Locally Measuring Cross-lingual Lexical Alignment: A Domain and Word Level Perspective

Taelin Karidi, Eitan Grossman, Omri Abend
Hebrew University of Jerusalem
{taelin.karidi,eitan.grossman,omri.abend}@mail.huji.ac.il

Abstract

NLP research on aligning lexical representation spaces to one another has so far focused on aligning language spaces in their entirety. However, cognitive science has long focused on a local perspective, investigating whether translation equivalents truly share the same meaning or the extent that cultural and regional influences result in meaning variations. With recent technological advances and the increasing amounts of available data, the longstanding question of cross-lingual lexical alignment can now be approached in a more data-driven manner. However, developing metrics for the task requires some methodology for comparing metric efficacy. We address this gap and present a methodology for analyzing both synthetic validations and a novel naturalistic validation using lexical gaps in the kinship domain. We further propose new metrics, hitherto unexplored on this task, based on contextualized embeddings. Our analysis spans 16 diverse languages, demonstrating that there is substantial room for improvement with the use of newer language models. Our research paves the way for more accurate and nuanced cross-lingual lexical alignment methodologies and evaluation.

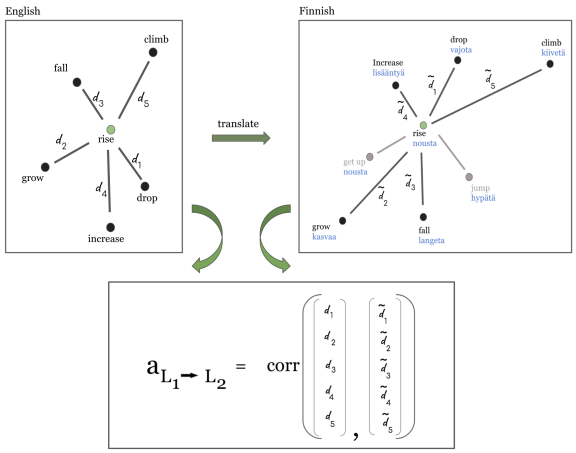

Figure 1: Distribution-based alignment. An illustration of the distribution-based alignment process for the word *rise* between English (L_1, upper left) and Finnish (L_2, upper right). First, we find the k nearest neighbors of the word *rise* in L_1 (for ease of visualization $k = 5$). We then translate the neighbors to L_2. The neighbors do not necessarily coincide with the nearest neighbors of the Finnish translation of *rise*. Here two of the nearest neighbors are not translations (marked in grey) and are ignored in the calculation). We then measure the correlation between the distance vectors. We repeat this in the opposite direction, $L_2 \rightarrow L_1$, and compute the average.

1 Introduction

Cross-lingual lexical semantic similarity can be approached from two complementary perspectives: *local* and *global*. The local perspective compares words or sets of words and characterizes how similarly meanings are lexicalized across languages (Wierzbicka, 1972; Majid et al., 2008; Berlin and Kay, 1991; Srinivasan and Rabagliati, 2015; Thompson et al., 2020; Georgakopoulos et al., 2022; Purves et al., 2023). For example, whether translation equivalents, such as the English *green* and French *vert*, encode the same meaning. In contrast, the global perspective focuses on how similar languages are as a whole, examining broader patterns, relationships, and structures within languages, rather than focusing on individual words. For instance, English and Bulgarian are both SVO languages, although the latter has a more flexible word-order structure. Such differences can influence their global alignment and the ability to transfer knowledge between them (Nikolaev et al., 2020; Arviv et al., 2023).

In the field of Natural Language Processing (NLP), most research on cross-lingual similarity focuses on global similarity measures (Conneau et al., 2017; Artetxe et al., 2018; Ruder et al., 2019; Vulić et al., 2021), as they facilitate cross-lingual transfer in a multitude of tasks, such as machine translation and bilingual lexicon induction (Schus-

Findings of the Association for Computational Linguistics: EACL 2024, pages 15877–15893
November 12-16, 2024

ter et al., 2019; Artetxe et al., 2019; Eronen et al., 2023). These approaches typically involve aligning entire vector spaces of different languages through methods such as linear transformations, aiming to create a unified semantic space where languages can be directly compared.

The local perspective, although often overlooked in NLP, has long been a topic of interest in cognitive sciences and linguistics (Whorf, 1956; Fodor, 1975; Frawley, 1998; Burns, 1994; Snedeker and Gleitman, 2004; Majid et al., 2008; Croft, 2010; Youn et al., 2016). To the extent that lexicons reflect the structure of human cognition, understanding how meaning is expressed across languages offers insight as to how humans categorize and represent the world. Examples of such inquiries span various semantic *domains* (a way of grouping words together based on common aspects of meaning or function), including colors (Berlin and Kay, 1991) and emotions (Jackson et al., 2019).

Traditionally, in linguistic and cognitive research, comparing the meaning of words across languages involves methodologies and approaches that are less data-driven in nature, but rather prioritize in-depth, relatively small-scale exploration of meaning, such as descriptive comparisons (Wierzbicka, 1972), elicitation studies (Barnett, 1977; Tokowicz et al., 2002; Moldovan et al., 2012; Allen and Conklin, 2013; Purves et al., 2023) and semantic maps (Haspelmath, 2003; Croft, 2022).

