Combining Noisy Semantic Signals with Orthographic Cues: Cognate Induction for the Indic Dialect Continuum

Niyati Bafna[1,3,4*], **Josef van Genabith**[1,2], **Cristina España-Bonet**[2], **Zdeněk Žabokrtský**[3]
[1]Saarland Informatics Campus, Saarland University, Germany
[2]DFKI GmbH, Germany
[3]Institute of Formal and Applied Linguistics, Charles University, Prague
[4]Inria, Paris, France
niyatibafna13@gmail.com,{josef.van_genabith,cristinae}@dfki.de,
zabokrtsky@ufal.mff.cuni.cz

Abstract

We present a novel method for unsupervised cognate/borrowing identification from monolingual corpora designed for low and extremely low resource scenarios, based on combining noisy semantic signals from joint bilingual spaces with orthographic cues modelling sound change. We apply our method to the North Indian dialect continuum, containing several dozens of dialects and languages spoken by more than 100 million people. Many of these languages are zero-resource and therefore natural language processing for them is non-existent. We first collect monolingual data for 26 Indic languages, 16 of which were previously zero-resource, and perform exploratory character, lexical and subword cross-lingual alignment experiments for the first time at this scale on this dialect continuum. We create bilingual evaluation lexicons against Hindi for 20 of the languages. We then apply our cognate identification method on the data, and show that our method outperforms both traditional orthography baselines as well as EM-style learnt edit distance matrices. To the best of our knowledge, this is the first work to combine traditional orthographic cues with noisy bilingual embeddings to tackle unsupervised cognate detection in a (truly) low-resource setup, showing that even noisy bilingual embeddings can act as good guides for this task. We release our multilingual dialect corpus, called HinDialect, as well as our scripts for evaluation data collection and cognate induction.[2]

1 Introduction

Hindi is listed as one of the 22 official languages of India, with the latest census showing 43.63% of Indians as having Hindi as their mother tongue.[3] However, this figure counts speakers of the languages of the whole Indic/Indo-Aryan (IA) dialect continuum, the "Hindi Belt", that stretches from Rajasthan in the West to Bihar and Jharkhand in the East, and of which modern standard Hindi is only a part.[4] This continuum, spread out over North and Central India, contains a wide variety of languages/dialects that may even be mutually incomprehensible, and form subfamilies of their own, e.g. the Rajasthani, Bihari, or Pahari subfamilies.[5]

Natural language processing (NLP) resources for these languages are sorely lacking; most of these languages, despite having millions of speakers, have little or no monolingual data, no linguistic resources such as lexicons, grammars, taggers, let alone more elaborate resources such as parallel data or pretrained embeddings.

We focus on 26 languages of the Hindi Belt written in the Devanagari script and make the following contributions: *(i)* we collect the first monolingual resources for many of these languages, and *(ii)* we develop a novel strategy for cognate lexicon induction in asymmetric truly low-resource scenarios, tackling this problem for the first time with the under-researched Indic dialect continuum. Cognate induction is an important first step towards obtaining bilingual lexicons, one of the most basic and all-purpose bilingual resources a language can have. Bilingual lexicons are especially useful in low-resource scenarios, e.g. for word-by-word translation, bilingual transfer, and as seeds for a variety of tasks; they also have applications in historical linguistics. Finally, in the case of severely under-supported languages, they are crucial for building dictionaries for speakers and language learners. In this work, we perform cognate induction for each language against Hindi, since Hindi

*This work was done at Charles University and Saarland University as a Masters' Thesis.

[2]See http://hdl.handle.net/11234/1-4839 and https://github.com/niyatibafna/north-indian-dialect-modelling, respectively.

[3]https://en.wikipedia.org/wiki/2011_Census_of_India

[4]We also see a shallower north-south dimension to the continuum, i.e. from Haryana to northern Maharashtra.

[5]See https://glottolog.org/resource/languoid/id/indo1321 for the full language tree.

Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL), pages 110 - 131
December 7-8, 2022

is the most well-studied and resource-rich of this set, and therefore the most logical language from which bilingual transfer may be attempted.

We crawl monolingual data for the continuum, forming the largest collection (in number of languages) of a dialect continuum as far as we know. This also introduces the first monolingual data for 16 zero-resource IA languages to the NLP community. Such a corpus has wide applications for work in transfer, historical linguistics, dialect continua, and building language support for these communities. We probe the resulting multilingual collection at a character, subword and lexical level, finding a general link between relatedness and genealogically and geographically proximal languages.

Secondly, we use the corpus for cognate/borrowing induction (CI) for each target language with Hindi:[6] identifying cognates from monolingual corpora containing fully inflected word forms in a completely unsupervised manner.[7] We work in an asymmetric data scarcity situation: we have abundant monolingual resources for Hindi, but only a few thousands/ten thousands of monolingual tokens for target languages. These constraints set this task apart from most of the previous literature on cognate identification (List, 2014; Fourrier et al., 2021; List, 2019; Artetxe et al., 2018); however, this setting is realistic when attempting to build resources for truly low-resource languages. We present two simple but novel strategies for cognate identification, evaluating on synthetically created test sets. We experiment with iteratively learning substitution probabilities within an edit distance paradigm, as well as combining noisy semantic signals from a subword embedding space with orthographic distance measures, reporting qualitative improvements over the baseline.

2 Related Work

Data and Resources. Languages in the continuum differ in the amount of resources available. For the highest resourced languages (this corresponds to Band 1 in Section 5) one can find raw and annotated corpora, pretrained embeddings, and evaluation resources (Kunchukuttan et al., 2020; Bojar et al., 2014; Nivre et al., 2016). For medium-resourced languages (Band 2), we have some collection efforts,[8] mostly monolingual (Ojha, 2019; Ojha et al., 2020; Goldhahn et al., 2012) but including some parallel data. Zampieri et al. (2018) presented a shared task for language identification for Awadhi, Braj, Bhojpuri, Magahi, and Hindi providing 15k sentences for each language. Mundotiya et al. (2021) collect monolingual corpora for Bhojpuri, Magahi, and Maithili, as well as POS-tagged annotated corpora and WordNets[9] aligned with the larger IndoWordNet effort (Sinha et al., 2006) Mundotiya et al. (2022) presents NER-annotated corpora and trained NER models for the same 3 languages. The least resourced languages (Band 3) lack any kind of systematic resource and are the main focus of our work.

Bi/Multilingual Lexicon Induction Much previous work has been based on *non-neural methods*. Batsuren et al. (2019) use semantic relationships from the Universal Knowledge Core (Giunchiglia et al., 2018) which is built from existing WordNets,[10] gold annotations as well as geographical-orthographic similarity measures for cognate identification. Çöltekin (2019) compares linear and neural models to predict the next edit-distance based action to perform crosslingual morphological inflection. In earlier works, Scherrer and Sagot (2014), inspired by Koehn and Knight (2002), induced cognate sets in a completely unsupervised manner using a character-based alignment algorithm, as well as co-occurrence-based context vectors. List (2012) induce cognate sets over aligned word lists of languages in a language family by iteratively learning phonological rules; this is implemented in the software LingPy (List, 2014). Hall and Klein (2010) work with unaligned word lists for languages in the same family, modelling transfer within a tree-based framework and learning edit-distance based transformation matrices for each vertical edge. Although the idea of learning edit distance matrices is quite old (Bilenko and Mooney, 2003), it has not been used in combination with modern embeddings-based methods for cognate identification as far as we know.

Recently, *neural and embeddings-based methods* have been gaining importance. Conneau et al. (2018) is one of the earliest works to link bilingual

[6] Henceforth, we use the term "cognate" as including borrowings.

[7] While we do have lexical resources for Band 1 and 2 languages including WordNets for *some* Band 1 languages (see Table 1 for bands), we simulate low-resource scenarios consistent with the truly low-resource Band 3 languages

[8] See `www.ldcil.org/resourcesTextCorp.aspx`

[9] Not publicly available yet

[10] CogNet contains only Band 1 Indic languages

lexicon induction (BLI) with bilingual embedding spaces, or the alignment of monolingual embeddings. This idea has been explored by other works that seek to adapt it to low-resource settings or relax its strong isometry assumption (Dou et al., 2018; Patra et al., 2019), sometimes using a bootstrapping strategy for embedding alignment and bilingual lexicon induction (Artetxe et al., 2018; Cao and Zhao, 2021). Fourrier et al. (2021) frame cognate induction as a machine translation problem, finding that SMT beats NMT over smaller datasets; Kanojia et al. (2019) identify cognate sets for (Band 1) Indian languages using the IndoWordNet combined with lexical similarity measures, training neural models over the resulting data.

3 Orthographic Distance for Cognate Induction

3.1 Baseline Approach

A straightforward approach for CI involves using orthographic distance as a stand-in for phonological distance, motivated by the fact that Devanagari is orthographically shallow, that is, spellings closely represent associated pronunciations. We consider source words from Hindi; the best cognate candidate in the other language is chosen by minimizing orthographic distance. We use two distances: normalized edit distance (**NED**), that is, the edit distance normalized by the maximum of the 2 word lengths, thus scaling to 0-1; and Jaro-Winkler (**JW**) distance (Winkler, 1990), which weights differences higher in the beginnings of strings.

For all approaches, we use a minimum source frequency of 5, maximum lexicon size of 5000, and we collect 5 best candidates per source word; this ensures identical recall over all approaches given a fixed source language corpus and test lexicon.

