Should All Cross-Lingual Embeddings Speak English?

Antonios Anastasopoulos and Graham Neubig
Language Technologies Institute, Carnegie Mellon University
{aanastas,gneubig}@cs.cmu.edu

Abstract

Most of recent work in cross-lingual word embeddings is severely Anglocentric. The vast majority of lexicon induction evaluation dictionaries are between English and another language, and the English embedding space is selected by default as the *hub* when learning in a multilingual setting. With this work, however, we challenge these practices. First, we show that the choice of hub language can significantly impact downstream lexicon induction and zero-shot POS tagging performance. Second, we both expand a standard English-centered evaluation dictionary collection to include all language pairs using triangulation, and create new dictionaries for under-represented languages.[1] Evaluating established methods over all these language pairs sheds light into their suitability for aligning embeddings from distant languages and presents new challenges for the field. Finally, in our analysis we identify general guidelines for strong cross-lingual embedding baselines, that extend to language pairs that do not include English.

1 Introduction

Continuous vectors for representing words (embeddings) (Turian et al., 2010) have become ubiquitous in modern, neural NLP. Cross-lingual representations (Mikolov et al., 2013) additionally represent words from various languages in a shared continuous space, which in turn can be used for Bilingual Lexicon Induction (BLI). BLI is often the first step towards several downstream tasks such as Part-Of-Speech (POS) tagging (Zhang et al., 2016), parsing (Ammar et al., 2016a), document classification (Klementiev et al., 2012), and machine translation (Irvine and Callison-Burch, 2013; Artetxe et al., 2018b; Lample et al., 2018).

Often, such shared representations are learned with a two-step process, whether under bilingual or multilingual settings (hereinafter BWE and MWE, respectively). First, monolingual word embeddings are learned over

---

[1]Available at `https://github.com/antonisa/embeddings`.

large swaths of text. Such pre-trained word embeddings, such as the fastText Wikipedia vectors (Grave et al., 2018), are available for many languages and are widely used. Second, a mapping between the languages is learned in one of three ways: in a supervised manner if dictionaries or parallel data are available to be used for supervision (Zou et al., 2013), under minimal supervision e.g. using only identical strings (Smith et al., 2017), or even in an unsupervised fashion (Zhang et al., 2017; Conneau et al., 2018). Both in bilingual and multilingual settings, it is common that one of the language embedding spaces is the target to which all other languages get aligned (hereinafter "the *hub*"). We outline the details in Section 2.

Despite all the recent progress in learning cross-lingual embeddings, we identify a major shortcoming to previous work: it is by and large English-centric. Notably, most MWE approaches essentially select English as the *hub* during training by default, aligning all other language spaces to the English one. We argue and empirically show, however, that English is a poor hub language choice. In BWE settings, on the other hand, it is fairly uncommon to denote which of the two languages is the *hub* (often this is implied to be the target language). However, we experimentally find that this choice can greatly impact downstream performance, especially when aligning distant languages.

This Anglocentricity is even more evident at the evaluation stage. The lexica most commonly used for evaluation are the MUSE lexica (Conneau et al., 2018) which cover 45 languages, but with translations only from and into English. Alternative evaluation dictionaries are also very English- and European-centric: (Dinu and Baroni, 2014) report results on English–Italian, (Artetxe et al., 2017) on English–German and English–Finnish, (Zhang et al., 2017) on Spanish–English and Italian–English, and (Artetxe et al., 2018a) between English and Italian, German, Finnish, Spanish, and Turkish. We argue that cross-lingual word embedding mapping methods should look beyond English for their evaluation benchmarks because, compared to all others, English is a language with disproportionately large available data and relatively poor inflectional morphology e.g., it lacks case, gender, and complex verbal inflection systems (Aronoff and Fudeman, 2011). These two factors allow for an

overly easy evaluation setting which does not necessarily generalize to other language pairs. In light of this, equal focus should instead be devoted to evaluation over more diverse language pairs that also include morphologically rich and low-resource languages.

With this work, we attempt to address these shortcomings, providing the following contributions:

- We show that the choice of the *hub* when evaluating on diverse language pairs can lead to significantly different performance for iterative refinement methods that use a symbolic-based seed dictionary (e.g., by more than 10 percentage points for BWE over distant languages). We also show that often English is a suboptimal hub for MWE.

- We identify some general guidelines for choosing a hub language which could lead to stronger performance; *less* isometry between the hub and source and target embedding spaces mildly correlates with performance, as does typological distance (a measure of language similarity based on language family membership trees). For distant languages, multilingual systems should be preferred over bilingual ones if the languages share alphabets, otherwise a bilingual system based on monolingual similarity dictionaries is preferable.

- We provide resources for training and evaluation on language pairs that do not include English. We outline a simple triangulation method with which we extend the MUSE dictionaries to an additional 4704 lexicons covering 50 languages (for a total of 4900 dictionaries, including the original English ones), and we present results on a subset of them. We also create new evaluation lexica for under-resourced, under-represented languages using Azerbaijani, Belarusian, and Galician as our test cases. Finally, we provide recipes for creating such dictionaries for any language pair with available parallel data.

2 Cross-Lingual Word Embeddings and Lexicon Induction

Bilingual Word Embeddings In the supervised BWE setting of Mikolov et al. (2013), given two languages $\mathcal{L} = \{l_1, l_2\}$ and their pre-trained *row-aligned* embeddings $\mathcal{X}_1, \mathcal{X}_2$, respectively, a transformation matrix M is learned such that:

$$M = \arg\min_{M \in \Omega} \|\mathcal{X}_1 - M\mathcal{X}_2\|.$$

The set Ω can potentially impose a constraint over M, such as the very popular constraint of restricting it to be orthogonal (Xing et al., 2015). Previous work has empirically found that this simple formulation is competitive with other more complicated alternatives (Xing et al., 2015). The orthogonality assumption ensures that there exists a closed-form solution through Singular Value Decomposition (SVD) of $\mathcal{X}_1 \mathcal{X}_2^T$.[2] Note that in this case only a single matrix M needs to be learned, because $\|\mathcal{X}_1 - M\mathcal{X}_2\| = \|M^{-1}\mathcal{X}_1 - \mathcal{X}_2\|$, while at the same time a model that minimizes $\|\mathcal{X}_1 - M\mathcal{X}_2\|$ is as expressive as one minimizing $\|M_1\mathcal{X}_1 - M_2\mathcal{X}_2\|$, with half the parameters.

In the minimally supervised or even the unsupervised setting, Zhang et al. (2017) and Conneau et al. (2018) reframe the task as an adversarial game, with a generator aiming to produce a transformation that can fool a discriminator. However, the most popular methods follow an iterative refinement approach (Artetxe et al., 2017). Starting with a seed dictionary (e.g. from identical strings (Zhou et al., 2019) or numerals) an initial mapping is learned in the same manner as in the supervised setting. The initial mapping, in turn, is used to expand the seed dictionary with high confidence word translation pairs. The new dictionary is then used to learn a better mapping, and so forth the iterations continue until convergence. The same iterative approach is followed by Artetxe et al. (2018a), with one important difference that allows their model (`VecMap`) to handle language pairs with different alphabets: instead of identical strings, the seed dictionary is constructed based on the similarity of the monolingual similarity distributions over all words in the vocabulary.[3]

Multilingual Word Embeddings In a multilingual setting, the simplest approach is to use BWE and align all languages into a target language (the *hub*). In this case, for N languages $\mathcal{L} = \{l_1, l_2, \ldots, l_N\}$ on has to learn $N - 1$ bilingual mappings (Ammar et al., 2016b). Rather than using a single hub space, Heyman et al. (2019) propose an incremental procedure that uses an Incremental Hub Space (`IHS`): each new language is included to the multilingual space by mapping it to all languages that have already been aligned (e.g. language l_3 would be mapped to the aligned space of $\{l_1, l_2\}$).

Alternatively, all mappings could be learned jointly, taking advantage of the inter-dependencies between any two language pairs. Importantly, though, there is no closed form solution for learning the joint mapping, hence a solution needs to be approximated with gradient-based methods. The main approaches are:

- Multilingual adversarial training with pseudo-randomized refinement (Chen and Cardie, 2018, `MAT+MPSR`): a generalization of the adversarial approach of Zhang et al. (2017); Conneau et al. (2018) to multiple languages, also combined with an iterative refinement procedure.[4]

- Unsupervised Multilingual Hyperalignment (Alaux et al., 2019, `UMH`): an approach

---

[2] We refer the reader to Mikolov et al. (2013) for details.

[3] We refer the reader to Artetxe et al. (2018a) for details.

[4] `MAT+MPSR` has the beneficial property of being as computationally efficient as learning $O(N)$ mappings (instead of $O(N^2)$). We refer the reader to Chen and Cardie (2018) for exact details.

which maps all languages to a single *hub* space,[5] but also enforces good alignments between all language pairs within this space.

Even though the architecture and modeling approach of all MWE methods are different, they share the same conceptual traits: one of the language spaces remains invariant and all other languages are effectively mapped to it. In all cases, English is by default selected to be the hub. The only exception is the study of triplets alignments in (Alaux et al., 2019), where Spanish is used as the Spanish–French–Portuguese triplet hub.

Lexicon Induction One of the most common downstream evaluation tasks for the learned cross-lingual word mappings is Lexicon Induction (LI), the task of retrieving the most appropriate word-level translation for a query word from the mapped embedding spaces. Specialized evaluation (and training) dictionaries have been created for multiple language pairs. Of these, the MUSE dictionaries (Conneau et al., 2018) are most often used, providing word translations between English (EN) and 48 other high- to mid-resource languages, as well as on all 30 pairs among 6 very similar Romance and Germanic languages (English, French, German, Spanish, Italian, Portuguese).

Given the mapped embedding spaces, the translations are retrieved using a distance metric, with Cross-Lingual Similarity Scaling (Conneau et al., 2018, CSLS) as the most commonly used in the literature. Intuitively, CSLS decreases the scores of pairs that lie in dense areas, increasing the scores of rarer words (which are harder to align). The retrieved pairs are compared to the gold standard and evaluated using precision at k (P@k, evaluating how often the correct translation is within the k retrieved nearest neighbors of the query). Throughout this work we report P@1, which is equivalent to accuracy; we provide P@5 and P@10 results in the Appendix.

3 New LI Evaluation Dictionaries

The typically used evaluation dictionaries cover a narrow breadth of the possible language pairs, with the majority of them focusing in pairs with English (as with the MUSE or Dinu et al. (2015) dictionaries) or among high-resource European languages. Glavaš et al. (2019), for instance, highlighted Anglocentricity as an issue, creating and evaluating on 28 dictionaries between 8 languages (Croatian, English, Finnish, French, German, Italian, Russian, Turkish) based on Google Translate. In addition, Czarnowska et al. (2019) focused on the morphology dimension, creating morphologically complete dictionaries for 2 sets of 5 *genetically related* languages (Romance: French, Spanish, Italian, Portuguese, Catalan; and Slavic: Polish, Czech, Slovak, Russian, Ukrainian). In contrast to these two (very valuable!) works, our method for creating dictionaries
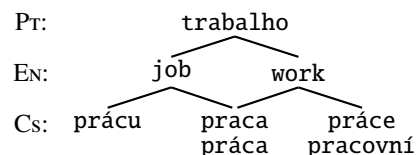

Figure 1: Transitivity example (Portuguese → English → Czech).

for low-resource languages (§3.1) leverages resources that are available for about 300 languages. In addition, we propose a simple triangulation process (§3.2), that makes it possible to create dictionaries for arbitrary language pairs, given that dictionaries into a pivot language (usually English) are available for both languages.

3.1 Low-Resource Language Dictionaries

Our approach for constructing dictionaries is straightforward, inspired by phrase table extraction techniques from phrase-based MT (Koehn, 2009). This is an automatic process, and introduces some degree of noise. Rather than controlling this through manual inspection, which would be impossible for all language pairs, we rely on fairly simple heuristics for controlling the dictionaries' quality.