However, defining and operationalizing the notion of cross-lingual lexical similarity is notoriously difficult and controversial. The difficulty in defining lexical similarity has motivated a turn from theory-driven to data-driven approaches. Indeed, considerable recent work pursued data-driven approaches to the quantification of equivalency between word pairs in different languages (Majid et al., 2014; Youn et al., 2016; Jackson et al., 2019; Georgakopoulos et al., 2022). While some studies employed NLP methods (Thompson et al., 2018, 2020; Rabinovich et al., 2020; Beinborn and Choenni, 2020), their application has been limited to static word representations and their validation have focused on converging evidence, i.e, getting the same (or similar) result in different means and looking for confounds. E.g, comparing to other tests from the cognitive science literature, such as picture naming (Duñabeitia et al., 2018) or translation norms (Tokowicz et al., 2002; Allen and Conklin, 2013). However, these test are only applicable to a very small set of languages and word lists. Moreover, in NLP, it is common to compare to some external reference point that we perceive is one of high quality. For brevity, we simply refer to it as ground truth (Section §5.2).

In this work, we focus on the challenging task of word-level semantic similarity across languages, addressing the question of how word meanings vary across languages. We use a range of metrics designed to assess and compare the nuanced meanings of translation equivalents in different languages, and extend them in a novel way to include contextualized word representations (Section 4.2).

One of the main challenges in developing metrics to quantify differences in lexical semantics across languages is the difficulty to validate them, given that there are no available resources tailored to define ground truth in this area. To address this, we not only conduct extensive synthetic analyses of the metrics (Section §5.1) but also establish a novel validation method, adapting a newly compiled linguistic resource of lexical gaps in the kinship domain (Section §5.2; Khishigsuren et al., 2023).

We perform a detailed comparison between the various metrics at two levels of granularity: word-level and domain level. We also analyze what features affect the alignment, using a combination of lexical and environmental features. We show that at the domain-level there is substantial similarity between the methods and that it offers a more stable level for such analysis. Moreover, while the contextualised embeddings (CE's) based metrics are substantially correlated with our naturalistic validation, the other methods are not, suggesting that there is definitely room for improvement in this area, using newer models.

To summarize our contributions we (1) formulate the question of cross-lingual lexical similarity as an NLP task; (2) compare and analyze various metrics for this task; (3) introduce novel metrics based on contextualised word embedding; (4) offer a comprehensive validation suite to support our findings, including a novel validation method against a high-quality linguistic resource specifically tailored to the kinship domain.

2 Related Work

A multitude of different approaches for computing distributional similarity have been explored in NLP, of which we select a number of representative examples. Distributional metrics can be clas-

sified based on whether they employ a joint space for the embeddings for the languages in question, or whether the spaces are trained monolingually and then aligned (Artetxe et al., 2018; Conneau et al., 2017).[1] The latter approaches have been a key facilitator of cross-lingual transfer in NLP and are especially important in low-resource settings. However, for identifying patterns of divergence and convergence in the usage of specific words and domains, this approach is suboptimal.[2] Globally optimal alignment (one that minimizes the distance between the image of one language in the space of another language) may distort the alignment of some words subsets, in the interest of improving the alignment of other, larger word sets.[3]

Local alignment, or the extent to which translation pairs like English *home* and Spanish *casa*, hold a similar meaning across languages, is a well studied open question in cognitive science (Berlin and Kay, 1991; Majid et al., 2008, 2014; Youn et al., 2016; Jackson et al., 2019), that had only recently been approached with NLP tools (Thompson et al., 2018, 2020). However, existing metrics are limited to static word embeddings and do not accommodate newer models that support contextualization.

Additionally, understanding the variability in meaning across languages can provide valuable insight into cultural differences, revealing how various societies conceptualize their unique experiences and worldviews (Qi, 2017; Khalilia et al., 2023; Shioiri et al., 2023; Tjuka et al., 2024). This line of research aligns with the recent surge in studies concerning multicultural knowledge in LLMs, which assess whether models like GPT variants or multimodal LMs possess diverse cultural knowledge or exhibit biases favoring Western cultures (Hershcovich et al., 2022; Havaldar et al., 2023; Ventura et al., 2023; Cao et al., 2023).

[1]We are aware of one study of cross-lingual lexical comparison that used global alignment to project languages to a shared space, and defined the degree of alignment between a translation pair to be the distance of the image of one word to the embedding of the other (Rabinovich et al., 2020). However, due to the reasons stated above, we do not consider their approach in this paper.

[2]To measure cross-lingual lexical alignment using global alignment, it is natural to define the distance (i.e., alignment) between a translation pair as their cosine distance in the shared (aligned) space. This definition is employed to (1) align the source and target language spaces, and (2) evaluate the accuracy of the alignment.

[3]Preliminary experiments conducted across various languages have shown that these methods do not correlate with other measures or with the validations we propose.