3.2 Expectation-Maximisation Approach

A limiting theoretical deficiency in the baseline approach is that it treats substitutions of any two characters equally (similarly for insertions and deletions). By contrast, the expectation-maximisation (EM) approach optimises substitution probabilities iteratively while simultaneously learning cognate pairs, given two lexicons, in an expectation-maximization style algorithm. We call it **EMT**, EM for “Transform probabilities".

Setup. Given two word lists (that may overlap) WL_s and WL_t, let the set of all characters of the source and target side be χ_s and χ_t respectively. We use a scoring function $S(c_i, c_j)$, that contains a “score” for replacing any character $c_i \in \chi_s$ with $c_j \in \chi_t$;[11] for a given character in a source word, S is modelled as a transformation probability distribution over χ_t. S is initialized by giving high probability (in practice, 0.5) to self-transforms and distributing the remaining probability mass equally over other characters.

Given that $C(a, b)$ is the number of times we have seen $a \to b$, and $T(a)$ is the total number of times we have seen a on the source side, our score is the conditional probability:

$$S(c_i, c_j) = \frac{C(c_i, c_j)}{T(c_i)} \quad (1)$$

We maintain a list of cognates found over all EM loops, so that we only update model parameters once per cognate pair. Note that a word may appear in many different cognate pairs in this setup.

The EMT Algorithm is composed of two steps.

1) Expectation step. Given a candidate source and target pair (s, t), we can find $Ops(s, t)$, which is the *minimal list* of the operations we need to perform to get from s to t. Each member in Ops is of the type (c_i, c_j). In addition to “insert”/“delete”/“replace” operations, we also use a “retain” operation, for characters that remain the same; we also want to estimate $S(a, a)\ \forall\ a$.

The score for the pair (s, t) is computed as

$$\zeta(s, t) = -\sum_{(a,b)\in Ops} log_{10}(S(a, b)), \quad (2)$$

where the lower the ζ the more probable a pair is a cognate. For a given s, we can then always find the word that is the most probable cognate as $t = min_{t_i \neq s}(\zeta(s, t_i))$.

Note that in the training phase, we disallow $s = t$, to mitigate exploding self-transform probabilities. Finally, we choose the best K of all cognate pairs i.e. those with the highest confidence, equivalent to the lowest ζ values.

2) Maximisation step. We update the model parameters based on the newly identified cognates in the previous step. This is done by increasing the counts of all observed edit distance operations:

$$C(a, b) := C(a, b) + 1 \quad \forall (a, b) \in Ops(s, t)$$

[11] We model insertion and deletion as special cases of replacement, by introducing a null character.

$$T(a) := T(a) + 1 \qquad \forall (a,b) \in Ops(s,t)$$

Inference is performed by choosing the K best target candidates that minimise $\zeta(s,t)$ as described above, now allowing self-matches.

4 Semantic Similarity for Cognate Induction

Orthographic matching, even with tailored and learnt substitution matrices for a given pair of languages, may be inherently inadequate, as it pays no heed to the shared semantics of cognates. We use bilingual subword embeddings (BE) to address this problem in the following way: we use the semantic space to narrow down possible candidates, and then apply orthographic matching in order to select the top K candidates. This is a two-stage approach that relies mainly on two separate metrics: first, the quality of semantic similarity judgments provided by a semantic embedding space, and second, orthographic similarity judgments provided by the distance/similarity metric we choose to use.

SEM_JW: BE+JW In this approach, we retrieve K nearest neighbours of each source word. These candidates are scored by an interpolation of semantic similarity and orthographic distance, with equal weighting. We use cosine similarity for the former, and JW for the latter. All words that are not within the K nearest neighbours (50 in our experiments) are discarded from consideration. The idea is to mitigate the effect of chance orthographic similarities.

For candidates, if $E(s)$ is the embedding vector for string s, we minimize:

$$D(a,b) = 1 - scos(E(a), E(b)) \cdot J(a,b), \quad (3)$$

where $scos(v_1, v_2)$ captures the cosine similarity (scaled to [0, 1]) between vectors v_1 and v_2, and $J(a,b)$ is the **JW** similarity.

SEM_EMT: BE+EMT We seek to combine the benefits of iteratively learning transformation probabilities with those of semantic spaces. This approach is almost identical to that in Section 3.2, except for the fact that only $K = 50$ nearest neighbours of a source word in the semantic space are used as its potential cognate candidates, both during training and inference.

5 Data Collection

We apply the methods described above to the Indic dialect continuum. Since these languages cover a range of resource situations, we divide them into three categories, Band 1, 2 and 3, based on amount of resources, with Band 1 containing the best resourced languages, and Band 3 containing (previously) zero-resource languages. See Table 1 for a description of the languages under consideration.

5.1 Monolingual Corpora Crawl

Digital presence of Band 3 languages is low to non-existent; automatic crawling for content faces the primary problems of scarcity, script handling, and automatic language identification between closely related variants.

Kavita Kosh,[12] translating roughly to "poetry collection", is an online collection of folksongs and poems in 31 languages from the IA continuum. Content is manually curated by the organization; the poetry consists of works by early contemporary writers, mostly from the late twentieth century. All content is in Devanagari (transliterated in case of e.g. Bengali content). The website categorizes pieces by type, language, author/theme, and possibly additional labels such as anthology. We collect data for a total of 31 languages, of which we have folksong data for 26 languages, and poetry data for 18 languages.[13,14] We leave out 5 languages for cognate induction: Bangla, Gujarati, Punjabi (written primarily in a different script), Sanskrit and Pali (extinct languages). The data is cleaned at a character-level, we filter out words with any character not within a specified UTF-8 code-point range and tokenization is performed by white-space splitting. See total counts in tokens in Table 1. Poem and token counts are reported in Appendix A.[15]

5.2 Evaluation Data for Cognate Induction

Band 3 languages lack standardized gold bilingual lexicons that may be used for supervision. After a survey of possible digital resources for this purpose (see Appendix B for a listing), we choose to use Languages Home, an online language learning website,[16] containing translations of 80–90 artificially simple English sentences (e.g. "He ate an apple",

[12] http://kavitakosh.org/kk/

[13] We also include Korku as an outlier datapoint; it is *not* an Indic language and therefore lacks the genealogical similarities of the others.

[14] We preserve the distinction made by the website between Khadi Boli and Hindi; the former is the closest to what we consider modern Hindi.

[15] We have been authorized by the organization to make the folksongs data available but not the poetry. However, our crawler is publicly available to use.

[16] https://www.languageshome.com

Language	Primary Regions	Language (Sub-)Family	Data (Tok.)	Collected (Tok.)	# native speakers
BAND 1					
Hindi	Uttar Pradesh*, Bihar*, Rajasthan*, 13 others	IA Central, Western Hindi	1.86B[1]	7127997	250M†
Marathi	Maharastra*, Goa*	IA Southern, Marathic	551M[1]	3327	73M
Nepali	Nepal*, West Bengal*	IA Northern, Eastern Pahari	14M[2]	692657	16M
Sindhi	Sindh*, Pakistan, Rajasthan, Gujarat	IA Northwestern, Sindhi-Lahnda	61M[5]	51458	25M
BAND 2					
Bhojpuri	Bihar, Jharkhand*	IA, Bihari	259K[3]	197639	40M
Awadhi	Bihar	IA, Bihari	123K[3]	500079	38M
Magahi	Bihar, Jharkand*	IA, Bihari	234K[3]	84754	40M
Maithili	Bihar*, Jharkhand*	IA, Bihari	300K[4]	218339	14M
Brajbhasha	Uttar Pradesh	IA Central, Western Hindi	249K[3]	160039	1M
BAND 3					
Rajasthani	Rajasthan	IA Central, Gujarati-Rajasthani	-	187724	50M
Hariyanvi	Haryana, Rajasthan	IA Central, Western Hindi	-	233003	13M
Bhili	Rajasthan, Gujarati, Madhya Pradesh	IA Central, Bhil	-	27326	3M
Korku	Madhya Pradesh, Maharashtra	Austro-Asiatic, North Munda	-	15509	0.7M
Baiga	Chattisgarh	IA Central, Chattisgarhi	-	13848	UNK
Nimaadi	Rajasthan, Madhya Pradesh	IA Central, Bhil	-	14056	2M
Malwi	Rajasthan, Madhya Pradesh	IA Central, Bhil	-	9626	5M
Bhadavari	Jammu Kashmir	IA Northern, Western Pahari	-	990	0.1M
Himachali	Himachal Pradesh	IA Northern, Himachali	-	466	2M
Garwali	Uttarakhand	IA Northern, Central Pahari	-	92668	6M
Kumaoni	Uttarakhand	IA Northern, Central Pahari	-	1028	2M
Kannauji	Uttar Pradesh	IA Central, Western Hindi	-	327	9.5M
Bundeli	Madhya Pradesh, Uttar Pradesh	IA Central, Western Hindi	-	26928	5.6M
Chattisgarhi	Chattisgarh*	IA Central, Eastern Hindi	-	83226	18M
Bajjika	Bihar	IA, Bihari	-	7414	12M
Angika	Bihar, Jharkhand*	IA, Bihari	-	1265146	15M
Khadi Boli	Delhi	IA Central, Western Hindi	-	4507	UNK

Table 1: Language bands. Note that Band 1 languages may have much more data available from other sources such as Wikipedia; for Band 2 languages, we may have other sources with the same order of magnitude of data. "Primary Regions" only mentions places in the Indian subcontinent; * indicates official status. Corpora from which data counts are taken: [1](Kakwani et al., 2020), [2](Yadava et al., 2008), [3](Zampieri et al., 2018), [4](Goldhahn et al., 2012) [5](Conneau et al., 2020). Speaker counts taken from (latest) 2011 census if available. †: probably inflated

"He will come") into 76 Indian languages (including some Dravidian languages and IA languages for which we do not have data). This resource has the best coverage as well as consistency over Band 3 languages. Of these, 20 languages are of our interest, including 12 Band 3 languages. This data is considerably noisy, with problems including the fact that it is written in "casual" Roman transliteration, inconsistent parenthetic explanations, and code-switching.