The first step is collecting publicly available parallel data between English and the low-resource language of interest. We use data from the TED (Qi et al., 2018), OpenSubtitles (Lison and Tiedemann, 2016), WikiMatrix (Schwenk et al., 2019), bible (Malaviya et al., 2017), and JW300 (Agić and Vulić, 2019) datasets.[6] This results in 354k, 53k, and 623k English-to-X parallel sentences for Azerbaijani (AZ), Belarusian (BE), and Galician (GL) respectively.[7] We align the parallel sentences using `fast_align` (Dyer et al., 2013), and extract symmetrized alignments using the `gdfa` heuristic (Koehn et al., 2005). In order to ensure that we do not extract highly domain-specific word pairs, we only use the TED, OpenSubtitles, and WikiMatrix parts for word-pair extraction. Also, in order to control for quality, we only extract word pairs if they appear in the dataset more than 5 times, *and* if the symmetrized alignment probability is higher than 30% in both directions.

With this process, we end up with about 6k, 7k, and 38k word pairs for AZ–EN, BE–EN, and GL–EN respectively. Following standard conventions, we sort the word pairs according to source-side frequency, and use the intermediate-frequency ones for evaluation, typically using the [5000–6500) rank boundaries. The same process can be followed for any language pair with a sufficient volume of parallel data (needed for training a reasonably accurate word alignment model).[8]

---

[5]Note that Alaux et al. (2019) use the term *pivot* to refer to what we refer to as the *hub* language.

[6]Not all languages are available in all these datasets.

[7]The anglocentricity in this step is by necessity – it is hard to find a large volume of parallel data in a language pair excluding English.

[8]In fact, we can produce similar dictionaries for a large number of languages, as the combination of the recently cre-

| Greek | | Italian | | Bridged Greek–Italian Lexicon | | |
word	tag	word	tag	Match	Greek	Italian
ειρηνικός	M;NOM;SG	pacifico	M;SG	M;SG	ειρηνικός	pacifico, ~~pacifici~~, pacifica
ειρηνική	F;NOM;SG	pacifici	M;PL	F;SG	ειρηνική	pacifica, pacifico, ~~pacifici~~
ειρηνικό	Neut;NOM;SG	pacifica	F;SG	SG	ειρηνικό	pacifica, pacifico, ~~pacifici~~
ειρηνικά	Neut;NOM;PL			PL	ειρηνικά	pacifici, ~~pacifica~~, ~~pacifico~~

Table 1: Triangulation and filtering example on Greek–Italian. All words are valid translations of the English word 'peaceful'. We also show ~~filtered-out translations~~.

3.2 Dictionaries for all Language Pairs through Triangulation

Our second method for creating new dictionaries is inspired by phrase table triangulation ideas from the pre-neural MT community (Wang et al., 2006; Levinboim and Chiang, 2015). The concept can be easily explained with an example, visualized in Figure 1. Consider the Portuguese (Pt) word trabalho which, according to the MUSE Pt–En dictionary, has the words job and work as possible En translations. In turn, these two En words can be translated to 4 and 5 Czech (Cs) words respectively. By utilizing the *transitive* property (which translation should exhibit) we can identify the set of 5 possible Cs translations for the Pt word trabalho. Following this simple triangulation approach, we create 4,704 new dictionaries over pairs between the 50 languages of the MUSE dictionaries.[9] For consistency, we keep the same train and test splits as with MUSE, so that the source-side types are equal across all dictionaries with the same source language.

Triangulating through English (which is unavoidable, due to the relative paucity of non-English-centric dictionaries) is suboptimal – English is morphologically poor and lacks corresponding markings for gender, case, or other features that are explicitly marked in many languages. As a result, several inflected forms in morphologically-rich languages map to the same English form. Similarly, gendered nouns or adjectives in gendered languages map to English forms that lack gender information. For example, the MUSE Greek–English dictionary lists the word peaceful as the translation for all ειρηνικός, ειρηνική, ειρηνικό, ειρηνικά, which are the male, female, and neutral (singular and plural) inflections of the same adjective. Equivalently, the English–Italian dictionary translates peaceful into either pacifico, pacifici, or pacifica (male singular, male plural, and female singular, respectively; see Table 1). When translating from or into English *lacking context*, all of those are reasonable translations. When translating between Greek and Italian, though, one should at least take number into account (gram-

matical gender is a more complicated matter: it is not uncommon for word translations to be of different grammatical gender across languages).

Hence, we devise a filtering method for removing blatant mistakes when triangulating morphologically rich languages. We rely on automatic morphological tagging which we can obtain for most of the MUSE languages, using the StanfordNLP toolkit (Qi et al., 2020).[10] The morphological tagging uses the Universal Dependencies feature set (Nivre et al., 2016) making the tagging comparable across almost all languages. Our filtering technique iterates through the bridged dictionaries: for a given source word, if we find a target word with the exact same morphological analysis, we filter out all other translations with the same lemma but different tags. In the case of feature mismatch (for instance, Greek uses 2 numbers, 4 cases and 3 genders while Italian has 2 numbers, 2 genders, and no cases) or if we only find a partial tag match over a feature subset, we filter out translations with disagreeing tags. We ignore the grammatical gender and verb form features, as they are not directly comparable cross-lingually. Coming back to our Greek–Italian example, this means that for the form ειρηνικός we would only keep pacifico as a candidate translation (we show more examples in Table 1).

Our filtering technique removes about 60.4% of the entries in 2964 of the 4900 dictionaries.[11] Unsurprisingly, we find that bridged dictionaries between morphologically rich languages require a lot more filtering. For instance more than 80% of the entries of the Urdu-Greek dictionary get filtered out. On average, the languages with more filtered entries are Urdu (62.4%), Turkish (61.1%), and German (58.6%). On the other hand, much fewer entries are removed from dictionaries with languages like Dutch (36.2%) or English (38.1%). Naturally, this filtering approach is restricted to languages for which a morphological analyzer is available. Mitigating this limitation is beyond the scope of this work, although it is unfortunately a common issue. For example, Kementchedjhieva et al. (2019) manually corrected five dictionaries (between English and German, Danish, Bulgarian, Arabic, Hindi) but one needs to rely

---

ated JW300 and WikiMatrix datasets provide an average of more than 100k parallel sentences in 300 languages. Before publication, we plan to create these dictionaries and make them publicly available, along with the corresponding code.

[9]Available at `https://github.com/antonisa/embeddings`.

[10]The toolkit has since been renamed to Stanza. See `https://stanfordnlp.github.io/stanfordnlp/`.

[11]Due to the lack of morphological analysis tools, we were unable to filter dictionaries in the following 11 languages: aze, bel, ben, bos, lit, mkd, msa, sqi, tam, tha, tel.

| src | Target | | | | | | | | | | μ^{BEST} | μ^{EN} |
	Az	Be	Cs	En	Es	Gl	Pt	Ru	Sk	Tr		
Az	–	17.2[EN]	35.1[ES]	35.7[ES]	48.0[TR]	32.7[RU]	41.5[EN]	29.8[PT]	31.7[CS]	32.0[PT]	33.7	31.7
Be	14.1[CS]	–	35.9[TR]	29.9[PT]	39.5[EN]	25.8[ES]	34.4[ES]	41.1[GL]	30.7[RU]	20.4[PT]	30.2	28.8
Cs	6.9[ES]	9.3[RU]	–	61.0[ES]	60.5[EN]	27.9[PT]	57.8[EN]	45.9[PT]	71.2[EN]	35.8[SK]	41.8	41.2
En	17.9[ES]	18.4[ES]	50.2[ES]	–	77.5[RU]	36.3[ES]	72.3[SK]	43.3[PT]	40.4[TR]	41.9[PT]	44.2	42.7
Es	12.1[EN]	10.1[RU]	47.4[PT]	74.6[SK]	–	37.5[ES]	83.1[GL]	41.9[TR]	40.0[ES]	38.6[SK]	42.8	41.4
Gl	5.5[EN]	3.6[AZ]	26.5[TR]	43.2[ES]	60.8[TR]	–	52.9[CS]	23.8[TR]	26.8[CS]	19.7[CS]	29.2	27.7
Pt	5.8[PT]	8.6[SK]	47.2[GL]	71.3[EN]	88.1[PT]	37.1[ES]	–	38.0[ES]	38.7[ES]	38.1[EN]	41.4	40.4
Ru	8.7[ES]	12.8[AZ]	50.3[GL]	55.5[ES]	54.8[CS]	23.0[PT]	52.4[EN]	–	45.5[TR]	27.0[BE]	36.7	35.9
Sk	4.0[BE]	10.9[RU]	72.5[BE]	55.6[TR]	53.9[EN]	28.4[EN]	52.0[ES]	44.0[GL]	–	28.5[EN]	38.9	37.9
Tr	12.1[SK]	9.0[AZ]	41.8[RU]	51.1[CS]	55.0[EN]	18.4[TR]	51.6[EN]	34.6[EN]	29.4[ES]	–	33.7	33.0
μ^{BEST}	9.7	11.1	45.2	53.1	59.8	29.7	55.3	38.0	39.4	31.3	37.3	
μ^{EN}	9.1	9.9	43.3	51.0	59.3	28.2	54.9	36.5	37.7	30.8		36.0

Table 2: Lexicon Induction performance (measured with P@1) over 10 languages (90 pairs). In each cell, the superscript denotes the hub language that yields the best result for that language pair. μ^{BEST}: average using the best hub language. μ^{EN}: average using the En as the hub. The lightly shaded cells are the language pairs where a bilingual `VecMap` system outperforms `MAT+MSPR`; in heavy shaded cells both `MUSEs` and `VecMap` outperform `MAT+MSPR`.

on automated annotations in order to scale to all languages. Our method that uses automatically obtained morphological information combined with the guidelines proposed by Kementchedjhieva et al. (2019) (e.g. removing proper nouns from the evaluation set) scales easily to multiple languages, allowing us to create more than 4 thousand dictionaries.

4 Lexicon Induction Experiments

The aim of our LI experiments is two-fold. First, the differences in LI performance show the importance of the hub language choice with respect to each evaluation pair. Second, as part of our call for moving beyond Anglo-centric evaluation, we also present LI results on several new language pairs using our triangulated dictionaries.

4.1 Methods and Setup

We train and evaluate all models starting with pre-trained Wikipedia FastText embeddings for all languages (Grave et al., 2018). We focus on the minimally supervised scenario which only uses similar character strings between any languages for supervision in order to mirror the hard, realistic scenario of not having annotated training dictionaries between the languages. We learn MWE with the `MAT+MPSR` method using the publicly available code,[12] aligning several language subsets varying the hub language. We decided against comparing to the incremental hub (`IHS`) method of Heyman et al. (2019), because the order in which the languages are added is an additional hyperparameter that would explode the experimental space.[13] We also do not compare to `UMH`, as we consider it conceptually similar to `MAT+MSPR` and no code is publicly available. For BWE

experiments, we use `MUSEs`[14] (MUSE, semisupervised) and `VecMap`[15] systems, and we additionally compare them to `MAT+MPSR` for completeness.

We compare the statistical significance of the performance difference of two systems using paired bootstrap resampling (Koehn, 2004). Generally, a difference of 0.4–0.5 percentage points evaluated over our lexica is significant with $p < 0.05$.

Experiment 1 We first focus on 10 languages of varying morphological complexity and data availability (which affects the quality of the pre-trained word embeddings): Azerbaijani (Az), Belarusian (Be), Czech (Cs), English (En), Galician (Gl), Portuguese (Pt), Russian (Ru), Slovak (Sk), Spanish (Es), and Turkish (Tr). The choice of these languages additionally ensures that for our three low-resource languages (Az, Be, Gl) we include at least one related higher-resource language (Tr, Ru, Pt/Es respectively), allowing for comparative analysis. Table 2 summarizes the best post-hoc performing systems for this experiment.

Experiment 2 In the second setting, we use a set of 7 more distant languages: English, French (Fr), Hindi (Hi), Korean (Ko), Russian, Swedish (Sv), and Ukrainian (Uk). This language subset has large variance in terms of typology and alphabet. The best performing systems are presented in Table 3.