ENGLISH FORM	CONCEPT	DOMAIN
mother	mutter::N	Kinship
mind	verstand::N	Cognition
go	gehen::V	Motion
today	heute::ADV	Time
towel	Handtuch::N	Clothing
business	Geschäft::N	Modern world
hold	halten::V	Possession
one	eins::NUM	Quantity
floor	Fußboden::N	The house
flower	Blume::N	Agriculture
middle	Mitte::N	Spatial relations
happiness	Glück::N	Emotions
horse	Pferd::N	Animals
red	rot::A	Sense perception
break	brechen::V	Basic actions
church	Kirche::N	Social
write	schreiben::V	Language
bread	Brot::N	Food and drink
skin	Haut::N	The body

Table 1: Concepts and their domains. Examples of concepts, labled according to the NEL dataset (§3). "Domain" designates the semantic domain the concept belongs to, and "English Form" designates the lexicalization of each concept in English. For space considerations, "Clothing" denotes "Clothing and grooming", "Agriculture" denoted "Agriculture and Vegitation", "Basic Actions" denotes "Basic actions and technology", "Social" denotes "Social and political relations", "Emotions" denotes "Emotions and values" and "Language" denotes "Speech and language".

3 Experimental Setup

We provide a brief description of our experimental setup, with full details available in Appendix §A.

Languages. We perform our analysis on a diverse set of 16 languages, spanning 7 different top-level language families from many geographical areas across Eurasia: English (eng), French (fra), Italian (ita), German (deu), Dutch (nld), Spanish (spa), Polish (pol), Finnish (fin), Estonian (est), Turkish (tur), Chinese (chn), Korean (kor), Japanese (jap), Hebrew (heb), Hindi (hin) and Arabic (arb).

Data. We conduct our analysis on the NorthEuraLex (NEL) dataset, a lexical resource compiled from dictionaries and other linguistic resources, such as concept lists, available for individual languages in Northern Eurasia. NEL comprises a list of 1016 distinct *concepts*[4] together with their word forms in 107 languages (Table 1). Rare cases where a concept does not have any realization in a given language are excluded for that language.

[4]The concepts in NEL are given in German (see Table 1).

Models & Corpus. In the main paper, for static word embeddings we use fastText[5] 300-dimension word embeddings, trained on Wikipedia using the skip-gram model (Bojanowski et al., 2017). For contextualised word embeddings (CE's) we use mBERT[6] (*bert-base-multilingual-uncased* model) 768-dimension vectors for the 16 languages. To extract sentences for SNC-CLOUD, we use the Leipzig corpus.[7] For concepts, their translations and semantic domains we use the North Euralex (NEL) dataset (Dellert et al., 2020). Other models and datasets that we experiment with, together with full explanations of how we use each dataset are detailed in Appendix §A.

4 Cross-lingual Lexicon Alignment

In this section, we lay the groundwork for cross-lingual lexicon alignment, focusing on word-level semantic similarity between languages. We use local metrics (Hamilton et al., 2016a; Thompson et al., 2018, 2020), extend them to novel ones and provide a framework for comparing word meanings across languages.[8]

4.1 Computational Framework

The computational framework we adopt in this paper, which we term Semantic Neighborhood Comparison (SNC), relies on the relationships between the nearest neighbors of translation equivalents to compare embeddings across different spaces. This approach has been used in various forms for both computational historical linguistics and lexical similarity tasks (Hamilton et al., 2016b; Thompson et al., 2020; Beinborn and Choenni, 2020). We experiment with several variants of this approach, including one based on contextualized word embeddings, which is, to our knowledge, novel.

Notation. Let $\mathcal{C}$ be the set of concepts in the NEL dataset (Dellert et al., 2019, see §3). We adopt the notion of a concept from the lexical typology literature (e.g., Dellert et al., 2019; Rzymski et al., 2020)[9], and take it to mean a word sense defined independently of any specific language. Let Ω be a set of languages. We denote the lexicon corresponding to $\mathcal{C}$ in a given language L with $\mathcal{L}$, and note that $|\mathcal{L}| \leq |\mathcal{C}|$ for every language. We assume that $\mathcal{C}$ is partitioned into domains, and denote the (non-overlapping) domains with $\mathcal{D}_1, \ldots, \mathcal{D}_m$.

Given a concept $c \in \mathcal{C}$, we denote its lexicalization (the word expressing that concept) in language L with $r_L(c) \in \mathcal{L}$. A translation pair between languages L_1 and L_2 is a pair of words $(w_1, w_2) \in \mathcal{L}_1 \times \mathcal{L}_2$, such that there exists $c \in \mathcal{C}$ such that $r_{L_1}(c) = w_1$ and $r_{L_2}(c) = w_2$. For example, the concept SONG gives rise to the English-French translation pair *(song,chanson)*. In principle, several translation pairs may correspond to a concept and language pair, but in the data we experiment with, this does not occur.

For a given word w in a given language L, we denote its embedding with $emb(w, L)$. We denote the embedding space corresponding to L with ℓ.