We develop a pipeline to extract the aligned lexicons. The pipeline consists of cleaning, transliteration of the Indic side into Devanagari with indictrans (Bhat et al., 2015), parallelizing with Hindi instead of English,[17] and finally extracting word-alignments over the given Hindi-parallel data with FAST-ALIGN (Dyer et al., 2013).

The resulting lexicons have an average size of 153.6 elements, a minimum size of 118, and a maximum of 177. We manually evaluate the Hindi-Marathi lexicon, finding that 73.5% of 130 source words contain at least one correct target.[18] Despite clear problems of noise, and acknowledging that these lexicons should be post-edited by native speakers, this is the best possible evaluation data that we can use, given its coverage and uniform format; however, we consider it as a relative rather than absolute indicator of performance.

6 Experiments and Results

6.1 Probing the Monolingual Corpora

We seek to capture a high-level picture of the data on the character, subword, and lexical level, comparing observations with language-specific characteristics from prior knowledge as well as with expected cross-lingual relationships. For this, we perform 3 types of experiments.

Character-level. We inspect the symmetric KL-Divergence[19] over characters as well as char-gram distributions of the languages. For the latter, the final metric is simply the average over divergence values for each char-gram length. Since IA languages are orthographically shallow, inspecting such distributions of a language may give us a fairly good idea of the general usage of consonants and vowels in the language.

Lexical Overlap. If L_i and L_j are the filtered lexicons of two languages i and j, we calculate

$$O_{ij} = \frac{|L_i \cap L_j|}{min(|L_i|, |L_j|)} \tag{4}$$

for all pairs. We apply a corpus-dependent frequency threshold to the data: we discard all words in a corpus with size N_L that occur with a frequency less than $T(N_L) = log_{100}(N_L) - 1$. The exponent 100 and the constant -1 were chosen such that the threshold does not grow too quickly, and that datasets with less than 1000 tokens are fully retained.

Subword-level. We calculate pairwise subword-level overlap measures, captured by character grams of length 2–4,[20] thinking of subwords as approximating morphemic units of the language. Let's define L_{ic} as the inventory/lexicon of c-length char-grams for language i, then the c-char-gram overlap O_{ijc} for languages i and j is calculated identically to lexical overlap in Eqn 4.

We would like to weight O_{ijc} according to c, capturing the idea that it is more of a similarity signal for two languages to share c-char-grams for a higher c. For this purpose, we calculate the "universe of possibilities" for each c; i.e. the total number U_c of unique c-char-grams that occur in the entire corpus. Since we want normalizing weights that are inversely related to the probability of an accidentally shared c-char-gram, we calculate subword similarity as follows:

$$O_{ij} = \sum_c \left(O_{ijc} \cdot \frac{U_c}{\sum_c U_c} \right) \tag{5}$$

Finally, we also calculate pairwise symmetric KL-Divergence over subword distributions.

Results. Figure 1 is generally representative of our results across character, subword and lexical results, both overlap-based and information-theoretic (see Appendix A for related heatmaps). The following general observations emerge from all the above experiments. The Purvanchal and eastern languages from Kannuaji to Angika (represented in the bottom right), show the highest similarity/overlap within themselves over all calculated measures. This is expected and confirms that the

[17]Word alignment of Indic languages with Hindi sentences as compared to English sentences is likelier to be accurate.

[18]Note that a word equivalent used here may not be a cognate even if a cognate does exist in the language.

[19]Specifically, for probability distributions P and Q, we calculate the symmetric quantity $D_{KL}(P||Q) + D_{KL}(Q||P)$

[20]Different ranges yield the same trend.

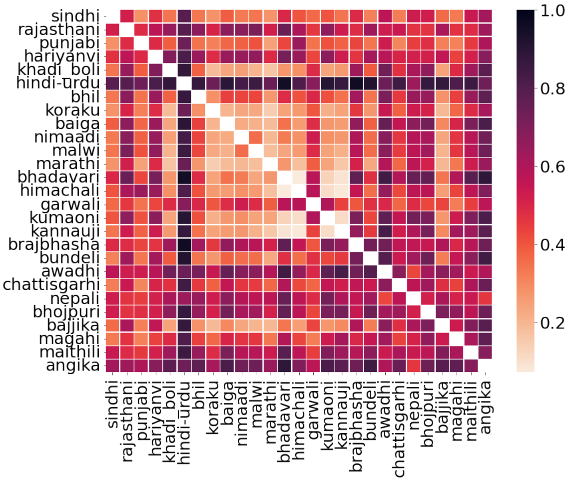


Figure 1: Overlap-based similarity over i-char-grams.

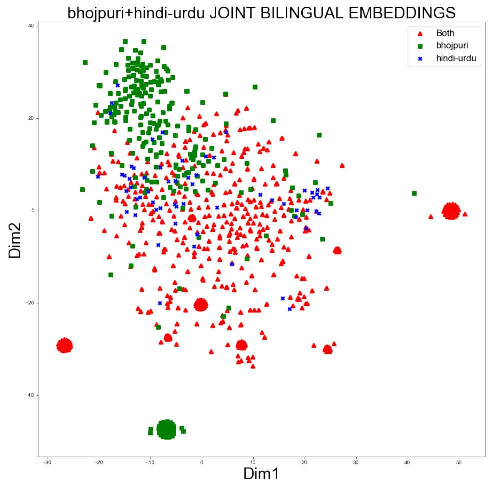


Figure 2: t-SNE visualization (Van der Maaten and Hinton, 2008). Bhojpuri words cluster together.

corpus represents the close genealogical and cultural ties between these languages.

We see that Hindi has high lexical/subword-level similarities with almost every language. This could be the result of the widespread use of Hindi, or its large dataset, including noise even after filtering. We also notice that some languages have consistently low lexical similarities with others. In the case of Korku, this is expected, given that Korku is a genealogical outlier. In other cases, such as with Malwi and Himachali, this is probably because the collected dataset is too small to be representative of the vocabulary of these languages. In general, and as expected, the eastern cluster as well as the western cluster of languages show close relationships with each other.

6.2 Bilingual Embeddings

We use FASTTEXT (Bojanowski et al., 2017) for training bilingual embeddings in a simple **joint** manner, with minimum corpus frequency according to the corpus-dependent threshold $T(N_L)$, described in Section 6.1; we hope to leverage its usage of subword information, given that that we are dealing with data-scarce morphologically rich languages.

Visualizations reveal that low-resource target language words often cluster around each other, whereas Hindi words and words belonging to both languages are more meaningfully distributed. (See Figure 2, Appendix C for other language plots.) A possible diagnosis is an effect pointed out by Gong et al. (2018) who show that low-frequency words tend to cluster together regardless of their semantics. This, along with the fact that we are unfairly applying the same minimum frequency threshold (better suited for the high-resource anchor) for both languages by mixing the data, may explain the poor quality of the target language embeddings. In order to mitigate the problem, we **upsample** the target language data to bring it to the same order of magnitude as the Hindi data.

Results We use the Nepali WordNet to extract a Hindi–Nepali bilingual lexicon, and we calculated Recall@50 (given 50 nearest neighbours). We also use basic visualizations and a crosslingual integration metric *cl_integ*, which measures the fraction of nearest neighbours per word that belong to the other language, to compare the two sets of embeddings, on average. That is, if $\nu_E(w, K)$ is the set of K nearest neighbours of w in the embedding space E and $\psi_n(L)$ is a sample of n words from a language with lexicon L, then

$$cl\_integ_{12} = \frac{1}{n \cdot K} \left(\sum_{\substack{w \in \\ \psi_n(L_1)}} \sum_{\substack{w' \in \\ \nu_E(w,K)}} I(w' \in L_2) \right)$$

We report scores as a percentage, with $n = 500$ and $K = 10$.

The UPSAMPLE Nepali model has better Recall@50 for the Hindi–Nepali gold lexicon (33% vs. 29%).[21] Representing *cl_integ* scores as a pair of integration values in either direction, i.e. (target-Hindi, Hindi-target), we find that the UPSAMPLE

[21] We also evaluated differently sized subsets of Nepali data for over the WordNet lexicon, which yielded consistent results; see Appendix C for details and more visualizations.