4.2 Analysis and Takeaways

MWE: English is rarely the best hub language In multilingual settings, we conclude that the standard practice of choosing English as the hub language is sub-optimal. Out of the 90 evaluation pairs from our 10-language experiment (Table 2) the best hub language is English in only 17 instances (less than 20% of the

---

[12] https://github.com/ccsasuke/umwe
[13] We refer the reader to Table 2 from Heyman et al. (2019) which compares to `MAT+MPSR`, and to Table 7 of their appendix which shows the dramatic influence of language order.

[14] https://github.com/facebookresearch/MUSE
[15] https://github.com/artetxem/vecmap

Source	Target							μ^{BEST}	μ^{EN}
	EN	FR	HI	KO	RU	SV	UK		
EN	–	76.3$^{\text{RU}}$	23.9$^{\text{UK}}$	10.4$^{\text{FR}}$	42.0$^{\text{UK}}$	59.0$^{\text{HI}}$	28.3$^{\text{RU}}$	40.0	38.5
FR	74.0$^{\text{UK}}$	–	19.0$^{\text{RU}}$	7.5$^{\text{SV}}$	40.8$^{\text{RU}}$	51.8$^{\text{EN}}$	28.8$^{\text{EN}}$	37.0	36.4
HI	31.4$^{\text{FR}}$	26.9$^{\text{RU}}$	–	2.1$^{\text{EN}}$	14.6$^{\text{UK}}$	17.3$^{\text{EN}}$	10.5$^{\text{FR}}$	17.1	16.2
KO	17.7$^{\text{SV}}$	13.6$^{\text{SV}}$	2.4$^{\text{FR}}$	–	7.9$^{\text{EN}}$	7.2$^{\text{RU}}$	3.6$^{\text{FR}}$	8.8	7.9
RU	53.4$^{\text{KO}}$	51.7$^{\text{KO}}$	15.3$^{\text{UK}}$	5.2$^{\text{EN}}$	–	41.3$^{\text{UK}}$	56.3$^{\text{KO}}$	37.2	36.2
SV	52.7$^{\text{UK}}$	48.2$^{\text{KO}}$	17.7$^{\text{RU}}$	5.1$^{\text{UK}}$	33.2$^{\text{FR}}$	–	24.1$^{\text{RU}}$	30.2	29.2
UK	41.4$^{\text{RU}}$	44.0$^{\text{HI}}$	14.4$^{\text{SV}}$	2.6$^{\text{EN}}$	59.7$^{\text{HI}}$	36.8$^{\text{KO}}$	–	33.2	32.4
μ^{BEST}	45.1	43.5	15.5	5.5	33.0	35.6	25.3	29.1	
μ^{EN}	42.7	42.5	14.5	5.1	32.4	34.9	24.5		28.1

Table 3: Lexicon Induction performance (P@1) over MWEs from 7 typologically distant languages (42 pairs). The lightly shaded cells are the only language pairs where a bilingual `MUSE` system outperforms `MAT+MSPR`; in heavy shaded cells a bilingual `VecMap` (but not `MUSEs`) system outperform `MAT+MSPR`.

time). In fact, the average performance (over all evaluation pairs) when using EN as the hub (denoted as μ^{EN}) is 1.3 percentage points worse than the optimal (μ^{BEST}). In our distant-languages experiment (Table 3) English is the best choice only for 7 of the 42 evaluation pairs (again, less than 20% of the time). As before, using EN as the hub leads to an average drop of one percentage point in performance aggregated over all pairs, compared to the averages of the optimal selection. The rest of this section attempts to provide an explanation for these differences.

Expected gain for a hub language choice As vividly outlined by the superscript annotations in Tables 2 and 3, there is not a single hub language that stands out as the best one. Interestingly, all languages, across both experiments, are the best hub language for some evaluation language pair. For example, in our 10-languages experiment, Es is the best choice for about 20% of the evaluation pairs, TR and EN are the best for about 17% each, while GL and BE are the best for only 5 and 3 language pairs respectively.

Clearly, not all languages are equally suited to be the hub language for many language pairs. Hence, it would be interesting to quantify how much better one could do by selecting the best hub language compared to a random choice. In order to achieve this, we define the expected gain G_l of using language l as follows. Assume that we are interested in mapping N languages into the shared space and p_l^m is the accuracy[16] over a specified evaluation pair m when using language l as the hub. The random choice between N languages will have an expected accuracy equal to the average accuracy when using all languages as hub:

$$\mathbb{E}[p^m] = \frac{\sum_l p_l^m}{N}.$$

The gain for that evaluation dataset m when using language l as hub, then, is $g_l^m = p_l^m - \mathbb{E}[p^m]$. Now, for a collection of M evaluation pairs we simply average their gains, in order to obtain the expected gain for using

---

Figure 2: Expected gain G_l for the MWE experiments.

language l as the hub:

$$G_l = \mathbb{E}[g_l] = \frac{\sum_m g_l^m}{M}.$$

The results of this computation for both sets of experiments are presented in Figure 2. The bars marked `'overall'` match our above definition, as they present the expected gain computed over all evaluation language pairs. For good measure, we also present the average gain per language aggregated over the evaluation pairs where that language was indeed the best hub language (`'when best'` bars). Perhaps unsurprisingly, Az seems to be the worst hub language choice among the 10 languages of the first experiment, with an expected loss (negative gain) of -0.4. This can be attributed to how distant Az is from all other languages, as well as to the fact that the Az pre-trained embeddings are of lower quality compared to all other languages (as the Az Wikipedia dataset is significantly smaller than the others). Similarly, HI and SV show expected loss for our second experiment.

Note that English is not a *bad* hub choice *per se* – it exhibits a positive expected gain in both sets of experiments. However, there are languages with larger expected gains, like Es and GL in the 10-languages experiment that have a twice-as-large expected gain, while RU has a 4 times larger expected gain in the distant-

---

[16]This could be substituted with any evaluation metric.

languages experiment. Of course, the language subset composition of these experiments could possibly impact those numbers. For example, there are three very related languages (Es, Gl, Pt) in the 10 languages set, which might boost the expected gain for that subset; however, the trends stand even if we compute the expected gain over a subset of the evaluation pairs, removing all pairs that include Gl or Pt. For example, after removing all Gl results, Es has a slightly lower expected gain of 0.32, but is still the language with the largest expected gain.

Identifying the best hub language for a given evaluation set The next step is attempting to identify potential characteristics that will allow us make educated decisions with regards to choosing the hub language, given a specific evaluation set. For example, should one choose a language typologically similar to the evaluation source, target, or both? Or should they use the source or the target of the evaluation set as the hub?

Our first finding is that the best performing hub language will very likely be neither the source nor the target of the evaluation set. In our 10-languages experiments, a language different than the source and the target yields the best accuracy for over 93% of the evaluation sets, with the difference being statistically significant in more than half such cases. Similarly, in the distant-languages experiment, there is only a single instance where the best performing hub language is either the source or the target evaluation language (for Fr–Ru), and for the other 97% of cases the best option is a third language. This surprising pattern contradicts the mathematical intuition discussed in Section 2 according to which a model learning a single mapping (keeping another word embedding space fixed) is *as expressive* as a model that learns two mappings for each of the languages. Instead, we find that in almost all cases, learning mappings for both language spaces of interest (hence rotating both spaces) leads to better BLI performance compared to when one of the spaces is fixed.

Our second finding is that the LI performance correlates with measures of distance between languages and language spaces. The typological distance (d^{gen}) between two languages can be approximated through their genealogical distance over hypothesized language family trees, which we obtain from the URIEL typological database (Littell et al., 2017). Also, Patra et al. (2019) recently motivated the use of Gromov-Hausdroff (GH) distance as an *a priori* estimation of how well two language embedding spaces can be aligned under an isometric transformation (an assumption most methods rely on). The authors also note that vector space GH distance correlates with typological language distance.

We find that there is a *positive* correlation between LI performance and the genealogical distances between the source–hub and target–hub languages. The average (over all evaluation pairs) Pearson's correlation coefficient between P@1 and d^{gen} is 0.49 for the distant languages experiment and 0.38 for the 10-languages one. A similar positive correlation of performance and the
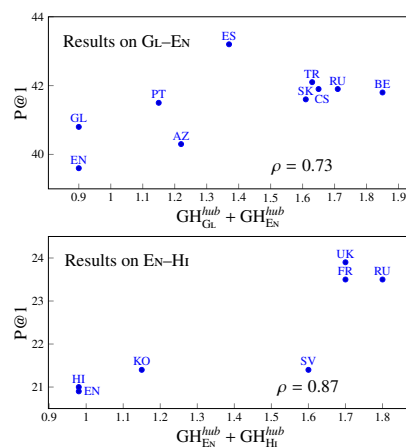
Figure 3: The Lexicon Induction accuracy generally correlates positively with the GH distance of the source and target language vector spaces to the hub language.

sum of the GH distances between the source–hub and target–hub spaces. On our distant languages experiment, the correlation coefficient between P@1 and GH is 0.45, while it is slightly lower (0.34) for our 10-languages experiment. Figure 3 shows two high correlation examples, namely Gl–En and En–Hi.

BWE: The hub matters for distant languages MUSEs implements a provably direction-independent closed form solution of the Procrustes problem, and we confirm empirically that the hub choice does not affect the outcome (we provide complete results on MUSEs in Table 7 in the Appendix). Similarly, because VecMap uses symmetric re-weighting and produces bidirectional dictionaries at its final step, the results are not dependent on the training direction. However, obtaining good performance with such methods requires the orthogonality assumption to hold, which for distant languages is rarely the case (Patra et al., 2019). In fact, we find that the gradient-based MAT+MPSR method in a bilingual setting over typologically distant languages exhibits better performance than MUSEs or VecMap. Across Table 2, in only a handful of examples (shaded cells) do VecMap or MUSEs systems outperform MAT+MPSR for BWE (with the majority being among En, Es, Gl, and Pt, all related high-resource languages).

In the 7 distant languages setting, however, the results are different: VecMap outperforms MUSEs and the multilingual MAT+MPSR in the vast majority of the language pairs. The difference is more stark when the languages of the pair use completely different alphabets, where the same-character strings heuristic for bootstrapping the initial dictionary mapping fails. Instead, the monolingual similarity approach employed by VecMap is definitely more appropriate for settings such as those posed by languages like Korean or Hindi. This highlights the importance of actually evaluating and reporting results on such language pairs.

On the one hand, we find that when aligning distant

Results on Az–Cs				Average
Bilingual with hub:	Az 22.7	Cs 29.1		25.8
Trilingual Az, Cs, +hub:				
Be 21.6	En 28.5	Es 31.8	Gl 23.0	28.2
Pt 29.6	Ru 27.4	Sk 30.4	Tr 32.9	
Trilingual Az, hub:Cs, +extra:				
En 30.1	Es 30.1	Pt 33.2	Ru 27.1	Tr 33.7 → 30.8
Multilingual (10 languages)				
Az 33.7	Be 34.0	Cs 32.3	En 34.5	Es **35.1**
Gl 34.0	Pt **34.8**	Ru 34.5	Sk 32.9	Tr 33.7 → 33.9

Results on Ru–Uk			Average
Bilingual with hub:	Ru 58.0	Uk 57.0	57.5
Trilingual Be, Ru, Uk with hub:			
Be **59.2**	Ru **58.9**	Uk 58.4	58.8
Trilingual Ru, Uk, +hub:			
Az 57.4, Cs 58.5, En 58.4, Es 58.3, Fr 58.0, Hi 57.0, Tr 57.2			57.8
Multilingual Be, Ru, Uk, +hub:			
Cs 58.0, En 58.1, Es 58.5, Gl **58.8**, Ko 57.0, Pt 58.3, Sv 58.2			58.1
Multilingual Ru, Uk, En, Fr, Hi, Ko, Sv, *with hub:*			
En 55.3, Fr 56.1, Hi 55.8, Ko 56.3, Ru 55.3, Sv 55.3, Uk 54.9			55.6

Table 4: Comparison of bilingual, trilingual, and multilingual systems for distant (left) and related (right) languages. Multilinguality boosts performance significantly on distant languages.