4.2 Alignment Metrics

Let $c \in \mathcal{C}$ be a concept and $w_1 = r_{L_1}(c) \in \mathcal{L}_1$, $w_2 = r_{L_2}(c) \in \mathcal{L}_2$ its lexicalizations, and $v_1 = emb(w_1, L_1) \in \ell_1$, $v_2 = emb(w_2, L_2) \in \ell_2$ their respective embeddings. We compute its k nearest neighbors in ℓ_1 with $\{n_1^{(1)}, ..., n_k^{(1)}\}$ ($k = 100$).[10] We then translate the nearest neighbors to L_2 using the NEL dataset[11], by taking their translation pairs, and denote the resulting vectors with $\{n_1^{(2)}, ..., n_k^{(2)}\} \in \ell_2$.

Neighbors Overlap (NO). A naïve approach to comparing the meaning of a concept across languages is to compare the number of overlapping nearest neighbors of a word and its direct translation across languages (Thompson et al., 2018). This approach is intuitive and stems from the distributional definition of meaning as the semantic neighborhood of the concept.

For a concept c, we back-translate its k nearest neighbors in ℓ_1 and ℓ_2 to $\mathcal{C}$[12] and define the alignment to be the amount of overlapping neighbors (in $\mathcal{C}$) divided by k.

Semantic Neighborhood Comparison (SNC). Although neighbors overlap has proven valuable for evaluating word-level similarities (Thompson

[5] https://fasttext.cc/docs/en/unsupervised-tutorial.html

[6] https://huggingface.co/bert-base-multilingual-uncased

[7] https://corpora.uni-leipzig.de/en?corpusId=deu_news_2021

[8] In §2 we provide a further explanation as to why we choose this set of metrics.

[9] A language $L \in \Omega$ may or may not lexicalize a concept $c \in \mathcal{C}$, and may lexicalize several concepts with one word (colexification).

[10] See Appendix §A for hyperparameters details.

[11] Translation retrievel method explained in Appendix §A.

[12] To enable the intersection computation, the concepts need to reside in a "joint space", here, the concept space $\mathcal{C}$ can be thought of as an interlingua.

et al., 2018), it falls short in capturing the intricate semantic relationships within the groups of neighbors. To address this limitation, we define the key metric in this paper, which serves as the foundation for all other variants. We define the unidirectional metric as

$$a_{L_1 \to L_2} = \rho\left(\left(cos(v_1, n_i^{(1)})\right)_{i=1}^{k}, \left(cos(v_2, n_i^{(2)})\right)_{i=1}^{k}\right)$$

ρ is the Pearson correlation coefficient.[13] The bidirectional metric as the arithmetic mean over the two directions:

$$a_{L_1 \leftrightarrow L_2} = \frac{a_{L_1 \to L_2} + a_{L_2 \to L_1}}{2}$$

We refer to this alignment strategy as SNC-STATIC.

Contextualised Word Embeddings. We now turn to detailing metrics that are analogous to SNC-STATIC, but instead use CEs.[14]

SNC-AVE For word $w \in \mathcal{L}$, we extract its representation from all layers (if w is tokenized to multiple subwords, we average over the subword representations). We average the outputs from layers 1-12 to define the final vector for w. We then proceed with the SNC process, as described with SNC-STATIC.[15]

SNC-CLOUD For word $w \in \mathcal{L}$, we extract all sentences (with a threshold of 1000) that w appears in, from an auxiliary corpus (see §3). We extract the CEs (from layer 12, if it is tokenized to subwords, we average over them) for w from each of the sentences. Denote these vectors with $V_w = \{v_{1_w}, ..., v_{k_w}\} \subseteq \mathbb{R}^{768}$. In this setting, each word w is represented by a point cloud of vectors V_w. Hence, the distance between two words is the distance between their corresponding point clouds. We define *point-cloud distance* as follows:

$$d(w, \tilde{w}) = min_{i,j}\ cos(v_{i_w}, v_{j_{\tilde{w}}})$$

We follow the SNC procedure (defined above) under this definition of distance.

[13] We conducted experiments with Spearman correlation, as well as Kendall τ. They present similar trends and are omitted due to space considerations.

[14] We denote contextualised word embeddings by CEs.

[15] We follow the work of (Vulić et al., 2020) on probing contextualized models for lexical semantics, averaging across layers bottom-to-top. However, we experimented with alternative settings, such as pooling from the top layer or averaging up to layer n (n<12), and found this method to be the most stable overall.

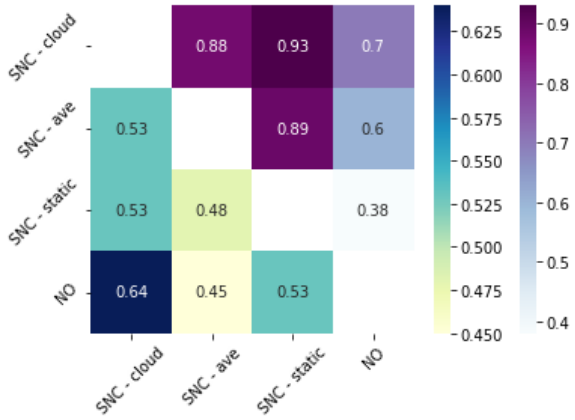

Figure 2: Correlation between the various metrics. Pearson correlation is computed for different aggregation methods. The **upper** matrix represents concept-level correlations, while the **bottom** matrix represents domain-level correlations. All correlation values are significant with $p < 0.05$.