	NED	JW	EMT	SEM_JW	SEM_EMT	Gold
1	कहा	कहा	कहा	कहा	कहा	कहलाह
2	कहना	कहना	कहना	कहात	क	कहल
3	कहाँ	कहाँ	कहाँ	कहाई	कहौं	-
4	एकहा	कहमा	एकहा	कहनाम	लजाते	-
5	कहमा	कहां	-	कह	पूछा	-

Figure 3: Hindi source word: /kəhaː/ (said). SEM_JW approach performs the best, resulting in Bhojpuri equivalents (except the third prediction) and inflections. SEM_EMT also results in semantically correct outputs (for all but the fourth prediction). The NED/JW approaches produce orthographically close words that are semantically unrelated, e.g. /kəhãː/ (where).

models show scores of (43%, 27%), and the JOINT models show (91%, 14%), averaged over all languages. We see that the UPSAMPLE models show less skew by direction, and higher scores for the latter direction (which is what we use).

Finally, visualizations for different languages (see Appendix C.1 for an example) show the target language words to be better distributed in the UPSAMPLE approach, with more meaningful collocations. All of these are good indications that upsampling did indeed improve the quality of the bilingual embedding space. We use these for the subsequent approaches.

6.3 Cognate Induction

Our main results are presented in Table 2. There is no clear quantitative winner; SEM_JW performs slightly better than the other approaches on average. Cognate identification methods usually work at a much higher accuracy (Beinborn et al., 2013; Fourrier et al., 2021), 70–90%. The low accuracies that we record are due to a number of factors: a much lower resource range, lack of aligned word lists, lemmatizers, or supervision and evaluation, as well as noise in the evaluation data. While most literature assumes lemmatized word lists as input for this task, we do not have lemmatizers for these languages and work with fully inflected word forms; this is a further challenge for our CI strategies.

Qualitatively, we observe significant differences across models. See Figure 3 for example outputs.

NED/JW: The NED/JW approaches are often able to capture the correct answer for longer words, because the closest candidate in edit distance is likely to be in the ballpark for closely related languages. However, we also often get outputs (especially the second or third prediction) that are entirely off, as is expected from this naive idea.

EMT: Taking a look into the substitution distributions learnt by EMT, we see that it learns some expected relationships e.g. the relationship between /i/ and /iː/, shifts between other vowels, or the fact that some rarely used characters are likely to be deleted. However, the approach is not able to produce good final outputs. We attribute this to a bad seed; this approach basically depends on the seed obtained from simple NED to get started, and if it meanders down a mistaken path, that error tends to magnify itself due to the iterative nature of the algorithm, sometimes resulting in even worse final outputs than simple NED/JW.

SEM_*: The SEM_* approaches are intended to address the fundamental inadequacy in the above approaches: the fact that they do not exploit the shared semantics of cognates. SEM_JW is accordingly better at producing outputs that are semantically related, besides the required cognates. Top predictions tend to be similar to those of NED/JW, but SEM_JW produces a better collection of outputs, from the perspective of bilingual lexicons, especially since it is less biased against a higher number of substitutions. However, for many words, the method produces rather Hindi-like outputs, probably as a result of the persisting problem of language-wise clustering in the spaces.[22] SEM_EMT still suffers from the same problems as before; we see therefore that a stronger orthographic distance metric such as JW is better able to spot the cognate from semantically related words.

7 Discussion and Conclusion

We analyse the performance of the approaches with respect to the different facets of cognacy.

Variant inflectional endings: Learning the correspondences between inflections in a dialect pair is a crucial task when it comes to cognate identification for fully inflected word forms. In terms of producing the right answer, we see an intuitive split between common and rare words when it comes to other approaches. For common words, SEM_JW

[22] This problem may be mitigated with a higher target frequency threshold.

	Total	Found	NED	JWM	EMT	SEM_JW	SEM_EMT
Kumaoni	138.0	118.0	**5.1**	4.2	**5.1**	**5.1**	4.2
Marathi	138.0	116.0	**7.8**	5.2	4.3	1.7	3.4
Bajjika	149.0	123.0	13.8	**15.4**	13.8	14.6	11.4
Malwi	153.0	125.0	**24.8**	22.4	20.0	20.0	15.2
Koraku	140.0	116.0	**1.7**	0.9	**1.7**	**1.7**	0.9
Bundeli	139.0	117.0	26.5	25.6	25.6	**30.8**	26.5
Bhil	156.0	128.0	19.5	**21.1**	17.2	18.8	18.0
Sindhi	134.0	114.0	10.5	**13.2**	7.9	10.5	9.6
Magahi	159.0	129.0	17.8	**20.9**	18.6	**20.9**	17.1
Chattisgarhi	136.0	115.0	25.2	26.1	24.3	**28.7**	26.1
Garwali	143.0	120.0	**15.8**	**15.8**	15.0	**15.8**	14.2
Brajbhasha	155.0	127.0	33.9	**34.6**	32.3	33.9	32.3
Rajasthani	144.0	120.0	30.8	29.2	27.5	**31.7**	30.0
Bhojpuri	139.0	115.0	31.3	28.7	**32.2**	30.4	29.6
Maithili	140.0	117.0	17.9	17.1	16.2	18.8	**20.5**
Hariyanvi	153.0	126.0	38.1	41.3	37.3	**43.7**	42.9
Awadhi	148.0	123.0	**28.5**	26.8	22.0	26.0	25.2
Nepali	105.0	95.0	**12.6**	**12.6**	9.5	9.5	7.4
Angika	141.0	116.0	21.6	20.7	21.6	**22.4**	21.6
Average	142.6	118.9	20.1	20.2	18.5	**20.3**	18.7

Table 2: Results for CI, precision (%) over bilingual lexicons presented in Section 5.2. A precision point is calculated per source word such that any predicted target exists in the evaluation target set.

is likely to perform better than the other approaches because the word is well embedded and the correct word form is likely to be nearby in the semantic space, and subsequently selected by JW. In these cases, especially for short words, NED/JW are likely to be derailed by irrelevant words.

Correct semantics: We would like to have semantically sensible outputs even if the predicted words are not cognates. Naturally, this is performed best by the SEM_* approaches, although the NED/JW approaches do better than expected.

Sound changes: Sound change is one of the fundamental phenomena of cognacy, and can be understood in the case of borrowing in the sense of changed pronunciations. Unfortunately, we do not have the theoretical data of attested sound changes across these dialects in order to be best able to check which approach performs best in this respect.

The SEM_JW produces overall the most respectable outputs, although this is more true for common words. The main inadequacy of all these approaches is their inability to capture language-pair specific correspondences. An extension of this work could focus on refining something akin to the SEM_EMT, which has the most theoretical potential in this direction. Improvements could include searching the hyperparameter space for better priors. An investigation into better bi/multilingual spaces is crucial to generalize good performance over rare words; future work can look into using orthographic similarities explicitly while training the space itself, as well as the utility of zero-shot multilingual contextual embeddings for this task.

We have presented a new approach to unsupervised cognate identification from monolingual corpora under conditions of asymmetric data scarcity. We collected monolingual data for 26 Indian languages of the Indic dialect continuum, 16 of which previously zero-resource, as well as synthetic evaluation data. Our experiments show the benefits of combining weak semantic signals from static bilingual embeddings with orthographic cues.

8 Acknowledgements

This work has been using data, tools and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2018101). The first author was supported by the Erasmus+ Programme, European Masters Program in Language and Communication Technologies, EU grant no. 2019-1508. The second and third author were supported by the German Federal Ministry of Education and Research (BMBF) under funding code 01IW20010 (CORA4NLP) and the German Research Foundation (Deutsche Forschungsgemeinschaft) under grant SFB 1102: Information Density and Linguistic Encoding.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. 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, Melbourne, Australia. Association for Computational Linguistics.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2019. CogNet: A Large-Scale Cognate Database. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3136–3145.

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2013. Cognate Production using Character-Based Machine Translation. In *Proceedings of the sixth international joint conference on natural language processing*, pages 883–891.

Irshad Ahmad Bhat, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. 2015. IIIT-H System Submission for FIRE2014 Shared Task on Transliterated Search. In *Proceedings of the Forum for Information Retrieval Evaluation*, FIRE '14, pages 48–53, New York, NY, USA. ACM.

Mikhail Bilenko and Raymond J Mooney. 2003. Employing Trainable String Similarity Metrics for Information Integration. In *IIWeb*, pages 67–72.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the association for computational linguistics*, 5:135–146.

Ondřej Bojar, Vojtěch Diatka, Pavel Rychlý, Pavel Straňák, Vít Suchomel, Aleš Tamchyna, and Daniel Zeman. 2014. HindMonoCorp 0.5. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Hailong Cao and Tiejun Zhao. 2021. Word Embedding Transformation for Robust Unsupervised Bilingual Lexicon Induction. *arXiv preprint arXiv:2105.12297*.

Çağrı Çöltekin. 2019. Cross-lingual Morphological Inflection with Explicit Alignment. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 71–79.

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

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word Translation without Parallel Data. In *Proceedings of the 6th International Conference on Learning Representations*.

Chakrabarti Debasri, Narayan Dipak Kumar, Pandey Prabhakar, and Bhattacharyya Pushpak. 2002. Experiences in building the Indo-Wordnet: A Wordnet for Hindi. In *Proceedings of the First Global WordNet Conference*.

Zi-Yi Dou, Zhi-Hao Zhou, and Shujian Huang. 2018. Unsupervised Bilingual Lexicon Induction via Latent Variable Models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 621–626, Brussels, Belgium. Association for Computational Linguistics.