Test	Hub SRC	Hub TRG	Test	Hub SRC	Hub TRG
Az–Cs	22.7	29.1	Gl–Pt	53.5	53.6
Az–En	13.2	20.7	Pt–Gl	39.0	36.7
Az–Tr	30.1	30.1	Uk–Ru	61.6	61.8

Table 5: The hub is important for BWE between distant languages with `MAT+MPSR`.

	Transfer from En				Transfer from Pt	
Hub	Es	Pt	Gl	Hub	Es	Gl
En	**38.7**	21.8	19.4	En	**48.4**	32.9
Es	26.5	16.1	28.5†	Es	41.4	25.5†
Pt	28.1	25.7	15.6	Pt	44.3†	**36.5**
Gl	35.4	22.8	23.1	Gl	48.1	23.8
Be	35.6	30.5	13.2			
Ru	28.6†	**30.6**	18.2	†: best train-test hub		
Sk	24.2	30.2†	14.6	for LI.		

Table 6: The choice of hub can significantly affect downstream zero-shot POS tagging accuracy.

languages with `MAT+MPSR`, the difference between hub choices can be significant – in Az–En, for instance, using En as the hub leads to more than 7 percentage points difference compared to using Az. We show some examples in Table 5. On the other hand, when aligning typologically similar languages, the difference is less pronounced. For example, we obtain practically similar performance for Gl–Pt, Az–Tr, or Uk–Ru when using either the source or the target language as the hub. Note, though, that non-negligible differences could still occur, as in the case of Pt–Gl. In most cases, it is the case that the higher-resourced language is a better hub than the lower-resourced one, especially when the number of resources differ significantly (as in the case of Az and Be against any other language). Since BWE settings are not our main focus, we leave an extensive analysis of this observation for future work.

Bi-, tri-, and multilingual systems This part of our analysis compares bilingual, trilingual, and multilingual systems, with a focus on the under-represented languages. Through multiple experiments (complete evaluations are listed in the Appendix) we reach two main conclusions. On one hand, when evaluating on typologically distant languages, one should use as many languages as possible. In Table 4 we present one such example with results on Az–Cs under various settings. On the other hand, when multiple related languages

are available, one can achieve higher performance with multilingual systems containing all related languages and one more hub language, rather than learning diverse multilingual mappings using more languages. We confirm the latter observation with experiments on the Slavic (Be, Ru, Uk) and Iberian (Es, Gl, Pt) clusters, and present an example (Ru–Uk) in Table 4.

5 Downstream Task Experiments

Differences in BLI performance do not necessarily translate to differences in other downstream tasks that use the aligned embeddings, so Glavaš et al. (2019) advocate for actual evaluation on such tasks. We extend our analysis to an example downstream task of zero-shot POS tagging using the aligned embeddings for select language pairs. We show that indeed the choice of the hub language can have dramatic impact. Using Universal Dependencies data (Nivre et al., 2016) we train simple bi-LSTM POS taggers on En and Pt using the respective embeddings produced from each `MAT+MPSR` run, and evaluate the zero-shot performance on Gl and

Es.[17] Although all taggers achieve consistent accuracies > 95% on English and Portuguese regardless of the original Eɴ or Pᴛ embeddings, the zero-shot performance on the test languages, as shown in Table 6, varies widely. For instance, using the embeddings produced from using Pᴛ as a hub, we obtain the highest zero-shot accuracy on Gʟ (36.5%), while using the ones from the Gʟ hub lead to significantly worse performance (23.8%). It should be noted that the best hub for POS-tagging does not always coincide with the best hub for LI, e.g. the best LI hub for Pᴛ–Gʟ is Es, which leads to 11 percentage points worse Gʟ POS tagging performance than the best system. In fact, for the language pairs that we studied we observe no correlation between the two tasks performance as we vary the hub (with an average Spearman's rank correlation $\rho = 0.08$).

6   Conclusion

With this work we challenge the standard practice in learning cross-lingual word embeddings. We empirically show that the choice of the hub language is an important parameter that affects lexicon induction performance in both bilingual (between distant languages) and multilingual settings. More importantly, we hope that by providing new dictionaries and baseline results on several language pairs, we will stir the community towards evaluating all methods in challenging scenarios that include under-represented language pairs. Towards this end, our analysis provides insights and general directions for stronger baselines for non-Anglocentric cross-lingual word embeddings. The problem of identifying the best *hub* language, despite our analysis based on the use of typological distance, remains largely unsolved. In the future, we will investigate a hub language ranking/selection model a la Lin et al. (2019).

Acknowledgements

References

Željko Agić and Ivan Vulić. 2019. Jw300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210.

Jean Alaux, Edouard Grave, Marco Cuturi, and Armand Joulin. 2019. Unsupervised hyperalignment for multilingual word embeddings. In *Proceedings of the International Conference on Learning Representations*.

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A Smith. 2016a. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. 2016b. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*.

Mark Aronoff and Kirsten Fudeman. 2011. *What is morphology?*, volume 8. John Wiley & Sons.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 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)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. 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)*, pages 789–798.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium.

Xilun Chen and Claire Cardie. 2018. Unsupervised multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 261–270. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of the Sixth International Conference on Learning Representations*.

Paula Czarnowska, Sebastian Ruder, Edouard Grave, Ryan Cotterell, and Ann Copestake. 2019. Don't forget the long tail! a comprehensive analysis of morphological generalization in bilingual lexicon 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)*, pages 974–983, Hong Kong, China. Association for Computational Linguistics.

Georgiana Dinu and Marco Baroni. 2014. How to make words with vectors: Phrase generation in distributional semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 624–633.

---

[17]Note that our goal is not to achieve SOTA in zero-shot POS-tagging, but to show that embeddings resulting from different hub choices have different qualities.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of the International Conference on Learning Representations, workshop track*.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 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*, pages 710–721, Florence, Italy. Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Geert Heyman, Bregt Verreet, Ivan Vulić, and Marie-Francine Moens. 2019. Learning unsupervised multilingual word embeddings with incremental multilingual hubs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1890–1902, Minneapolis, Minnesota. Association for Computational Linguistics.

Ann Irvine and Chris Callison-Burch. 2013. Combining bilingual and comparable corpora for low resource machine translation. In *Proceedings of the eighth workshop on statistical machine translation*, pages 262–270.

Yova Kementchedjhieva, Mareike Hartmann, and Anders Søgaard. 2019. Lost in evaluation: Misleading benchmarks for bilingual dictionary induction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*, pages 1459–1474.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Philipp Koehn. 2009. *Statistical machine translation*. Cambridge University Press.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh system description for the 2005 iwslt speech translation evaluation. In *International Workshop on Spoken Language Translation (IWSLT) 2005*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium.

Tomer Levinboim and David Chiang. 2015. Multi-task word alignment triangulation for low-resource languages. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1221–1226, Denver, Colorado. Association for Computational Linguistics.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2015: Extracting large parallel corpora from movie and tv subtitles. In *International Conference on Language Resources and Evaluation*.

Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Copenhagen, Denmark.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. arXiv:1309.4168.

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

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. 2019. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *The 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, Florence, Italy.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. arXiv:2003.07082.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, New Orleans, USA.

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

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of the Fifth International Conference on Learning Representations*.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394, Uppsala, Sweden. Association for Computational Linguistics.

Haifeng Wang, Hua Wu, and Zhanyi Liu. 2006. Word alignment for languages with scarce resources using bilingual corpora of other language pairs. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 874–881, Sydney, Australia. Association for Computational Linguistics.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, Denver, Colorado. Association for Computational Linguistics.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970.

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016. Ten pairs to tag–multilingual pos tagging via coarse mapping between embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1307–1317.

Chunting Zhou, Xuezhe Ma, Di Wang, and Graham Neubig. 2019. Density matching for bilingual word embedding. In *Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Minneapolis, USA.

Will Y Zou, Richard Socher, Daniel Cer, and Christopher D Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398.

A Does evaluation directionality matter?

We also explored whether there are significant differences between the evaluated quality of aligned spaces, when computed on both directions (src–trg and trg–src). We find that the evaluation direction indeed matters a lot, when the languages of the evaluation pair are very distant, in terms of morphological complexity and data availability (which affects the quality of the original embeddings). A prominent example, from our European-languages experiment, are evaluation pairs involving Az or Be. When evaluating on the Az–XX and Be–XX dictionaries, the word translation P@1 is more than 20 percentage points higher than when evaluating on the opposite direction (XX-Az or XX-Be). For example, Es–Az has a mere P@1 of 9.9, while Az–Es achieves a P@1 of 44.9. This observation holds even between very related languages (cf. Ru–Be: 12.8, Be–Ru: 41.1 and Tr–Az: 8.4, Az–Tr: 32.0), which supports our hypothesis that this difference is also due to the quality of the pre-trained embeddings. It is important to note that such directionality differences are not observed when evaluating distant pairs with presumably high-quality pre-trained embeddings e.g. Tr–Sk or Tr–Es; the P@1 for both directions is very close.

B Complete results for all experiments

Here we provide complete evaluation results for our multilingual experiments. Table 7 presents the P@1 of the bilingual experiments using MUSE, and Table 8 presents accuracy using VecMap. Tables 9–14 present P@1, P@5, and P@10 respectively, for the experiment on the 10 European languages. Similarly, results on the distant languages experiment are shown in Tables 15, 16, and 17.

Table 7: BWE results (P@1) with `MUSE`

Source	Target									
	Az	Be	Cs	En	Es	Gl	Pt	Ru	Sk	Tr
Az	–	4.8	21.4	23.6	32.6	13.6	26.7	10.4	15.0	31.8
Be	4.0	–	26.1	3.8	12.3	9.3	11.3	42.0	23.1	2.9
Cs	2.6	5.4	–	57.1	55.5	11.9	52.3	44.7	71.2	31.6
En	12.2	2.5	47.3	–	79.3	32.0	72.9	39.7	34.3	40.6
Es	7.8	2.4	45.0	76.7	–	37.1	83.4	38.9	34.3	38.2
Gl	2.7	1.8	14.0	38.5	61.2	–	53.3	11.4	12.9	8.5
Pt	2.9	2.3	44.9	72.2	88.7	36.3	–	33.7	33.7	34.6
Ru	1.7	12.0	48.6	50.2	49.4	6.6	46.8	–	44.6	21.1
Sk	0.3	5.2	71.8	48.0	46.4	9.3	44.4	43.2	–	21.2
Tr	10.8	0.3	35.8	48.0	50.9	3.5	45.9	26.9	20.3	–

Source	Target						
	En	Fr	Hi	Ko	Ru	Sv	Uk
En	–	80.3	17.9	9.5	39.7	60.0	25.9
Fr	76.6	–	11.9	5.1	38.0	52.4	26.8
Hi	24.2	17.0	–	0.4	3.1	3.3	2.3
Ko	12.4	7.1	0.4	–	2.5	2.2	0.6
Ru	50.2	47.3	3.2	1.6	–	35.8	58.8
Sv	53.3	47.8	5.2	2.3	27.8	–	19.9
Uk	37.4	40.3	4.1	0.3	60.7	30.2	–

Table 8: BWE results (P@1) with `VecMap`

Source	Target									
	Az	Be	Cs	En	Es	Gl	Pt	Ru	Sk	Tr
Az	–	15.86	32.43	32.38	37.81	28.48	37.29	26.58	29.38	28.71
Be	14.41	–	35.31	32.74	43.67	30.56	36.58	43.49	30.0	20.77
Cs	6.78	8.65	–	57.45	56.75	35.66	54.09	44.59	73.49	34.75
En	20.12	18.06	46.41	–	69.91	40.83	63.5	40.13	40.19	37.7
Es	11.66	8.9	45.09	69.49	–	39.37	81.19	40.52	40.7	39.89
Gl	5.34	2.44	29.14	46.11	58.44	–	51.64	26.57	28.53	22.47
Pt	6.72	6.97	43.48	66.21	85.68	41.17	–	38.29	39.81	36.61
Ru	8.06	10.33	52.43	59.03	59.29	29.87	55.55	–	49.93	27.73
Sk	2.92	9.26	70.16	56.73	52.35	36.62	50.96	45.47	–	31.23
Tr	14.2	9.74	42.37	45.51	50.21	28.06	49.11	32.33	34.57	–

Source	Target						
	En	Fr	Hi	Ko	Ru	Sv	Uk
En	–	69.82	35.0	19.2	40.56	56.49	23.63
Fr	68.44	–	28.27	15.53	38.18	49.71	26.95
Hi	44.61	38.52	–	14.01	20.39	26.26	14.72
Ko	32.69	18.32	12.93	–	11.72	18.45	7.21
Ru	59.24	55.18	21.65	10.65	–	47.58	55.12
Sv	51.94	46.92	27.46	12.66	34.29	–	26.96
Uk	42.61	47.82	17.92	5.21	57.64	43.23	–

Table 9: All results from the European-languages MWE experiment: P@1 (part 1).