5 Validation Experiments for SNC

Due to the opaque nature of distributional metrics, it is important to verify their validity as metrics for cross-lingual alignment of translation pairs. While some of the metrics we use are taken from the literature, others have not yet been used for lexical alignment , as they were either adapted from other tasks of lexical similarity in NLP, or are novel adaptations of existing metrics. Unlike in many NLP papers and tasks, there is no gold standard definition of the extent to which the meanings of two words in different languages align, which prohibits direct validation of the metrics (Schlechtweg et al., 2020). Instead, we conduct two sets of experiments (synthetic and naturalistic) to verify both the self-consistency of the metrics (Sec §5.1) and its validity against a high quality reference (§5.2).

5.1 Synthetic Validation

Given the lack of human-evaluated data quantifying the similarity between translation equivalents across languages, we conduct synthetic experiments to verify the metrics' internal consistency. These experiments evaluate the proposed metrics without relying on any external resources.

Shuffle Baseline. To verify that the degree of alignment of words (or semantic domains) is a product of their similar semantic structure with neighboring words, rather than some unexpected artifact, we conduct a shuffle baseline test. This test evaluates whether the observed degree of align-

ment is a result of words being in a dense/sparse part of the embedding space, which may skew the results. To compute the shuffled alignment, let $\{n_1^{(1)}, ..., n_k^{(1)}\}$ be the k nearest neighbors of $w_1 \in L_1$, and $\{n_1^{(2)}, ..., n_k^{(2)}\}$ their translation equivalents in L_2. We perform a random permutation of the translations indices, such that, $\{n_1^{(2)}, ..., n_k^{(2)}\} \mapsto \{n_{p(1)}^{(2)}, ..., n_{p(k)}^{(2)}\}$, with p a permutation over $\{1, ..., k\}$. We then compute $a_{L_1 \to L_2}$ (as defined in §4.2), and do the same for the other direction $a_{L_2 \to L_1}$). If a high alignment is an artifact of the neighbors being in a dense/sparse part of the embedding space, the correlation between the alignment and its shuffled version should be close to 1 (and to 0 if it is not the case). We find that the correlation is always $r \in [-0.5, 0.5]$ with p-values $p > 0.5$, suggesting that the density does not play a major role in the observed correlations.

Sensitivity. A key step in SNC computation (Section §4.2) is the k nearest neighbors search, which is restricted to the NEL lexicon. We verify that results are robust to removal of semantic domains from the data by removing j domains ($j = 5, 10, 15$) and computing the correlation between the results before[16] and after the removal (we do this 1000 times per j). The results are highly stable to such removal, with $0.87 \leq r \leq 0.99$ and $p \leq 0.05$.[17]

5.2 Naturalistic Validation

In this section, we present a novel external validation method for cross-lingual lexicon alignment, using a recently developed, extensive resource that identifies lexical gaps across languages. A key notion in capturing lexical diversity across languages is that of the *lexical gap*, which refers to the lack of lexicalization of a particular concept in a particular language. For example, many languages lack an equivalent of the English word *cousin*, and instead employ several more specific words that distinguish male and female, elder and younger or paternal and maternal cousins (Khishigsuren et al., 2023). We use a newly released lexical resource for the kinship domain, which contains 37370 gaps in 699 languages.[18] The resource focuses on the domain of kinship as it is universally represented in human languages, but is also known to be incredibly diverse across languages and cultures. This is a unique linguistic resource, as it is the first extensive resource to cover lexical gaps across a diverse set of languages. As such, it allows for a reliable external validation of the alignment methods discussed in this paper.

We consider similar gap patterns as indicative of greater alignment between languages. Keeping this in view, we establish a metric for alignment derived from these gap patterns, which we view as a high quality reference point.

Lexical Gaps. The notion of a lexical gap is closely related to that of untranslatability (Catford, 1978). For example, Wierzbicka (2008) considered that the concept of *color* is a lexical gap in Warlpiri, an Australian Indigenous language, as it lacks a word for it. A lexical gap is defined as the absence of a specific word or term in a language to express a particular concept or idea (Bentivogli and Pianta, 2023).

Interlingua. In order to compare lexical gaps across languages, an interlingual conceptual space of the kinship domain is defined (Figure 3). It consists of 198 concepts and 347 is-a relations (e.g., *parent's male sibling*), covering the six subdomains that Kinship is usually divided into: grandparents, grandchildren, siblings, uncles and aunts, nephews and nieces, and cousins (Murdock, 1970). We denote the subdomains by $\mathcal{S}$. For subdomain $s \in \mathcal{S}$ (e.g, *sibling*), let $\mathcal{C}_s$ be the list of concepts associated with it (i.e, all the possible lexicalizations in the interlingual space, e.g, *female elder sister*). For a language L and subdomain $s \in \mathcal{S}$, $c \in \mathcal{C}_s$ is a *gap* if it is not distinctly lexicalized in L. We define the **gap pattern** of subdomain s as the set of lexical gaps for this concept, and denote it with $\xi_{L,s}$.