Pankaj Dwivedi and Somdev Kar. 2016. Sociolinguistics and Phonology of Kanauji. In *International Conference on Hindi Studies*.

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.

Clémentine Fourrier, Rachel Bawden, and Benoît Sagot. 2021. Can Cognate Prediction Be Modelled as a Low-Resource Machine Translation Task? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Bangkok, Thailand.

Fausto Giunchiglia, Khuyagbaatar Batsuren, and Abed Alhakim Freihat. 2018. One World–Seven Thousand Languages. In *Proceedings 19th international conference on computational linguistics and intelligent text processing, CiCling2018*, pages 18–24.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765.

Chengyue Gong, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2018. FRAGE: Frequency-agnostic Word Representation. *Advances in neural information processing systems*, 31.

David Hall and Dan Klein. 2010. Finding Cognate Groups using Phylogenies. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1030–1039. Citeseer.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Diptesh Kanojia, Kevin Patel, Malhar Kulkarni, Pushpak Bhattacharyya, and Gholemreza Haffari. 2019. Utilizing Wordnets for Cognate Detection among Indian Languages. In *Proceedings of the 10th Global Wordnet Conference*, pages 404–412.

Philipp Koehn and Kevin Knight. 2002. Learning a Translation Lexicon from Monolingual Corpora. In *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition*, pages 9–16.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Avik Bhattacharyya, Mitesh M Khapra, Pratyush Kumar, et al. 2020. AI4Bharat-IndicNLP Corpus: Monolingual corpora and Word Embeddings for Indic languages. *arXiv preprint arXiv:2005.00085*.

Johann-Mattis List. 2012. LexStat: Automatic Detection of Cognates in Multilingual Wordlists. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 117–125.

Johann-Mattis List. 2019. Automatic Inference of Sound Correspondence Patterns across Multiple Languages. *Computational Linguistics*, 45(1):137–161.

Mattis List. 2014. Sequence Comparison in Historical Linguistics. In *Sequence Comparison in Historical Linguistics*. Düsseldorf university press.

Rajesh Kumar Mundotiya, Shantanu Kumar, Ajeet Kumar, Umesh Chandra Chaudhary, Supriya Chauhan, Swasti Mishra, Praveen Gatla, and Anil Kumar Singh. 2022. Development of a Dataset and a Deep Learning Baseline Named Entity Recognizer for Three Low Resource Languages: Bhojpuri, Maithili and Magahi. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*

Rajesh Kumar Mundotiya, Manish Kumar Singh, Rahul Kapur, Swasti Mishra, and Anil Kumar Singh. 2021. Linguistic Resources for Bhojpuri, Magahi, and Maithili: Statistics about Them, Their Similarity Estimates, and Baselines for Three Applications. *Transactions on Asian and Low-Resource Language Information Processing*, 20(6):1–37.

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'16)*, pages 1659–1666.

Atul Kr Ojha. 2019. English-Bhojpuri SMT System: Insights from the Karaka Model. *arXiv preprint arXiv:1905.02239*.

Atul Kr. Ojha, Valentin Malykh, Alina Karakanta, and Chao-Hong Liu. 2020. Findings of the LoResMT 2020 Shared Task on Zero-Shot for Low-Resource languages. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 33–37, Suzhou, China. Association for Computational Linguistics.

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 *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 184–193.

Yves Scherrer and Benoît Sagot. 2014. A Language-independent and Fully Unsupervised Approach to Lexicon Induction and Part-of-Speech Tagging for Closely Related Languages. In *Language Resources and Evaluation Conference*.

Manish Sinha, Mahesh Reddy, and Pushpak Bhattacharyya. 2006. An Approach Towards Construction and Application of Multilingual Indo-Wordnet. In *3rd Global Wordnet Conference (GWC 06), Jeju Island, Korea*. Citeseer.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

William E Winkler. 1990. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. *Proceedings of the Section on Survey Research Methods (American Statistical Association)*.

Yogendra P Yadava, Andrew Hardie, Ram Raj Lohani, Bhim N Regmi, Srishtee Gurung, Amar Gurung, Tony McEnery, Jens Allwood, and Pat Hall. 2008. Construction and Annotation of a Corpus of Contemporary Nepali. *Corpora*, 3(2):213–225.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shon, James Glass, Yves Scherrer, Tanja Samardzic, Nikola Ljubešić, Jörg Tiedemann, et al. 2018. Language identification and Morphosyntactic Tagging: The second VarDial evaluation campaign. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 1–17.

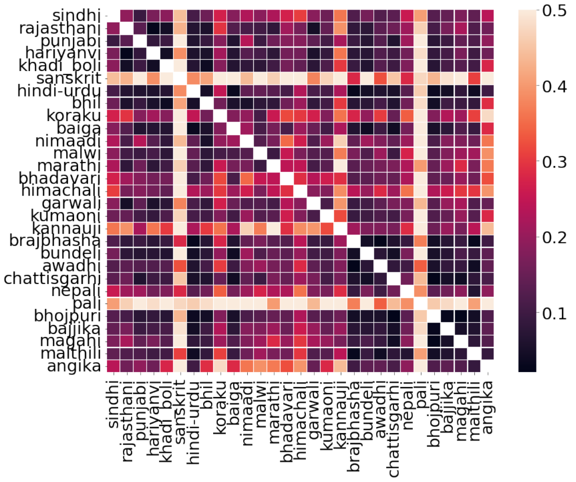


Figure 4: Character-level symmetric KL-Divergence for all languages

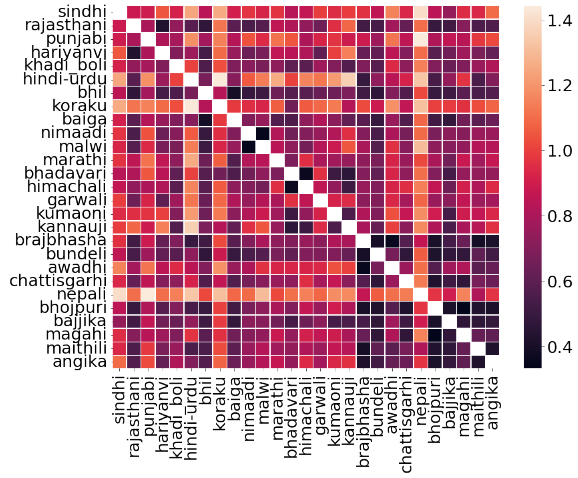


Figure 5: Pairwise KL-Divergence over distributions of i-char-grams. Lower is better.

A Data Collection and Probing

We record counts of tokens from the folksongs and poetry in Table 3.

A.1 Character-level probes

We inspect a table of character distributions over the language data after it has been cleaned. As expected, the commonest and most widely used consonants and vowels in the IA family form the bulk of the distributions of most languages, e.g. /t/, /ð/, /a/, /e/. We see some conspicuously low numbers, e.g. /ʃ/, /v/, and /ɳ/, fairly common consonants in the rest of the languages, seem to be very little used (in this corpus) from Kannauji. This is in part corroborated by Dwivedi and Kar (2016), who say that the first two are not native to Kannuaji but borrowed from Hindi.

We also see spikes in more endemic consonants as expected, for example /ɭ/ only shows reasonable percentages in Marathi and Nimaadi. Finally, the "avagraha" symbol /s/, used in Sanskrit to denote the deletion of the inherent vowel of the previous consonant, has only been inherited into the scripts of certain languages like Nepali and Magahi; in Hindi, it is sometimes used to denote the elongation of the previous vowel especially in lyrical texts. See Figure 4 for a heatmap over pairwise symmetric KL-divergence for character distributions.

A.2 Lexical measures

See Figure 6 for a depiction of pairwise lexical overlap. We also take a "close-up" look at sections of the pairwise results for language clusters that we expect to have closer relationships within the cluster. See Figures 7a,7b,7c. There are 3 such geographically motivated bands that we are interested in.

Firstly, we observe the "north" band, including Sindhi, Haryanvi, Punjabi, and the Pahari languages. Then we have the "north-central" band, which follows the heartland of the Gangetic plains, from Rajasthan (Rajasthani) across Delhi (Khadi Boli), Uttar Pradesh (Awadhi, Kannauji), Chattisgarh (Chattisgarhi), and Bihar (Bhojpuri, Magahi, Angika). Finally, we have the "central" band across southern Rajasthan (Bhili), Madhya Pradesh (Nimaadi, Malwi) and Maharashtra (Marathi).