Test	Hub language										μ
	Az	Be	Cs	En	Es	Gl	Pt	Ru	Sk	Tr	
Az–Be	13.7	12.6	14.2	17.2	16.4	13.9	15.0	15.6	14.5	15.8	14.9
Az–Cs	33.7	34.0	32.3	34.5	35.1	34.0	34.8	34.5	32.9	33.7	33.9
Az–En	31.1	34.7	32.8	32.6	35.7	34.2	33.6	33.6	34.0	33.2	33.5
Az–Es	42.7	46.6	45.2	46.1	44.9	44.4	44.9	43.3	46.1	48.0	45.2
Az–Gl	25.9	27.2	29.0	26.5	29.0	24.7	27.2	32.7	31.5	25.9	28.0
Az–Pt	37.5	41.5	39.3	41.5	39.8	39.0	39.8	41.5	38.5	40.0	39.8
Az–Ru	27.9	27.1	27.1	27.4	27.7	29.0	29.8	26.3	26.3	28.5	27.7
Az–Sk	28.8	30.1	31.7	29.1	30.4	30.4	28.8	28.5	29.5	30.4	29.8
Az–Tr	29.8	30.8	32.0	30.1	31.3	30.8	32.0	31.1	32.0	31.8	31.2
Be–Az	10.4	13.3	14.1	13.0	11.9	12.7	12.4	13.0	13.3	13.0	12.7
Be–Cs	30.5	31.6	33.3	33.0	30.8	31.6	32.5	32.2	33.0	35.9	32.5
Be–En	24.8	26.5	27.8	27.8	28.2	24.8	29.9	28.2	26.5	25.6	27.0
Be–Es	36.4	38.1	36.4	39.5	35.5	38.1	39.0	37.0	36.1	34.4	37.0
Be–Gl	24.4	24.4	22.9	24.9	25.8	22.6	24.9	23.5	22.6	24.4	24.0
Be–Pt	33.2	33.2	32.7	33.7	34.4	31.7	33.9	31.7	31.9	31.4	32.8
Be–Ru	40.9	40.9	40.6	40.3	40.0	41.1	39.1	38.9	39.7	40.0	40.1
Be–Sk	30.1	27.7	30.7	27.4	28.6	29.2	28.9	30.7	27.7	27.4	28.8
Be–Tr	17.7	17.2	18.9	19.9	17.4	18.9	20.4	18.7	16.9	18.4	18.5
Cs–Az	3.5	4.6	4.9	6.0	6.9	4.9	3.7	4.9	4.0	6.0	4.9
Cs–Be	8.6	7.8	8.6	8.6	8.8	7.8	8.8	9.3	9.3	8.6	8.6
Cs–En	59.7	60.5	59.4	59.2	61.0	60.4	60.1	59.7	60.2	58.8	59.9
Cs–Es	59.0	59.1	57.5	60.5	59.2	58.7	58.9	59.6	59.1	57.6	58.9
Cs–Gl	27.1	26.9	27.1	27.6	27.0	21.4	27.9	27.1	26.5	26.1	26.5
Cs–Pt	56.9	55.6	55.4	57.8	55.5	56.9	55.6	57.3	56.1	54.1	56.1
Cs–Ru	44.2	45.5	45.5	45.0	45.5	45.3	45.9	45.0	45.2	45.9	45.3
Cs–Sk	69.8	69.8	70.2	71.2	70.6	70.2	70.4	69.7	68.4	70.2	70.0
Cs–Tr	35.3	35.2	34.6	35.1	34.7	34.7	35.1	35.0	35.8	34.2	35.0
En–Az	15.8	17.7	16.6	17.5	17.9	16.9	17.5	16.1	16.6	17.2	17.0
En–Be	16.4	15.1	17.6	14.9	18.4	17.4	15.6	17.1	15.9	16.4	16.5
En–Cs	49.2	49.0	47.6	47.4	50.2	49.8	50.1	48.3	48.8	49.3	49.0
En–Es	76.3	77.5	77.2	77.0	76.8	76.5	76.6	77.5	77.3	76.6	76.9
En–Gl	35.0	35.8	36.0	35.2	36.3	31.9	35.9	36.2	35.3	35.0	35.3
En–Pt	71.3	71.8	71.3	72.1	71.5	72.0	71.0	71.5	72.3	71.3	71.6
En–Ru	42.5	43.3	42.7	40.8	43.1	43.3	43.3	41.3	41.4	42.8	42.4
En–Sk	38.7	39.6	40.2	38.0	40.4	39.3	38.5	38.6	36.8	40.4	39.0
En–Tr	40.5	41.7	41.3	41.6	39.4	40.9	41.9	41.0	41.3	40.9	41.0
Es–Az	8.4	10.8	9.0	12.1	10.5	10.5	10.8	9.6	11.8	11.8	10.5
Es–Be	9.9	7.2	8.5	9.3	7.5	9.9	9.9	10.1	9.1	8.8	9.0
Es–Cs	45.3	46.0	44.2	43.4	45.8	45.5	47.4	46.3	45.4	44.7	45.4
Es–En	73.0	74.5	73.8	73.2	74.0	74.1	73.1	73.5	74.6	73.6	73.7
Es–Gl	37.1	37.0	37.1	36.9	37.5	33.7	36.8	37.0	36.8	36.7	36.7
Es–Pt	82.1	82.9	82.7	83.0	83.1	83.1	82.5	83.0	82.9	83.0	82.8
Es–Ru	41.4	41.5	41.2	39.4	41.3	41.9	40.9	40.3	40.2	41.9	41.0
Es–Sk	37.0	39.2	38.8	37.4	40.0	39.2	39.5	39.5	35.2	38.8	38.5
Es–Tr	37.5	38.0	37.7	38.2	37.6	37.8	38.4	37.8	38.6	37.9	38.0

8671

Table 10: All results from the European-languages MWE experiment: P@1 (part 2).

Test	Hub language										μ
	Az	Be	Cs	En	Es	Gl	Pt	Ru	Sk	Tr	
Gl–Az	4.0	4.6	4.3	5.5	5.0	4.1	5.2	4.7	4.8	5.0	4.7
Gl–Be	3.6	3.0	2.4	3.0	3.0	2.4	3.0	2.4	1.2	3.0	2.7
Gl–Cs	23.2	25.7	25.0	23.8	26.5	23.0	25.6	25.4	25.6	26.5	25.0
Gl–En	40.3	41.8	41.9	39.6	43.2	40.8	41.5	41.9	41.6	42.1	41.5
Gl–Es	60.0	60.5	60.1	59.9	60.4	59.0	60.0	60.3	59.6	60.8	60.1
Gl–Pt	52.5	52.5	52.9	52.0	52.0	50.4	52.5	51.9	52.1	52.0	52.1
Gl–Ru	22.5	22.7	22.9	21.7	23.3	21.9	23.7	22.7	22.5	23.8	22.8
Gl–Sk	26.0	26.3	26.8	25.6	26.4	23.4	25.5	25.1	23.2	26.4	25.5
Gl–Tr	18.5	19.3	19.7	18.6	17.8	18.3	18.9	19.2	19.4	17.6	18.7
Pt–Az	3.8	4.7	5.8	5.0	5.0	3.2	5.8	5.0	5.5	4.7	4.8
Pt–Be	7.3	5.3	7.3	7.3	6.1	7.1	6.8	6.1	8.6	7.1	6.9
Pt–Cs	45.5	47.0	46.3	45.0	45.5	47.2	45.5	46.7	46.5	45.6	46.1
Pt–En	69.9	70.9	70.2	71.3	71.1	70.5	70.6	71.3	70.6	70.8	70.7
Pt–Es	87.4	88.1	87.7	87.6	88.0	87.4	88.1	87.8	87.6	88.1	87.8
Pt–Gl	35.7	36.9	36.3	36.3	37.1	32.7	36.0	35.9	35.2	36.4	35.8
Pt–Ru	37.4	37.7	36.4	36.5	38.0	38.0	36.2	37.0	37.1	37.4	37.2
Pt–Sk	37.6	37.0	37.3	36.7	38.7	37.7	38.3	37.9	33.6	38.0	37.3
Pt–Tr	36.5	37.4	37.2	38.1	35.9	36.4	35.5	37.2	36.2	36.3	36.7
Ru–Az	5.0	6.4	6.2	7.8	8.7	7.3	7.5	7.3	6.7	7.5	7.0
Ru–Be	12.8	9.9	10.7	11.5	11.2	11.0	11.5	12.3	11.0	11.8	11.4
Ru–Cs	49.2	50.0	49.2	50.1	49.7	50.3	50.3	49.8	50.1	50.1	49.9
Ru–En	53.6	53.8	54.4	52.7	54.7	55.5	54.8	52.0	54.5	55.5	54.1
Ru–Es	53.7	53.4	54.8	54.5	52.3	53.5	54.0	53.2	53.9	51.2	53.4
Ru–Gl	20.9	21.3	22.1	22.3	22.9	17.2	23.0	21.8	21.7	21.9	21.5
Ru–Pt	50.4	50.3	50.4	52.4	51.1	51.1	49.6	49.8	51.0	47.6	50.4
Ru–Sk	45.0	44.7	44.7	45.2	45.2	44.7	44.3	43.7	43.7	45.5	44.7
Ru–Tr	25.9	27.0	26.2	26.9	26.0	25.9	26.1	25.6	26.8	24.7	26.1
Sk–Az	2.8	4.0	1.5	3.7	2.1	2.8	3.4	3.1	1.8	3.4	2.9
Sk–Be	10.2	7.5	9.9	9.4	9.6	8.3	10.4	10.9	10.9	9.1	9.6
Sk–Cs	71.4	72.5	70.9	70.8	70.5	71.1	71.3	70.6	71.0	71.4	71.1
Sk–En	54.8	55.0	54.0	52.9	55.4	54.7	54.8	54.6	53.0	55.6	54.5
Sk–Es	52.5	51.6	52.2	53.9	52.3	52.0	50.4	50.5	51.5	51.1	51.8
Sk–Gl	27.0	27.3	27.2	28.4	27.8	20.6	26.2	26.0	27.0	27.0	26.4
Sk–Pt	49.3	50.3	48.2	50.4	52.0	49.2	49.1	48.7	48.5	47.7	49.3
Sk–Ru	43.8	43.4	43.5	43.2	43.7	44.0	42.8	42.9	41.2	43.4	43.2
Sk–Tr	28.2	27.5	27.2	28.5	27.1	26.1	26.2	27.6	27.4	26.0	27.2
Tr–Az	9.8	12.1	10.1	11.1	10.1	11.4	11.4	10.8	12.1	11.1	11.0
Tr–Be	9.0	4.8	8.7	8.1	7.8	7.5	8.1	6.9	7.5	7.2	7.6
Tr–Cs	40.3	41.6	40.3	41.6	41.6	40.8	41.6	41.8	40.9	39.2	41.0
Tr–En	51.1	49.3	51.1	50.2	50.4	48.5	50.5	50.2	50.7	50.1	50.2
Tr–Es	53.8	53.6	55.0	55.0	52.5	53.0	54.6	52.9	54.1	53.3	53.8
Tr–Gl	17.0	17.3	17.3	15.9	16.8	11.6	17.5	17.1	17.1	18.4	16.6
Tr–Pt	50.1	50.1	51.4	51.6	49.3	48.9	48.7	49.9	50.5	49.5	50.0
Tr–Ru	34.0	34.3	32.3	34.6	34.3	33.6	33.2	32.0	33.0	32.9	33.4
Tr–Sk	27.5	29.2	27.9	28.5	29.4	27.7	27.9	27.5	25.2	27.9	27.9

Table 11: All results from the European-languages MWE experiment: P@5 (part 1).