Comparing Lexical Gaps. To use the information about the lexical gaps as validation, we define a metric for alignment of language pairs, based on their gap patterns. Let $L_1, L_2 \in \Omega$ be a pair of languages, we define:

$$\lambda_{L_1 \leftrightarrow L_2} = \frac{1}{|\mathcal{S}|} \Sigma_{s \in \mathcal{S}} \frac{|\xi_{L_1,s} \cap \xi_{L_2,s}|}{|\mathcal{C}_s|}$$

We denote the concatenation over all pairs of languages in $\binom{\Omega}{2}$ with λ. To compare the above alignment with SNC and COLEXA, we manually

[16] The results before are the results aggregated by domain (see §6.1), computed on the full domain list, and then the domains that are selected to be removed are removed from the vector before making the comparison.

[17] This holds true for all SNC variants, models and datasets we experiment in the paper.

[18] https://github.com/kbatsuren/KinDiv

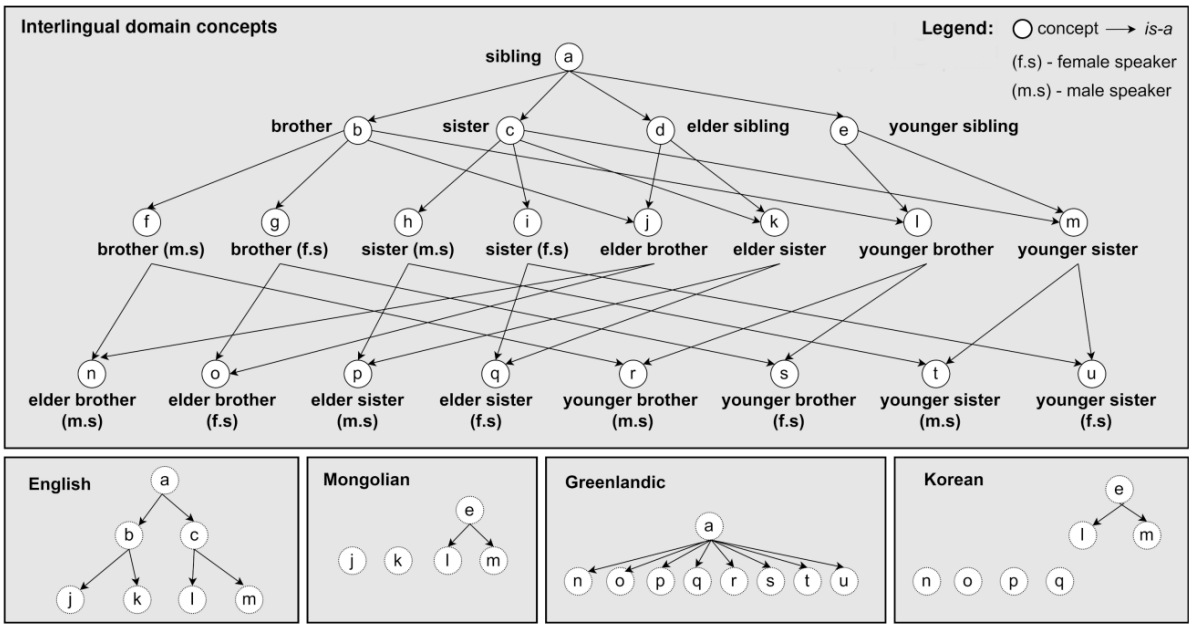

Figure 3: Lexical Gaps. Interlingual conceptual layer of sibling domain, reproduced with permission from the authors of (Khishigsuren et al., 2023).

filter the concepts $\mathcal{C}$ that SNC and COLEXA is computed on (i.e, $\mathcal{C}$) to contain only relevant concepts that are both in the Kinship domain and that can be manually mapped into a subdomain $s \in \mathcal{S}$ (e.g, the concept "onkel::N" for *uncle* in $\mathcal{C}$ is mapped to the subdomain *uncles and aunts*, however the concept "ihr::PRN" for *you* cannot be mapped to any subdomain $s \in \mathcal{S}$). We denote the filtered concepts with $\tilde{\mathcal{C}}$ and restrict the computation of SNC and COLEXA to $\tilde{\mathcal{C}}$. We compute correlation at the word-level (the domain-level is not relevant here, as we are restricted to one domain, kinship) between the various metrics as detailed in §6.1. For each alignment measure and pair of languages $(L_j, L_p) \in \binom{\Omega}{2}$, we have a vector in $\mathbb{R}^{|\tilde{\mathcal{C}}|}$. We perform aggregation over its components which results in $\mu_{L_j,L_p} \in \mathbb{R}$. The final vector $\mu \in \mathbb{R}^{\binom{\Omega}{2}}$ is a concatenation over all pairs of languages. For each alignment measure we compute the Pearson correlation between μ and λ.

6 Analysis

Having established a set of metrics and an extensive validation suite, we now turn to analyzing the alignment results. This analysis is conducted at two levels of granularity: the word level and the domain level. We compare various metrics, identifying which words and domains are most and least aligned. Furthermore, we examine factors affecting alignment, such as lexical properties and environmental features, to uncover the underlying causes of semantic divergence across languages.

	SNC-STATIC	SNC-AVE	SNC-CLOUD
Top 3	March	twelve	thirty
	August	eleven	fifty
	January	five	twelve
Bottom 3	rise	be afraid	corner
	groan	soft	soft
	set	be noisy	round

Table 2: Most and least aligned words. Word-level alignment, averaged across languages.

