Neural CRF Model for Sentence Alignment in Text Simplification, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 7943–7960. https://www.aclweb.org/anthology/2020.acl-main.709.

Karakanta, Alina, Jon Dehdari, and Josef van Genabith (2018), Neural machine translation for low-resource languages without parallel corpora, *Machine Translation* **32** (1-2), pp. 167–189. http://link.springer.com/10.1007/s10590-017-9203-5.

Lample, G. and A. Conneau (2019), Cross-lingual language model pretraining, *CoRR*. http://arxiv.org/abs/1901.07291.

Laville, Martin, Emmanuel Morin, and Philippe Langlais (2022), About Evaluating Bilingual Lexicon Induction, *Proceedings of the 15th Workshop on Building and Using Comparable Corpora (BUCC 2022) @LREC2022*, European Language Resources Association, pp. 8–14.

Liu, Fangyu, Ivan Vulić, Anna Korhonen, and Nigel Collier (2021), Learning domain-specialised representations for cross-lingual biomedical entity linking, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Association for Computational Linguistics, Online, pp. 565–574. https://aclanthology.org/2021.acl-short.72.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov (2019), Roberta: A robustly optimized BERT pretraining approach, *CoRR*. http://arxiv.org/abs/1907.11692.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013), Efficient estimation of word representations in vector space, *Proceedings of the ICLR Workshop Papers*.

Molchanov, Alexander, Vladislav Kovalenko, and Fedor Bykov (2021), PROMT Systems for WMT21 Terminology Translation Task, *Proceedings of the Sixth Conference on Machine Translation (WMT)*, Association for Computational Linguistics, pp. 835–841.

Nguyen, Dong and Jack Grieve (2020), Do word embeddings capture spelling variation?, *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Span, pp. 870–881.

Nguyen, Dong, Arnold Overwijk, Claudia Hauff, Dolf R. B. Trieschnigg, Djoerd Hiemstra, and Franciska de Jong (2009), Wikitranslate: Query translation for cross-lingual information retrieval using only wikipedia, *in* Peters, Carol, Thomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J. F. Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas, and Vivien Petras, editors, *Evaluating Systems for Multilingual and Multimodal Information Access*, Vol. 5706 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 58–65.

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

Rigouts Terryn, Ayla, Véronique Hoste, and Els Lefever (2018), A Gold Standard for Multilingual Automatic Term Extraction from Comparable Corpora: Term Structure and Translation Equivalents, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, European Language Resources Association, pp. 1803–1808.

Shakurova, Lena, Beata Nyari, Chao Li, and Mihai Rotaru (2019), Best practices for learning domain-specific cross-lingual embeddings, *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, Association for Computational Linguistics, Florence, Italy, pp. 230–234. https://aclanthology.org/W19-4327.

Sharoff, Serge (2018), Language Adaptation Experiments via Cross-lingual Embeddings for Related Languages, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association, pp. 844–849.

Sharoff, Serge (2020), Finding next of kin: Cross-lingual embedding spaces for related languages, *Natural Language Engineering* **26** (2), pp. 163–182.

Søgaard, Anders, Željko Agić, Héctor Martínez Alonso, Barbara Plank, Bernd Bohnet, and Anders Johannsen (2015), Inverted indexing for cross-lingual NLP, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Association for Computational Linguistics, Beijing, China, pp. 1713–1722. https://aclanthology.org/P15-1165.

Sorg, P. and P. Cimiano (2012), Exploiting Wikipedia for Cross-lingual and Multilingual Information Retrieval, *Data & Knowledge Engineering* **74**, pp. 26–45. https://linkinghub.elsevier.com/retrieve/pii/S0169023X12000213.

Vulić, I. and M.F. Moens (2016), Bilingual distributed word representations from document-aligned comparable data, *Journal of Artificial Intelligence Research* **55** (1), pp. 953–994.

Wu, Shijie and Mark Dredze (2020), Are all languages created equal in multilingual bert?, *Workshop on Representation Learning for NLP*.