Test	Hub language										μ
	Az	Be	Cs	En	Es	Gl	Pt	Ru	Sk	Tr	
Az–Be	26.0	22.5	26.5	26.0	26.5	25.2	25.7	26.0	25.7	25.7	25.6
Az–Cs	53.4	54.8	53.7	57.5	54.8	55.9	55.6	54.5	53.2	54.8	54.8
Az–En	44.7	48.0	47.6	45.7	45.9	47.4	46.8	46.3	46.1	47.2	46.6
Az–Es	60.1	62.6	60.7	62.6	60.7	60.4	60.7	62.4	61.8	62.9	61.5
Az–Gl	38.3	37.7	40.1	41.4	38.9	35.8	40.1	41.4	38.9	39.5	39.2
Az–Pt	52.8	55.3	55.3	56.3	55.8	55.3	55.8	57.8	55.3	56.8	55.7
Az–Ru	45.2	46.5	46.8	48.1	49.2	47.3	48.4	45.5	46.8	50.0	47.4
Az–Sk	43.9	46.1	47.0	48.3	49.2	48.3	49.2	48.3	46.7	46.7	47.4
Az–Tr	45.2	49.1	51.3	49.1	46.7	48.7	49.1	49.4	49.6	49.4	48.8
Be–Az	20.6	20.6	23.4	23.2	24.6	22.0	22.9	24.9	22.3	24.6	22.9
Be–Cs	44.5	44.8	47.6	48.5	46.5	47.9	48.7	46.8	45.7	47.9	46.9
Be–En	42.3	42.3	42.7	41.5	44.4	42.7	42.3	42.7	41.0	43.2	42.5
Be–Es	50.4	53.0	54.2	53.3	50.4	53.6	54.4	51.0	54.2	52.4	52.7
Be–Gl	38.8	36.5	37.7	38.8	38.0	36.5	38.3	38.0	38.6	37.7	37.9
Be–Pt	49.5	50.8	52.8	51.5	52.0	50.0	49.0	49.0	50.5	49.5	50.5
Be–Ru	53.0	53.2	52.1	51.8	53.8	52.7	53.0	53.0	53.2	51.8	52.8
Be–Sk	43.8	40.1	44.7	43.5	41.6	43.8	44.4	43.5	40.1	43.5	42.9
Be–Tr	33.4	33.2	34.6	37.8	32.2	34.4	36.9	33.4	33.2	32.2	34.1
Cs–Az	10.3	11.2	11.2	13.8	14.1	11.8	12.1	10.6	11.2	12.6	11.9
Cs–Be	14.8	15.5	15.5	16.3	16.3	16.6	16.1	16.1	14.8	15.8	15.8
Cs–En	75.6	76.4	75.1	75.7	76.2	76.9	76.1	75.8	75.9	76.0	76.0
Cs–Es	75.5	75.3	74.1	76.5	75.9	74.9	74.3	75.5	75.9	74.1	75.2
Cs–Gl	40.8	41.8	43.0	43.7	43.1	36.5	42.1	42.6	42.1	41.2	41.7
Cs–Pt	72.9	74.1	72.2	74.3	73.1	73.7	72.7	73.8	72.7	71.6	73.1
Cs–Ru	64.5	64.4	63.6	63.9	63.9	64.5	64.9	64.5	64.3	65.5	64.4
Cs–Sk	81.7	82.9	83.2	82.8	82.5	83.0	83.2	82.7	81.6	82.7	82.6
Cs–Tr	56.2	56.0	55.1	57.1	56.4	54.2	54.9	55.5	54.9	53.8	55.4
En–Az	28.3	29.1	30.3	29.9	28.9	29.2	30.2	29.1	28.8	30.6	29.4
En–Be	32.8	28.3	34.0	31.5	34.0	34.5	30.3	32.8	33.3	32.8	32.4
En–Cs	74.7	74.9	73.4	74.5	76.1	76.5	74.8	75.1	73.8	75.5	74.9
En–Es	88.9	89.5	88.8	89.3	89.1	89.3	89.1	89.3	89.0	89.1	89.1
En–Gl	49.0	50.4	50.5	50.4	51.3	47.8	50.9	51.4	49.1	50.7	50.1
En–Pt	86.0	86.6	86.2	86.6	86.2	86.4	86.3	86.3	86.4	85.8	86.3
En–Ru	68.0	68.1	68.2	66.0	68.6	69.6	68.7	67.7	67.4	68.2	68.1
En–Sk	62.3	62.7	62.5	60.8	62.5	62.1	63.5	62.7	59.9	63.2	62.2
En–Tr	63.6	62.6	64.3	62.4	62.4	63.8	63.8	63.0	63.2	63.2	63.2
Es–Az	16.3	16.9	16.9	17.5	18.4	17.8	17.2	17.2	19.0	18.1	17.5
Es–Be	16.8	15.5	17.1	18.9	16.3	18.9	18.7	17.1	18.1	16.5	17.4
Es–Cs	64.4	65.7	63.5	65.2	66.1	65.5	65.9	66.0	65.8	65.9	65.4
Es–En	85.2	86.3	86.0	85.5	85.8	85.5	85.8	86.1	86.0	86.0	85.8
Es–Gl	45.6	46.0	45.7	46.1	46.4	43.2	45.9	45.7	45.8	46.2	45.7
Es–Pt	90.8	91.1	90.7	91.3	91.4	91.1	91.3	90.7	90.9	90.9	91.0
Es–Ru	61.5	62.5	61.4	62.5	62.1	61.7	62.2	60.8	61.6	62.9	61.9
Es–Sk	57.9	59.1	58.7	58.5	59.1	57.8	58.1	57.6	57.0	58.5	58.2
Es–Tr	57.0	57.4	57.2	56.7	55.0	56.3	56.3	55.5	56.6	56.5	56.5

Table 12: All results from the European-languages MWE experiment: P@5 (part 2).

Test	Hub language										μ
	Az	Be	Cs	En	Es	Gl	Pt	Ru	Sk	Tr	
Gl–Az	8.4	9.0	8.8	9.8	9.6	10.0	9.7	9.4	9.2	9.7	9.4
Gl–Be	7.3	6.1	6.1	6.7	6.7	6.7	7.9	6.1	6.1	7.3	6.7
Gl–Cs	41.8	42.1	43.0	42.3	44.5	40.2	42.5	42.5	42.0	43.0	42.4
Gl–En	56.8	57.4	58.6	56.3	59.7	57.6	57.2	57.8	56.7	58.1	57.6
Gl–Es	68.3	68.8	68.1	68.8	68.6	67.9	68.3	68.8	68.2	68.8	68.5
Gl–Pt	63.9	64.3	63.4	64.1	63.2	62.8	63.4	64.0	63.7	63.9	63.7
Gl–Ru	40.2	39.8	39.3	39.6	39.5	37.0	40.0	39.5	39.3	40.8	39.5
Gl–Sk	41.6	42.4	41.1	41.9	43.7	38.5	41.0	41.4	39.2	41.5	41.2
Gl–Tr	33.5	33.4	34.9	33.9	33.3	29.4	32.4	32.6	34.0	31.5	32.9
Pt–Az	8.7	11.1	10.2	12.5	11.1	10.2	10.5	9.9	12.0	11.1	10.7
Pt–Be	14.4	12.1	14.4	17.4	14.1	15.9	14.9	14.9	14.9	14.6	14.8
Pt–Cs	65.6	66.6	64.7	65.8	66.5	66.6	65.9	66.3	65.5	65.1	65.9
Pt–En	81.3	82.1	82.0	82.1	81.9	82.0	81.5	81.7	81.5	82.0	81.8
Pt–Es	92.1	92.6	92.4	92.1	92.0	91.8	92.4	92.4	92.0	92.3	92.2
Pt–Gl	45.4	46.4	46.2	46.9	46.8	43.5	45.8	45.4	45.2	46.7	45.8
Pt–Ru	57.6	57.8	57.7	58.7	58.1	58.5	57.0	57.5	57.6	57.6	57.8
Pt–Sk	57.2	56.9	57.0	57.8	56.6	55.4	56.6	56.8	53.1	56.4	56.4
Pt–Tr	53.9	54.8	54.2	56.3	53.3	53.6	52.7	54.5	54.4	54.6	54.2
Ru–Az	12.0	15.6	15.9	15.6	15.9	14.8	15.4	14.2	14.2	15.9	15.0
Ru–Be	20.1	18.3	20.6	20.1	20.9	20.6	20.6	20.9	21.1	20.4	20.4
Ru–Cs	65.7	65.0	65.1	64.7	65.0	66.7	66.1	65.8	65.1	65.5	65.5
Ru–En	72.8	73.0	73.9	72.0	73.8	73.5	72.7	72.3	72.9	73.5	73.0
Ru–Es	70.1	69.8	69.7	71.3	69.2	70.3	71.2	68.8	70.7	68.4	69.9
Ru–Gl	36.1	35.9	36.1	36.8	37.1	30.9	36.5	36.6	35.9	35.3	35.7
Ru–Pt	66.8	66.8	67.0	69.3	67.9	67.6	65.8	66.6	67.3	65.2	67.0
Ru–Sk	61.1	62.6	61.4	61.1	62.0	61.8	61.8	60.9	59.8	61.6	61.4
Ru–Tr	48.0	48.0	47.6	49.9	47.1	47.5	48.0	46.0	47.0	47.4	47.7
Sk–Az	7.7	9.2	7.1	9.5	7.4	8.3	8.9	8.9	8.3	8.6	8.4
Sk–Be	17.4	16.7	18.5	18.2	17.7	18.5	18.2	19.3	19.3	18.5	18.2
Sk–Cs	82.1	82.1	81.3	81.6	82.1	82.4	81.6	81.6	81.3	81.9	81.8
Sk–En	70.7	71.7	71.3	69.6	71.2	71.4	71.5	70.9	70.3	71.4	71.0
Sk–Es	69.2	69.7	70.2	71.2	70.1	68.8	70.0	68.6	69.2	69.4	69.6
Sk–Gl	43.4	43.3	42.9	45.1	43.7	36.0	42.9	42.0	43.0	42.7	42.5
Sk–Pt	68.2	67.5	67.5	68.7	69.9	67.6	66.1	67.6	66.7	66.7	67.7
Sk–Ru	59.2	58.1	58.2	58.8	59.4	59.5	58.8	58.5	57.5	59.5	58.8
Sk–Tr	47.2	48.7	47.6	48.7	47.1	46.7	48.2	47.8	46.7	46.2	47.5
Tr–Az	19.5	22.2	19.9	21.2	20.9	20.9	20.5	19.5	21.9	20.2	20.7
Tr–Be	17.1	12.3	16.2	17.1	16.8	15.6	16.5	16.5	16.2	16.2	16.1
Tr–Cs	61.6	62.1	60.1	61.8	62.4	61.9	61.6	61.5	61.4	60.1	61.4
Tr–En	68.0	68.2	68.1	67.2	67.8	67.5	69.6	67.7	67.9	67.2	67.9
Tr–Es	69.8	69.0	70.4	70.5	68.0	69.2	70.5	69.4	69.8	69.5	69.6
Tr–Gl	30.5	30.7	31.1	30.0	30.4	23.6	31.4	31.1	29.7	30.7	29.9
Tr–Pt	67.1	66.9	66.9	67.9	66.5	65.9	65.2	67.1	67.5	66.6	66.8
Tr–Ru	55.4	55.9	54.0	55.4	55.3	55.1	55.1	53.0	52.9	53.5	54.6
Tr–Sk	48.2	49.9	48.9	49.7	48.7	47.8	48.9	48.1	44.2	47.7	48.2

Table 13: All results from the European-languages MWE experiment: P@10 (part 1).