6.1 Word-level & Domain-level Comparison

Word-level Comparison. We start with a comparison between the metrics themselves. The most straightforward level of comparison between metrics is their word-level correlation[19].

Figure 2 presents the Pearson correlation between the metrics, and Table 2 shows the top/bottom aligned words for the metrics. Results show that SNC methods are moderately correlated among themselves (r around 0.5), meaning there is a substantial variability in their predictions at the word level. Moreover, manual inspection of the data reveals that it is challenging to infer conclu-

[19]Each metric, a concept and language pair, give rise to a vector of alignment scores (full details in Appendix §B.2).

sions at the word level off handily (Section §7).

Domain-level Comparison Alignment metrics between languages are often used to compare the degree of alignment across different domains. For example, Thompson et al. (2020) argue, based on findings with a SNC-STATIC metric, that more structured domains (e.g, Quantity) tend to be better aligned across languages. To examine the alignment at the domain level, for every measure $m \in \mathcal{M}$, we aggregate the word-level alignment over each domain (without aggregating over languages). Strikingly, as opposed to the word-level comparison, here the similarity between the methods is very high, reaching $r = 0.93$ (between SNC-CLOUD and SNC-STATIC). This finding encourages the formulation of conclusions at the domain level, as it presents to be more stable.

Relative Degree of Alignment across Domains. We examine what domains are the most/least aligned across languages. Figure 4 shows both the distribution and median of alignment values for each language pair across the semantic domains, for SNC-CLOUD. The most aligned domains are Quantity, Time and Kinship[20], whereas the least aligned domains are Motion, Basic Actions, and Technology and Possession. Similar trends are reported by Thompson et al. (2020), who argue that the high degree of alignment of these domains is related to their structure and organization along explicit dimensions (e.g., generation: grandmother/mother/daughter)y capture different notions of similarity. Table 4 presents a few examples of the differences.

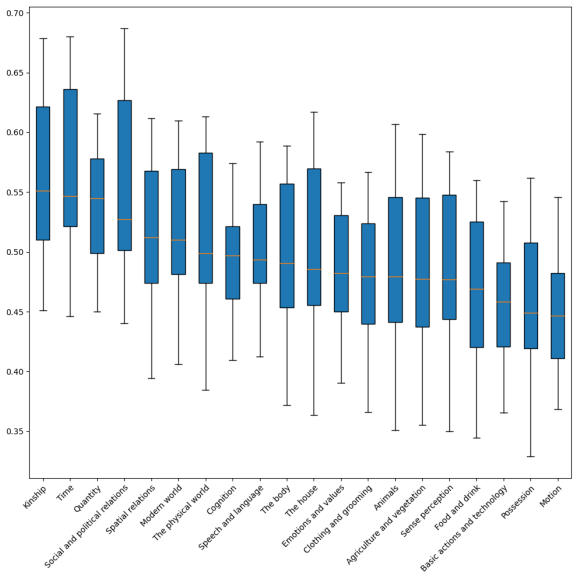

Figure 4: Alignment of domains under SNC-AVE. The domains are ranked according to the mean value of the alignment. Each box represents the distribution of alignment values (per language pair), for a specific domain (concepts-level alignment is aggregated within each domain). The centre line is the median, the box limits are the upper and lower quartiles, and the whiskers represents the $1.5\times$ interquartile range.

6.2 Determinants Of Semantic Similarity

In this section we examine the effect of such features on the measured correlations. A combination of lexical (frequency, concreteness, rate of change) and environmental (cultural and geographic distance) features was selected. See §3.

Correlation With Lexical Features. At the word-level there is no correlation with respect to frequency and concreteness, and weak-moderate negative correlation with rate of lexical change. When aggregating over domains concreteness is still not correlated with any of the alignment methods; however, the correlation goes up for frequency (albeit still weakly for SNC), and interestingly a substantial correlation for NO ($r = 0.6$). and jumps for rate of change ($r \approx -0.6$ for SNC). This interesting result means that words that undergo faster lexical change are less aligned across languages. This echoes the findings that polysemy has an important role in the rate of lexical change (Brown and Witkowski, 1983; Thompson et al., 2020), and coincides with the findings that rate of change is correlated negatively with prototypicality (how representative a word is of its category) (Dubossarsky et al., 2017). Our experiments demonstrate that **polysemy** also plays a role in lexical alignment, as further elaborated in Appendix F.

Correlation With Environmental Features. The question of how **geographical** and **cultural** factors influence the alignment of words across languages is a matter of ongoing discussion among scholars (Youn et al., 2016; Josserand et al., 2021, e.g.,). Table 3 shows a significant correlation with geographic and cultural distance for SNC, with cultural distance playing a more prominent role.

[20]This trend persists for all SNC methods, across all model architectures, and various k values.