We see that the "north-central" band indeed has the highest inter-similarities with some pairs (even excluding Hindi) showing similarities at around 70% (Bundeli-Angika, Kannauji-Awadhi). The "north" band follows; we see that Haryanvi and Nepali generally have high overlap with surrounding languages. Finally, the "central" band shows Rajasthani as having high lexical similarity with languages spoken in nearby regions, e.g. Bhili and Nimaadi; this makes sense, since Rajasthani is a catch-all for many related languages with high influence over nearby languages. Baiga shows generally low similarities except with Chattisgarhi, of which it is supposed to be a variant.[23]

Also see a dendrogram induced from lexical similarity measures in Figure 8. We see that some languages expected to be similar are grouped in the same subtrees e.g. Haryanvi and Rajasthani, {Awadhi, Angika, Bhojpuri}, as well as {Nimaadi,

[23] https://glottolog.org/resource/languoid/id/baig1238

Language	Band	Folksongs	Poetry	Folksongs tokens	Poetry tokens	Total Pieces	Total tokens
Rajasthani	3	67	1790	7404	180320	1857	187724
Gujarati	1	14	624	1795	73363	638	75158
Himachali	3	3	0	466	0	3	466
Hindi-Urdu	1	1	54408	100	7127897	54409	7127997
Magahi	2	340	376	37587	47167	716	84754
Awadhi	2	47	1333	4942	495137	1380	500079
Punjabi	1	754	0	69595	0	754	69595
Koraku	3	177	0	15509	0	177	15509
Baiga	3	35	0	13848	0	35	13848
Nimaadi	3	157	0	14056	0	157	14056
Khadi Boli	3	42	0	4507	0	42	4507
Bhojpuri	2	131	1275	20350	177289	1406	197639
Garwali	3	128	449	33380	59288	577	92668
Chattisgarhi	3	92	378	33504	49722	470	83226
Brajbhasha	2	83	1441	8883	151156	1524	160039
Bhil	3	155	0	27326	0	155	27326
Sanskrit	3	2	248	184	95450	250	95634
Angika	3	96	6773	21419	1243727	6869	1265146
Hariyanvi	3	554	930	49122	183881	1484	233003
Kannauji	3	6	0	327	0	6	327
Bundeli	3	326	0	26928	0	326	26928
Bangla	1	12	0	838	0	12	838
Malwi	3	129	0	9626	0	129	9626
Marathi	1	5	30	1412	1915	35	3327
Kumaoni	3	9	0	1028	0	9	1028
Bhadavari	3	8	0	990	0	8	990
Nepali	1	0	4753	0	692657	4753	692657
Maithili	2	0	1552	0	218339	1552	218339
Pali	3	0	27	0	5859	27	5859
Bajjika	3	0	71	0	7414	71	7414
Sindhi	1	0	500	0	51458	500	51458

Table 3: Showing crawled corpus counts for all collected languages.

Malwi, Bhili, and Baiga}. More distantly related languages like Gujarati, Pali, Bangla and Sanskrit are placed on the outer parts of the tree. However, we would have also expected to see Khadi Boli closer to Haryanvi, and Bajjika closer to Angika and Bhojpuri.

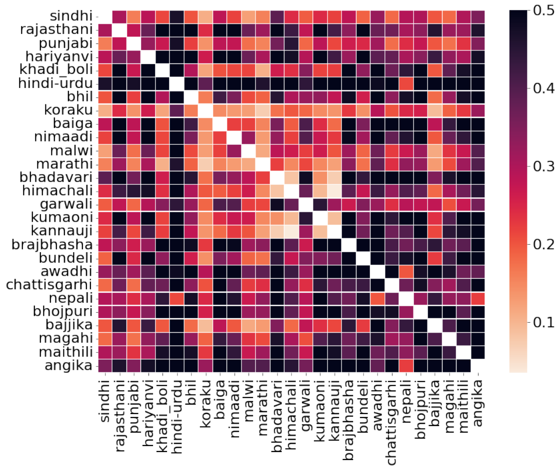


Figure 6: Lexical Overlap, all languages

A.3 Subword-level

See Figure 5 for a heatmap capturing pairwise symmetric KL-Divergence over subword distributions. Trends are similar to those seen in overlap-based measures; however, we see that the similarities against Hindi are lower, suggesting lower influence of corpus size on the measure.

B Evaluation Data

B.1 Existing resources

For some Band 1 languages (specifically, Hindi, Nepali, and Marathi), we have WordNets from the IndoWordNet project (Sinha et al., 2006; Debasri et al., 2002), from which we can extract equivalents across languages. We are not concerned, therefore, with searching for multilingual lexical resources for Band 1 languages. For some Band 2 languages (Bhojpuri, Magahi, and Maithili), WordNets are under way (Mundotiya et al., 2021) but as yet unavailable.

For Band 3, as discussed, we do not have any pre-existing bilingual or multilingual lexical resources in a convenient format. We therefore look for bilingual lexicons in the "wild"; that is, blogs, websites, scanned dictionaries, etc. We list all such raw material that we found that could be potentially useful for this purpose in Table 4. The names of these resources are listed separately in Table 5.

We exclude a few other resources we found due to too small a length ($<$ 30 word pairs), or too unstructured a format; these are unlikely to be of much help to the NLP community.

B.2 Overview of existing resources

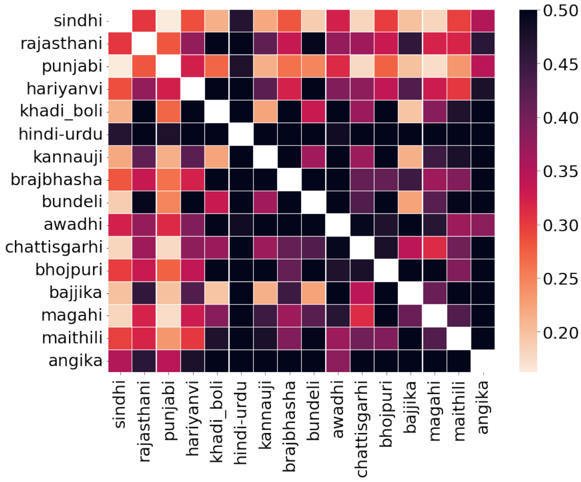
The listed resources cover 4 Band 2 languages and 7 Band 3 languages: this is counting "Bihari" as the same as Bhojpuri, and Rajasthani the same as Marwari. Note that these resources may cover more languages; we have only listed the ones relevant to this project in the "Languages" column. These resources have widely different domains, content types, and formats.

Four of the listed websites disable copying and webpage inspection, discouraging crawling or reusing their data; this means that 3 Band 3 languages are once more resource-less.

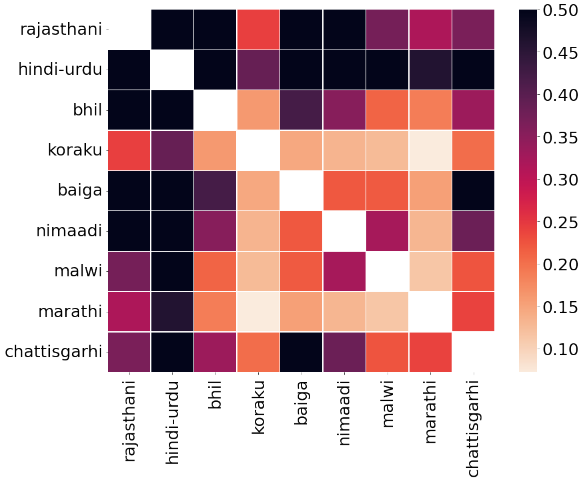
Content-wise, we see that many resources have explanations on the target side (Hindi or English), rather than equivalents. For this project, that means that the resource is not really ready-to-use as a bilingual lexicon, but will require further work in terms of extracting equivalents from the explanations for the target side, or recasting it as a lexicon of similar words on the target side, etc. $R11$ for Rajasthani also requires transliteration for the source side before it is useful. Finally, we note that even the resources listed as containing equivalents in Table 4 usually contain a mixture of equivalents, explanations, and examples. That is, each resource would require considerable processing, possibly manual, to yield a relatively noiseless bilingual lexicon.

As we discussed, for the purposes of this project, we would like to have not only bilingual lexicons per language with an anchor (preferably Hindi), but also considerable intersections between the lexicons to allow the potential of testing multilingual interactions beyond Hindi-*lang* tasks. This too, unfortunately, is likely to be a problem when gathering resources from different sources with rather small lists, although we can hope to find some common words.

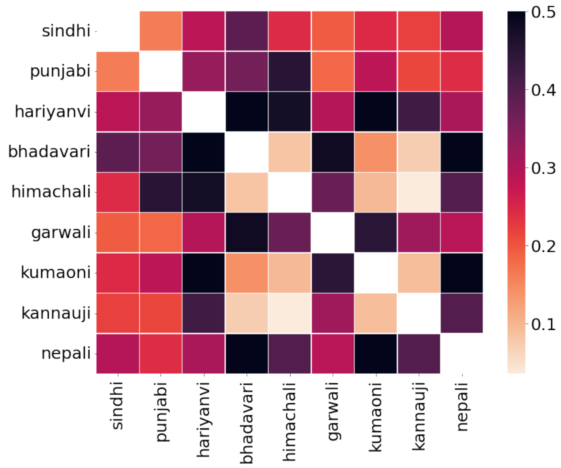
Given the above problems, including potential extensive manual efforts to the above individual resources usable, probable multilingual mismatch, and low coverage of Band 3 languages despite it all, we decided not to attempt garnering lexicons from these different resources for individual languages with the intention of putting them together.