Test	Az	Be	Cs	En	Es	Gl	Pt	Ru	Sk	Tr	μ
					Hub language						
Az–Be	31.1	27.1	30.8	31.4	31.9	31.1	29.8	30.3	32.2	31.1	30.7
Az–Cs	60.3	62.5	60.8	62.7	63.6	61.4	62.7	61.1	60.3	63.6	61.9
Az–En	49.3	51.1	52.6	50.5	49.5	50.7	51.4	50.3	50.1	50.7	50.6
Az–Es	63.8	65.7	65.4	67.1	65.2	66.3	68.0	64.6	66.6	67.4	66.0
Az–Gl	42.6	42.6	45.1	45.1	43.8	39.5	45.1	43.8	42.6	43.8	43.4
Az–Pt	58.5	61.2	62.7	62.5	61.5	61.7	61.0	61.2	61.7	62.5	61.5
Az–Ru	50.8	52.7	52.9	50.8	54.0	53.2	54.3	51.6	51.9	54.5	52.7
Az–Sk	48.9	52.0	53.0	52.0	53.9	54.2	53.0	52.4	51.7	51.7	52.3
Az–Tr	53.3	55.5	56.7	57.0	55.0	55.3	55.7	56.5	57.0	56.7	55.9
Be–Az	25.7	25.4	29.7	28.5	29.4	26.8	27.7	28.2	26.8	28.0	27.6
Be–Cs	50.7	51.0	52.1	51.3	51.8	53.8	52.7	51.8	50.7	51.8	51.8
Be–En	46.6	48.7	50.0	46.2	48.3	50.9	46.2	48.3	46.2	47.9	47.9
Be–Es	54.7	57.3	58.7	58.7	56.2	57.9	57.9	55.9	58.5	57.9	57.4
Be–Gl	47.0	45.2	44.6	46.1	43.8	41.4	43.5	43.8	44.3	42.9	44.3
Be–Pt	55.3	55.8	57.0	57.8	57.0	56.5	55.8	54.5	55.5	56.0	56.1
Be–Ru	56.3	56.3	56.1	56.1	56.9	56.1	56.3	56.3	56.9	55.5	56.3
Be–Sk	48.0	45.6	48.3	47.7	48.0	48.6	49.8	48.6	46.2	48.0	47.9
Be–Tr	38.3	40.5	41.5	43.2	40.3	40.3	41.8	41.5	40.3	38.3	40.6
Cs–Az	13.8	14.9	15.5	16.1	17.5	14.9	15.8	14.1	14.9	15.5	15.3
Cs–Be	18.9	17.9	19.2	19.9	19.4	19.9	19.9	19.2	17.9	19.2	19.1
Cs–En	80.2	80.5	79.8	80.0	80.1	81.0	80.2	80.5	80.5	81.1	80.4
Cs–Es	80.1	79.6	78.8	80.0	79.9	79.4	79.9	79.3	80.2	79.0	79.6
Cs–Gl	47.2	48.0	47.9	49.9	49.3	42.4	48.2	48.3	49.1	47.1	47.7
Cs–Pt	77.5	78.7	77.5	78.3	77.1	77.7	76.9	77.7	76.9	76.8	77.5
Cs–Ru	70.1	70.3	69.1	69.6	69.4	70.7	69.6	69.5	69.5	70.5	69.8
Cs–Sk	85.5	85.6	85.7	85.2	84.9	85.1	86.2	85.2	84.9	85.6	85.4
Cs–Tr	63.2	62.7	62.5	63.5	62.7	62.5	62.7	63.4	62.6	61.6	62.7
En–Az	32.2	33.3	34.3	34.3	33.8	32.5	34.4	33.0	34.3	33.8	33.6
En–Be	38.5	34.0	40.4	39.0	40.0	41.2	38.7	38.2	38.7	38.5	38.7
En–Cs	81.2	81.1	79.9	80.7	81.9	82.5	80.6	80.7	80.7	81.5	81.1
En–Es	91.3	92.1	91.7	91.5	91.9	91.7	91.8	91.6	91.9	91.7	91.7
En–Gl	53.9	56.3	56.4	55.7	55.8	53.2	55.9	56.2	54.9	55.5	55.4
En–Pt	89.4	90.0	89.2	89.5	89.1	89.5	89.3	89.0	89.4	89.0	89.3
En–Ru	74.6	74.0	75.8	72.2	74.8	76.0	74.8	73.8	74.0	74.4	74.4
En–Sk	69.3	69.7	69.9	68.0	69.6	68.7	69.9	69.5	67.1	69.9	69.2
En–Tr	69.9	70.1	71.0	69.3	69.5	69.8	70.3	71.1	70.0	69.2	70.0
Es–Az	20.2	20.8	20.2	21.1	20.8	20.2	19.3	20.2	21.1	21.1	20.5
Es–Be	20.8	18.9	20.8	22.9	21.3	22.4	21.1	23.2	21.3	21.3	21.4
Es–Cs	70.5	70.7	70.8	70.9	71.0	71.1	71.3	71.8	72.2	70.9	71.1
Es–En	88.5	88.4	88.5	88.3	88.5	88.5	88.5	88.5	88.5	88.4	88.5
Es–Gl	49.5	49.4	49.4	49.8	50.0	46.0	49.6	49.6	49.4	50.2	49.3
Es–Pt	92.7	92.5	92.5	92.5	93.0	92.9	92.8	92.4	92.1	92.7	92.6
Es–Ru	67.5	67.1	67.4	68.9	67.4	67.6	67.8	66.8	68.7	68.5	67.8
Es–Sk	64.5	64.3	63.9	65.4	65.4	63.5	64.3	64.8	63.0	63.8	64.3
Es–Tr	63.6	63.8	64.3	62.7	61.6	62.6	63.7	62.2	63.8	61.7	63.0

Table 14: All results from the European-languages MWE experiment: P@10 (part 2).

Test	Hub language										μ
	Az	Be	Cs	En	Es	Gl	Pt	Ru	Sk	Tr	
Gl–Az	11.5	11.2	11.1	12.5	12.6	12.3	13.1	12.1	12.5	12.3	12.1
Gl–Be	8.5	7.3	8.5	9.1	8.5	7.9	7.9	7.9	8.5	9.7	8.4
Gl–Cs	48.0	49.0	48.8	49.0	50.7	46.6	48.3	49.1	49.0	49.0	48.8
Gl–En	64.1	64.4	64.7	62.2	64.4	62.5	63.4	64.4	62.4	63.0	63.6
Gl–Es	71.3	71.5	71.5	72.1	71.7	71.1	71.0	71.6	71.4	72.5	71.6
Gl–Pt	66.9	67.1	67.4	67.6	67.5	67.7	67.1	67.6	66.8	68.1	67.4
Gl–Ru	46.7	46.5	45.9	45.0	46.3	42.8	45.8	44.8	44.7	45.7	45.4
Gl–Sk	48.2	48.1	47.2	48.5	48.8	45.3	47.6	46.7	45.5	48.2	47.4
Gl–Tr	39.7	39.3	39.3	39.1	38.2	35.9	38.8	38.9	38.3	38.0	38.5
Pt–Az	11.7	14.6	13.4	14.6	15.2	12.5	13.4	13.1	13.4	15.7	13.8
Pt–Be	18.9	17.2	18.2	21.0	18.7	20.2	18.7	19.7	18.4	18.7	19.0
Pt–Cs	71.6	72.0	70.6	71.7	71.7	72.0	71.5	71.9	71.2	70.7	71.5
Pt–En	84.0	84.3	84.1	85.1	84.2	84.9	84.1	83.9	84.7	84.3	84.4
Pt–Es	92.8	93.2	93.2	93.2	93.6	93.0	93.4	93.3	93.2	93.4	93.2
Pt–Gl	49.3	49.6	48.9	50.1	49.9	46.8	49.3	48.9	47.9	49.6	49.0
Pt–Ru	63.6	64.3	62.8	64.7	64.4	64.3	63.0	63.4	63.8	62.4	63.7
Pt–Sk	63.6	62.4	62.6	63.9	63.0	62.6	62.4	62.1	59.7	62.2	62.4
Pt–Tr	60.4	60.8	60.4	62.3	59.5	60.4	60.3	60.9	60.5	60.9	60.6
Ru–Az	15.4	17.0	18.7	20.1	18.4	18.4	19.0	17.9	17.3	19.8	18.2
Ru–Be	25.1	22.2	24.5	23.8	24.3	24.0	24.5	24.3	25.3	24.3	24.2
Ru–Cs	70.8	70.3	70.9	70.4	70.8	71.3	71.0	70.5	70.8	71.1	70.8
Ru–En	76.9	77.8	78.6	76.6	78.4	77.8	77.4	76.8	77.1	77.5	77.5
Ru–Es	75.2	75.2	75.3	76.3	75.6	75.3	76.3	74.8	76.4	74.5	75.5
Ru–Gl	43.1	42.2	42.1	43.3	43.5	37.1	41.9	41.7	41.3	40.5	41.7
Ru–Pt	72.6	71.8	72.6	74.5	72.5	72.6	71.5	71.5	72.2	70.2	72.2
Ru–Sk	65.5	66.8	66.3	66.5	66.3	66.4	67.0	66.5	64.7	66.9	66.3
Ru–Tr	56.1	56.2	55.2	57.7	56.8	57.0	56.1	54.8	57.3	54.8	56.2
Sk–Az	11.0	11.0	10.7	13.8	10.7	13.2	13.2	10.4	11.3	12.0	11.7
Sk–Be	23.2	20.8	21.1	22.1	21.1	22.9	22.7	22.9	23.4	22.1	22.2
Sk–Cs	85.1	85.5	84.6	84.4	85.3	85.9	85.6	84.9	85.0	85.0	85.1
Sk–En	74.5	76.3	76.6	73.9	75.7	76.0	75.6	75.4	75.3	75.8	75.5
Sk–Es	75.7	75.5	74.9	76.2	74.4	74.2	74.6	74.4	74.7	74.7	74.9
Sk–Gl	49.1	48.7	48.9	51.7	50.1	40.9	49.4	48.5	49.6	49.7	48.7
Sk–Pt	73.7	73.2	72.6	74.7	74.0	73.1	71.7	72.8	72.9	72.0	73.1
Sk–Ru	63.5	64.4	62.8	64.0	64.0	64.2	64.0	62.6	62.6	64.6	63.7
Sk–Tr	55.4	57.0	56.2	57.4	55.7	55.4	57.0	56.0	54.4	55.2	56.0
Tr–Az	22.9	24.6	23.9	23.2	23.6	24.9	23.6	23.2	24.6	24.9	23.9
Tr–Be	22.2	16.8	21.6	20.7	21.3	21.6	23.4	19.8	19.5	21.3	20.8
Tr–Cs	68.5	68.0	66.7	67.2	68.0	68.1	68.4	67.1	67.8	66.3	67.6
Tr–En	73.5	74.0	73.7	73.2	73.0	73.2	74.2	74.0	72.9	72.2	73.4
Tr–Es	74.4	74.0	74.6	75.5	73.2	73.8	74.6	74.7	74.8	74.4	74.4
Tr–Gl	36.1	36.6	35.9	36.4	35.9	29.7	36.7	36.7	35.0	36.8	35.6
Tr–Pt	72.2	71.8	71.8	72.8	71.3	71.4	70.8	71.8	72.4	72.1	71.8
Tr–Ru	61.3	61.8	60.0	61.8	61.7	61.8	60.5	60.0	59.5	59.9	60.8
Tr–Sk	55.4	56.8	56.8	57.0	56.2	54.9	56.4	55.8	51.6	55.4	55.6

Table 15: All results from the distant languages MWE experiment (P@1).