(a) Lexical Overlap, "North central" cluster of languages



(b) Lexical Overlap, "Central" cluster of languages



(c) Lexical Overlap, "Northern" cluster of languages

Figure 7: Pairwise lexical overlap for different subsets of languages

Resource	Languages	Anchor language	Content notes	Format	Approx. length
R1	Rajasthani^r	Eng.^{r}	Explanations in English	Simple list	>500
R2	Rajasthani^d	Hin^d, Eng^r	Hindi equivalents, English explanation	Webpages by initial letter	> 500
R3	Angika^d	Hin^d, Eng^r	Explanations	Each word on diff. page, disabled copying	102
R4	Bundeli^d	Hin^d	Equivalents	Simple listing, disabled copying	Few 100s
R5	Haryanvi^d	Hin^d	Equivalents	Simple list	< 100
R6	Chattisgarhi^d	Hin^d	Explanations	Webpage per word, disabled copying	< 100
R7	Chattisgarhi^d	Hin^d	Equivalents	List, disabled copying	Few 100s
R8	Kumaoni$^{d\ r}$	Hin^d, Eng^r	Equivalents, categorized by themes	Simple list	< 100
R9	Brajbhasha^d	Hin^d	Equivalents/ explanations	Mixture of paragraphs and lists, rather disorganized	Few 100s
R10	Bhojpuri^d	Hin^d	Mostly equivalents, also Hindi synonyms	Simple list	400
R11	Hindi^r, Marathi^i, Nepali^i, "Bihari"^{i}, Magahi^{d,i}, Marwari^i	-	Cognates	Swadesh list	207
R12	{Bhojpuri, Garwali, Hindi, Marathi, Nepali, Magahi, Maithili, Sindhi}^{d,i}	Eng^r	Short phrase translations	Simple list	45 phrases (on avg.)

Table 4: Raw resources found for different languages. The superscripts ^{d}, ^{r} and ^{i} indicate that the script used for the language is Devanagari, Roman or IPA respectively. The lexicon length given is an approximation because some of these formats make it difficult to get the exact number of entries.

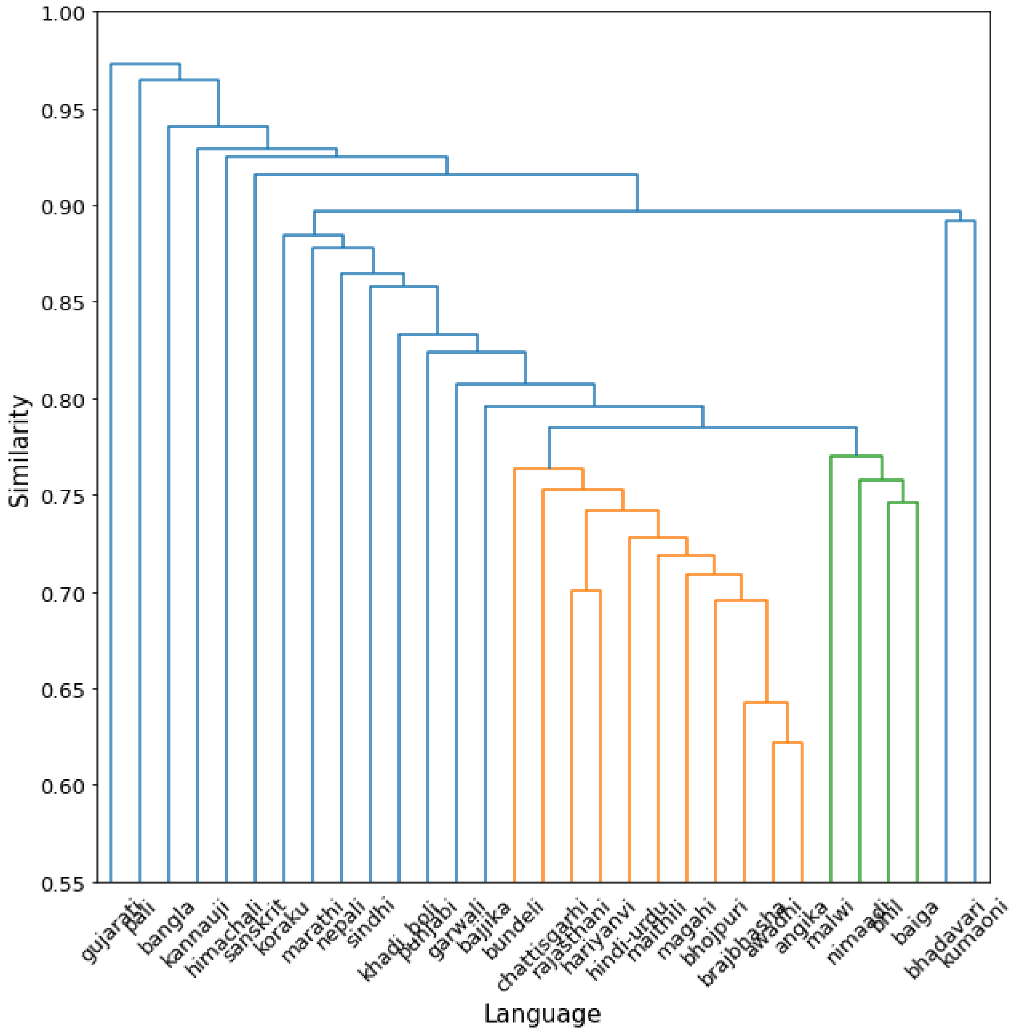


Figure 8: Dendrogram based on lexical overlap.

R11 is naturally exactly what we would have liked to find, although, again, it may require transliteration from IPA from most languages to be useful (and for Hindi, from a "casual" Roman script). The main problem, however, is that it deals with 3 Band 1 languages (for which we already have lexicons), 2 Band 2 languages, and only 1 Band 3 language, making it a low-coverage resource for our situation.

R12 is another interesting multilingual resource, highly similar to the resource that we finally decided to use, discussed in Section 5.2.

Note that a couple of these resources are valuable on their own, e.g. *R10* for Bhojpuri is extensive, simply formatted, and relatively neat and consistent; it will not require too much manual work to convert it into a usable resource for linguists. Similarly, *R1* and *R2* in Rajasthani provide the raw material for good bilingual lexicons, although they will first require a good quality transliteration into Devanagari for the Rajasthani side.

B.3 Collected data

Example of parallel sentence from "Languages Home":

```
English: Will you give me your pen?
Hindi: Kya tum mujhe apna pen doge?
```

We see that the word "pen" is code-switched in Hindi, rather than using the Hindi word "kalam". However, in other languages such as Bagheli, we see the word "kalam" used instead.[24] Therefore, although the word "kalam" exists in both languages, this relationship is not obscured because the trans-

[24] By itself, this difference is not a bad thing given that the purpose of this website is language learning. In Hindi, the given parallel sentence is absolutely natural-sounding - people do often code-switch the word "pen". Code-switching with English may be less common in less urban languages such as Bagheli; thus accounting for the use of the native word "kalam".

Resource	Name
R1	Rajasthani Language Dictionary | Rangrasiya
R2	Glossary of Rajasthani Language - Jatland Wiki
R3	Angika Shabdkosh
R4	Bundeli Shabdkosh
R5	(Blog post) Learn Harayanvi Language Through Hindi Language
R6	Chattisgarhi-Hindi online dictionary
R7	(Post) HS MiXX Entertainment
R8	Kumaoni Boli
R9	(Blog post) Learn Brajbhasha Vocabulary
R10	(Blog post) Bhojpuri dictionary
R11	(Blog post) Swadesh Word List of Indo-European languages
R12	Omniglot

Table 5: Resource websites: indexed according to Table 4

lator chose to use a different equivalent instead (in this case, code-switched, but not necessarily so in other sentences).

We report per-language statistics of the Hindi-parallel transliterated data in Table 6.

C CI: Using semantic similarity

C.1 Training embeddings: Visualizations

We use t-SNE (Van der Maaten and Hinton, 2008) to obtain the following visualizations; we performed these for *joint* models of Bhojpuri, Rajasthani, Hariyanvi, Magahi, and Korku (with Hindi-Urdu). See Figure 10 for Bhojpuri (the others are similar).

The main observations we can make for this type of model, common to all the languages, is that the low-resource target language words seem to be clustered around each other, whereas Hindi words and words belonging to both languages are better situated according to their semantics.

For the UPSAMPLE models, we visualize the same words for these languages; we present a representative (Bhojpuri) plot in Figure 10 (lower figure). While it is not clear from the visualization that the JOINT_UPSAMPLED models are less language-wise clustered than the JOINT, the target language words seem at least much better distributed, and we see more meaningful collocations (both monolingual in the target language, and cross-lingual) that we did not see before, such as “we”, “our” (cross-lingual) in the Bhojpuri. However, it is difficult to say from such visualizations which space is better

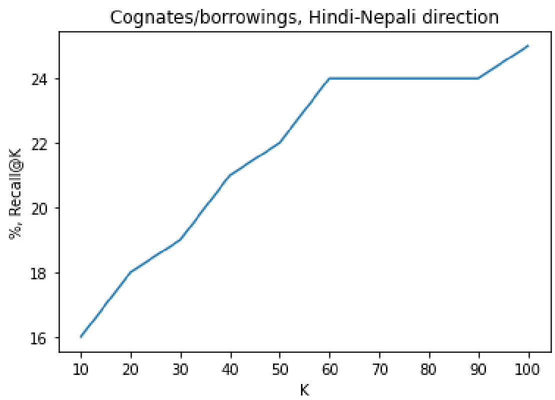


Figure 9: Recall@K for the bilingual FASTTEXT Nepali embeddings.

embedded.