Test	Hub language							μ
	EN	FR	HI	KO	RU	SV	UK	
EN–FR	75.1	75.3	75.2	75.8	76.3	75.5	75.4	75.5
EN–HI	20.9	23.5	21.0	21.4	23.5	21.4	23.9	22.2
EN–KO	9.2	10.4	9.1	9.8	9.8	10.1	10.0	9.8
EN–RU	41.8	42.0	41.8	41.5	42.0	41.8	42.0	41.8
EN–SV	57.0	57.5	59.0	56.6	57.8	57.6	58.4	57.7
EN–UK	26.9	27.5	26.9	26.9	28.3	27.8	26.2	27.2
FR–EN	72.5	72.0	71.6	72.7	72.9	73.4	74.0	72.7
FR–HI	18.7	16.0	14.8	17.3	19.0	17.8	17.5	17.3
FR–KO	6.9	6.7	5.8	5.5	5.8	7.5	6.0	6.3
FR–RU	39.9	38.3	40.3	40.4	40.8	40.0	39.6	39.9
FR–SV	51.8	49.3	50.5	51.1	49.4	48.2	51.8	50.3
FR–UK	28.8	27.0	27.8	28.5	28.7	27.7	26.1	27.8
HI–EN	27.8	31.4	27.9	28.6	30.4	29.3	29.3	29.3
HI–FR	25.6	23.1	25.1	23.3	26.9	25.5	24.2	24.8
HI–KO	2.1	1.7	1.3	1.6	1.6	1.4	1.8	1.6
HI–RU	13.9	14.2	14.3	13.6	14.3	13.5	14.6	14.0
HI–SV	17.3	16.8	16.3	15.9	17.0	15.9	16.6	16.6
HI–UK	10.3	10.5	9.1	9.1	9.8	9.5	9.6	9.7
KO–EN	15.1	16.6	15.2	17.0	16.6	17.7	16.4	16.4
KO–FR	11.9	10.2	10.9	10.9	12.6	13.6	10.8	11.6
KO–HI	1.8	2.4	1.2	1.6	2.0	1.8	2.0	1.9
KO–RU	7.9	6.6	6.0	5.7	6.9	6.8	7.3	6.7
KO–SV	6.8	6.6	5.9	5.9	7.2	5.6	7.2	6.5
KO–UK	3.5	3.6	3.4	3.2	3.5	3.5	3.1	3.4
RU–EN	50.2	53.2	52.2	53.4	52.5	52.6	52.1	52.3
RU–FR	51.1	49.6	50.7	51.7	51.0	50.6	50.3	50.7
RU–HI	14.6	15.0	12.0	14.6	13.3	14.8	15.3	14.2
RU–KO	5.2	4.6	4.4	3.6	4.3	4.1	5.0	4.4
RU–SV	40.7	40.9	40.1	41.0	39.8	36.7	41.3	40.1
RU–UK	55.3	56.1	55.8	56.3	55.3	55.3	54.9	55.6
SV–EN	51.2	51.1	52.3	51.9	52.0	50.7	52.7	51.7
SV–FR	47.9	45.7	46.8	48.2	47.1	46.6	47.4	47.1
SV–HI	17.2	16.3	15.0	16.0	17.7	15.9	17.0	16.4
SV–KO	4.9	4.2	4.0	3.8	5.0	4.0	5.1	4.4
SV–RU	31.5	33.2	32.4	33.0	31.8	30.2	31.8	32.0
SV–UK	22.4	23.8	23.0	23.5	24.1	21.0	21.9	22.8
UK–EN	39.5	40.8	40.3	40.7	41.4	40.2	40.2	40.4
UK–FR	43.6	42.3	44.0	43.3	43.0	43.3	40.6	42.9
UK–HI	13.8	13.8	12.8	12.8	12.7	14.4	13.0	13.3
UK–KO	2.6	2.5	2.4	2.0	2.0	2.4	2.6	2.4
UK–RU	59.4	58.9	59.7	58.7	59.1	58.4	58.6	59.0
UK–SV	35.8	35.5	35.8	36.8	35.4	32.7	35.1	35.3

Table 16: All results from the distant languages MWE experiment (P@5).

Test	Hub language							μ
	EN	FR	HI	KO	RU	SV	UK	
EN–FR	87.3	88.2	87.8	88.4	88.3	88.0	87.7	88.0
EN–HI	37.2	39.4	36.5	37.1	39.3	38.7	39.9	38.3
EN–KO	23.4	24.6	22.6	23.4	24.3	25.9	25.0	24.2
EN–RU	63.5	65.3	65.1	64.8	66.9	64.6	65.9	65.2
EN–SV	74.8	76.1	76.3	75.8	75.4	75.6	76.5	75.8
EN–UK	47.7	49.8	49.3	47.9	49.3	48.5	47.7	48.6
FR–EN	85.3	84.5	83.7	84.5	85.4	85.1	84.6	84.7
FR–HI	32.7	30.0	29.5	30.6	33.4	32.2	31.6	31.4
FR–KO	14.9	14.5	14.0	14.6	16.0	15.3	15.2	14.9
FR–RU	61.0	59.5	61.9	61.7	62.1	60.6	60.9	61.1
FR–SV	69.6	68.1	68.8	69.1	68.6	68.0	71.1	69.0
FR–UK	45.6	44.2	44.8	45.6	45.8	45.0	44.1	45.0
HI–EN	44.5	47.0	46.3	44.3	47.0	46.3	46.7	46.0
HI–FR	41.7	39.3	41.6	39.6	42.7	41.2	42.3	41.2
HI–KO	5.3	4.8	3.4	3.5	4.7	5.1	5.0	4.5
HI–RU	27.6	29.6	27.6	28.1	27.9	28.8	29.5	28.4
HI–SV	31.7	31.7	30.8	30.7	32.7	30.2	32.0	31.4
HI–UK	21.4	21.9	19.9	20.1	20.8	20.4	20.2	20.7
KO–EN	28.9	28.7	27.0	28.1	30.1	33.1	28.6	29.2
KO–FR	21.9	21.6	19.7	20.4	24.0	24.4	21.3	21.9
KO–HI	4.3	4.8	3.9	4.1	4.6	4.8	5.0	4.5
KO–RU	16.2	15.3	12.9	13.4	15.8	15.7	16.3	15.1
KO–SV	16.2	14.1	13.9	13.8	15.6	13.9	16.3	14.8
KO–UK	9.7	8.0	8.6	8.6	9.3	8.2	8.8	8.8
RU–EN	69.8	71.1	70.9	71.0	70.2	71.1	71.3	70.8
RU–FR	65.7	66.2	67.7	67.9	67.0	66.6	67.2	66.9
RU–HI	27.3	27.6	24.7	26.7	25.6	26.6	28.7	26.7
RU–KO	12.1	10.4	10.1	10.0	11.1	10.4	12.4	10.9
RU–SV	58.8	58.9	58.2	58.2	58.8	56.1	59.9	58.4
RU–UK	68.3	68.8	69.2	68.0	68.8	68.6	66.9	68.4
SV–EN	65.4	66.2	66.3	65.7	65.1	64.4	65.9	65.6
SV–FR	62.5	60.1	60.3	61.1	60.7	59.8	61.3	60.8
SV–HI	28.2	28.0	26.6	27.4	29.3	27.1	28.6	27.9
SV–KO	11.7	10.7	10.9	9.8	11.5	11.6	11.4	11.1
SV–RU	50.5	51.0	50.7	50.9	50.3	47.8	49.9	50.2
SV–UK	40.2	42.1	41.6	41.6	41.7	38.3	39.2	40.6
UK–EN	56.3	58.1	57.5	57.2	59.1	58.1	56.1	57.5
UK–FR	58.3	56.4	58.5	58.7	58.9	58.0	56.4	57.9
UK–HI	27.2	25.8	24.0	25.4	26.5	25.8	25.3	25.7
UK–KO	7.4	7.2	6.8	6.0	7.3	7.3	7.3	7.0
UK–RU	71.0	71.0	71.2	70.1	70.4	70.7	70.5	70.7
UK–SV	53.3	53.3	52.5	53.1	53.7	48.9	53.1	52.5

Table 17: All results from the distant languages MWE experiment (P@10).

Test	Hub language							μ
	EN	FR	HI	KO	RU	SV	UK	
EN–FR	90.8	91.3	90.1	91.0	91.1	91.1	90.7	90.9
EN–HI	44.0	45.9	43.3	43.1	45.0	45.2	45.6	44.6
EN–KO	31.1	31.5	28.4	30.5	31.6	33.7	32.1	31.3
EN–RU	70.1	71.7	71.0	70.7	72.4	71.1	72.3	71.3
EN–SV	80.0	81.1	80.9	80.4	80.8	80.4	81.2	80.7
EN–UK	55.3	57.5	56.5	55.2	57.4	56.4	54.6	56.1
FR–EN	87.6	87.8	86.6	87.7	88.0	87.9	88.0	87.6
FR–HI	39.1	35.3	35.5	36.5	38.6	38.1	38.5	37.4
FR–KO	20.1	18.4	18.4	19.6	20.3	19.4	19.7	19.4
FR–RU	67.1	65.9	68.1	67.5	66.8	66.8	67.4	67.1
FR–SV	74.4	73.3	74.2	74.8	73.3	73.3	75.5	74.1
FR–UK	51.7	49.7	51.3	51.8	52.0	51.2	49.9	51.1
HI–EN	50.0	52.3	53.0	50.8	52.7	51.7	52.3	51.8
HI–FR	49.0	45.5	46.8	46.8	48.3	48.1	48.9	47.6
HI–KO	7.9	7.2	5.1	5.1	6.4	6.6	7.2	6.5
HI–RU	34.5	35.3	34.5	34.7	33.6	35.3	36.3	34.9
HI–SV	38.0	37.5	36.1	37.9	38.9	36.3	38.5	37.6
HI–UK	27.3	27.6	25.8	25.4	26.2	25.9	25.5	26.3
KO–EN	34.2	34.3	32.3	35.2	37.1	38.4	35.4	35.3
KO–FR	27.0	25.9	23.7	24.6	28.5	30.1	26.4	26.6
KO–HI	6.2	6.9	5.6	6.0	6.7	6.7	6.9	6.4
KO–RU	21.2	19.3	16.4	18.2	20.4	20.9	20.8	19.6
KO–SV	20.9	18.1	17.8	17.5	21.1	18.4	20.6	19.2
KO–UK	12.9	12.1	11.5	11.3	12.6	12.0	11.7	12.0
RU–EN	74.9	75.8	75.4	75.5	75.5	76.2	75.6	75.6
RU–FR	71.8	72.5	73.0	72.2	72.7	72.7	72.6	72.5
RU–HI	33.0	32.9	30.1	32.1	31.9	32.1	34.6	32.4
RU–KO	17.2	14.6	13.2	13.5	15.9	15.0	16.7	15.2
RU–SV	64.7	64.7	63.6	64.6	64.2	62.5	64.6	64.1
RU–UK	73.3	72.8	73.1	72.0	73.1	72.9	71.7	72.7
SV–EN	69.5	70.4	71.0	70.6	70.9	69.3	70.0	70.2
SV–FR	67.0	64.2	65.0	65.3	65.5	64.2	65.7	65.3
SV–HI	33.6	32.6	32.0	30.9	33.3	31.9	33.2	32.5
SV–KO	15.7	14.7	14.0	12.9	15.7	14.9	15.6	14.8
SV–RU	57.2	56.4	56.5	56.2	56.4	53.8	56.4	56.1
SV–UK	47.5	47.9	47.7	47.7	48.5	44.8	46.4	47.2
UK–EN	61.6	63.4	62.9	62.2	63.5	62.7	61.1	62.5
UK–FR	63.5	62.4	63.9	63.4	64.3	63.5	61.9	63.3
UK–HI	32.7	32.3	28.6	30.2	31.7	31.5	30.7	31.1
UK–KO	10.6	10.2	9.5	8.7	10.1	10.4	10.2	10.0
UK–RU	74.5	73.8	74.1	73.9	74.5	74.1	73.9	74.1
UK–SV	59.1	58.8	58.8	58.7	59.3	55.2	57.8	58.2