C.2 Evaluating embeddings

C.2.1 Measuring Integration: *cl_integ*

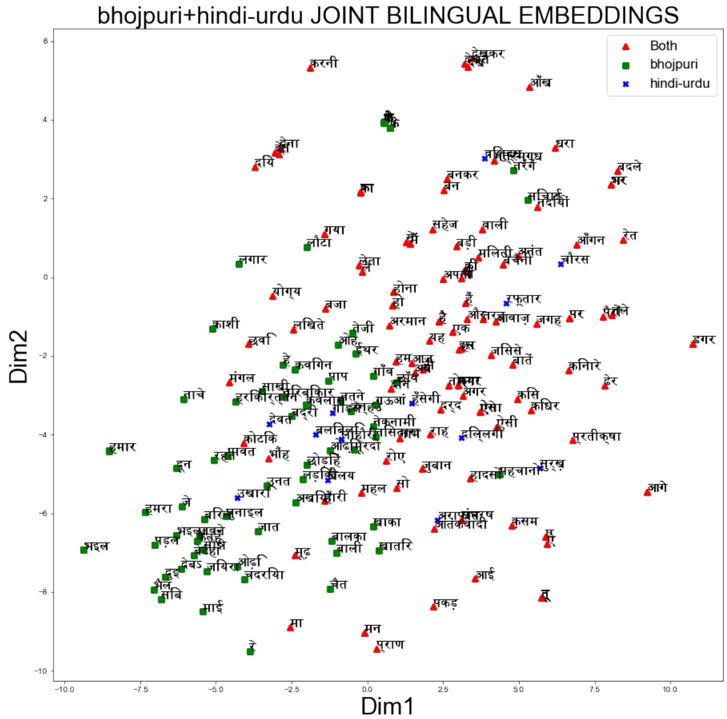
See Table 7 for the evaluation for JOINT as well as UPSAMPLE embeddings for all languages over the *cl_integ* metric.

C.2.2 Evaluating embeddings: Nepali WordNet

As mentioned before, we do in fact have WordNets from the IndoWordNet project (Kakwani et al., 2020) for Nepali and Marathi, from which bilingual lexicons can easily be extracted. While the Marathi dataset in our current collection is not very representative as previously discussed, we evaluate the Nepali-Hindi bilingual space using the

Language	Total in corpus	Unique in corpus	Total in test	Unique in test	Common in corpus and test	Frac. covered in corpus[1]	Frac. covered in test[2]
Brajbhasha	156986	30194	299	161	93	0.12	0.65
Angika	1253545	91757	310	165	102	0.09	0.60
Maithili	218491	41434	273	147	81	0.09	0.54
Magahi	79405	16942	326	172	81	0.11	0.64
Hindi-Urdu	7100394	197355	336	171	165	0.25	0.98
Awadhi	490877	53103	281	145	109	0.05	0.82
Rajasthani	187708	34360	312	161	124	0.11	0.84
Hariyanvi	232526	27431	298	156	123	0.13	0.86
Bhil	27246	5557	319	177	68	0.12	0.48
Chattisgarhi	83073	14463	267	134	95	0.16	0.76
Nepali	688865	104687	203	118	65	0.04	0.62
Bajjika	7412	2788	317	149	55	0.13	0.53
Koraku	15508	2278	262	132	17	0.04	0.23
Malwi	9626	2883	325	163	51	0.12	0.46
Sindhi	52659	11850	250	141	55	0.09	0.51
Bhojpuri	196513	34051	303	146	110	0.16	0.83
Garwali	90234	22655	275	161	86	0.07	0.64
Marathi	3109	1685	230	130	29	0.05	0.37
Kumaoni	1013	441	250	171	16	0.10	0.16
Bundeli	26902	7991	272	147	82	0.12	0.63

Table 6: Evaluation token data statistics post-transliteration, after aligning with Hindi. [1] This reports the fraction of the corpus (token-wise) that is contained in the test, vice-versa for [2].



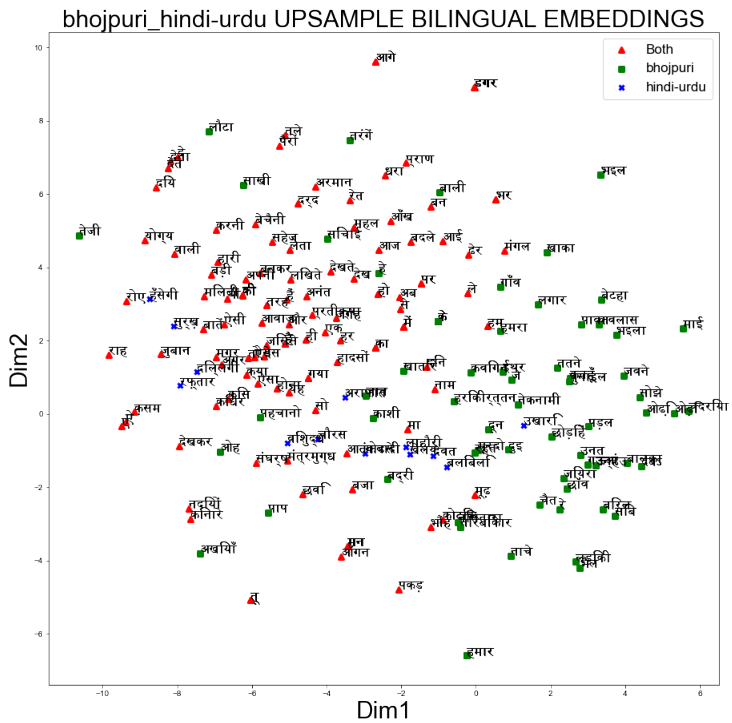


Figure 10: t-SNE (Van der Maaten and Hinton, 2008) Visualization of Bhojpuri-Hindi bilingual space, JOINT (up) and UPSAMPLE (down)

	J_12	J_21	U_12	U_21
Sindhi	0.53	0.23	0.31	0.33
Rajasthani	0.78	0.33	0.62	0.40
Punjabi	0.58	0.19	0.40	0.27
Hariyanvi	0.75	0.30	0.66	0.36
Khadi Boli	0.99	0.18	0.76	0.13
Sanskrit	0.33	0.28	0.12	0.26
Bhil	0.92	0.24	0.53	0.34
Koraku	0.59	0.13	0.34	0.10
Baiga	0.97	0.21	0.73	0.31
Nimaadi	0.87	0.16	0.47	0.21
Malwi	0.88	0.14	0.45	0.13
Marathi	0.95	0.20	0.32	0.15
Bhadavari	1.00	0.12	0.81	0.30
Himachali	1.00	0.07	0.48	0.07
Garwali	0.64	0.25	0.25	0.39
Kumaoni	0.97	0.09	0.74	0.05
Kannauji	1.00	0.04	0.66	0.14
Brajbhasha	1.00	0.32	0.74	0.38
Bundeli	0.99	0.21	0.58	0.36
Awadhi	0.69	0.34	0.45	0.43
Chattisgarhi	0.86	0.29	0.51	0.36
Nepali	0.37	0.39	0.31	0.48
Pali	0.57	0.11	0.07	0.10
Bhojpuri	0.91	0.32	0.74	0.41
Bajjika	1.00	0.20	0.74	0.30
Magahi	0.84	0.21	0.44	0.42
Maithili	0.85	0.38	0.57	0.49
Angika	0.63	0.44	0.50	0.40

Table 7: cl_integ values reported as 0-1 measure for both sets of embedding spaces, in both directions. 12 indicates that we consider the non-Hindi language as source, and look for the fraction of nearby Hindi words, 21 is vice versa.

# tokens	integ_12	integ_21	bl_12	bl_21
JOINT				
5000	0.43	0.37	0.30	0.21
50000	0.33	0.38	0.29	0.21
100000	0.29	0.37	0.29	0.20
500000	0.33	0.44	0.29	0.20
UPSAMPLE				
500000	0.29	0.42	0.33	0.15

Table 8: Recall@50 for Nepali data splits of different sizes against Hindi-Nepali lexicon obtained from IndoWordNet. 12: Nepali as source, 21: Hindi as source. We also show results for *cl_integ* and bilingual lexicon tests for UPSAMPLE Nepali model

Nepali WordNet. We used the WordNet to extract a Hindi/Urdu-Nepali bilingual lexicon, and we calculated Recall@K, in the following way: for each Hindi-Urdu word, we extract its K nearest neighbours. If any of those are the gold target, we count a full point for that word. Finally, we report the total such points as a percentage of the length of the gold bilingual lexicon.

See the results for the *joint* Nepali model in Figure 9.

Nepali is in the highest range of availability in our current dataset, so we do not expect these results to be representative for other languages with less data. We therefore also look at these results over artificially smaller cuts of the Nepali dataset. See Table 8. We also report these numbers for the UPSAMPLE Nepali model (all data included) in the same table.

C.2.3 Discussion

There are a couple of interesting things to note about the above results. We see that cl_integ shows high values from the LRL to Hindi direction, but not vice versa. Nepali happens to be an outlier in this case, which is perhaps unfortunate since it is unlikely to be representative of the other languages, and it is the only language we can evaluate with more detail.

We notice in Table 8 that the results for the WordNet bilingual lexicon test seem to be stable across different data splits. This is rather suspicious; however, a possible explanation is that the positives accrue from frequent words anyway, possible also present in the Hindi-Urdu data; therefore, reducing the number of Nepali tokens does not seem to affect this number. Note that this is not at all an indication that the resulting embeddings are of the same quality, simply that this metric is not able to capture possible underlying damage.