		SNC-CLOUD	SNC-AVE	SNC-STATIC	NO
CLT	C	0.14*	0.1*	0.25*	-0.08
	D	0.2*	0.49*	0.13*	0.11*
GEO	C	0.03*	0.09*	0.22*	-0.05*
	D	0.16*	0.41*	0.05	0.1*
frequency	C	0.04*	0.06	0.06	0.1*
	D	0.33*	0.18*	0	0.6*
concreteness	C	0.03	0	0	0
	D	0.18*	0.06	0.1*	0.1
rate-change	C	-0.32*	-0.22*	-0.25*	-0.14*
	D	-0.57*	-0.62*	-0.62*	-0.42*

Table 3: Correlation with lexical and enviromental features. Columns represent the features (CLT denotes cultural distance and GEO denotes geographical distance) and subcolumns represents concept-level aggregation (C) vs. domain-level aggregation (D). NO represents Neighbors Overlap metric. significant correlation with $p < 0.05$ are marked by $\star$.

7 Qualitative Analysis

We conduct a small scale qualitative analysis to further understand the results.[21]

Drawing conclusions at the word level solely from the raw data is challenging, as evidenced by the discrepancies between methodologies in our empirical analysis. We give a few examples of factors that impact the alignment.[22] The first is cultural differences. To illustrate, the word "pig" is less aligned between English and Arabic than between English and German. Manual examination of the data reveals that the nearest neighbors of "pig" in English and German include animals and food items, such as "butter" and "salt," whereas in Arabic, they are exclusively animals. This difference might arise from the cultural context in Islam, where pork is prohibited. Another reason for divergence is illustrated by the word "soft," which is one of the least aligned words across languages. The neighbors of "soft" are highly varied and come from different semantic domains, even within the same language, making their connection to the word itself less immediate. For example, in English, some neighbors for the word "soft" are "red", "sea" and "hand," while in German, they include "meat," "beautiful" and "rich". This diversity in semantic association contributes to the lower alignment for words like "soft".

[21]The qualitative analysis was conducted by one of the authors. The full details of the setup are in Appendix §E.

[22]Analysis done for SNC-CLOUD.

Differences in word senses also contribute to misalignment. For example, the word "brother" is less aligned between English and Hebrew, which might be attributed to the homonymy in Hebrew between the senses of "brother" and "hearth". This difference is reflected in their nearest neighbors. This issue is directly related to the influence of polysemy on alignment, which we address in Appendix §F and leave to future work.

8 Discussion

How and why languages vary in their use of words to carve up the semantic space has long been a question of central interest in cognitive sciences and linguistics, but has often been overlooked in NLP. Recent advances in NLP allow addressing this gap, and performing such analyses well and at scale. In this paper, we formulate this question as an NLP task, and provide a methodology for computing the efficacy of different metrics in addressing the task. We use existing metrics and extend them to contextualized word representations. We evaluate the metrics across multiple scenarios, using both synthetic and naturalistic validation approaches. We observe consistent trends across all metrics and architectures: the rate of change is a strong predictor of alignability. Additionally, internally structured domains such as Time, Quantity, and Kinship show the highest degree of alignment across languages, consistent with (Thompson et al., 2020).

One of the major challenges in analyzing cross-lingual alignment for individual words or domains is the lack of ground-truth data for validation. This impedes the comparison between different metrics, and thereby hinder progress on this task. To address this, we provide both synthetic and naturalistic validations using a newly created linguistic resource, based on the kinship domain. Our validation shows that all the SNC metrics we propose effectively capture cross-lingual semantic variability, with a slight preference for the metrics that are based on contextualized representations.

In future work, we plan to leverage cross-lingual alignment to investigate its impact on cross-lingual transfer on various NLP tasks. Additionally, we aim to use this approach to examine the extent to which LLMs encode cultural knowledge, a topic that has recently garnered significant attention (Havaldar et al., 2023; Li et al., 2024; Rao et al., 2024; Zhou et al., 2024).

Limitations

While we introduce a novel validation for the task of cross-linguistic lexicon alignment, it is currently limited to the kinship domain. A more comprehensive validation spanning multiple domains would provide a more thorough verification of the metrics. Additionally, we use the NEL dataset to analyze 1,016 concepts across various languages. In future work, we plan to expand this list to include more words and languages, enhancing the robustness of our analysis.

Moreover, throughout our study, we use the semantic domains defined in the NEL dataset. Although these manually defined domains are widely used in cognitive science and linguistic research, manual examination has shown that they are not optimal, due to noise and what seems to be inconsistent rationale as to what to include and what not. They could benefit from additional filtration and refinement to improve their accuracy and relevance.

Our work is readily applicable to both static word representations and contextualized representations; however, it is not currently suited for autoregressive models such as GPT and its variants. In future work, we aim to expand our metrics and evaluation to accommodate these architectures as well.

References

David Allen and Kathy Conklin. 2013. Cross-linguistic similarity norms for japanese-english translation equivalents. *Behavior research methods*, 46.

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.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*.

Ofir Arviv, Dmitry Nikolaev, Taelin Karidi, and Omri Abend. 2023. Improving cross-lingual transfer through subtree-aware word reordering. *arXiv preprint arXiv:2310.13583*.

George Barnett. 1977. Bilingual semantic organizationa multidimensional analysis. *Journal of Cross-cultural Psychology*, 8:315–330.

Lisa Beinborn and Rochelle Choenni. 2020. Semantic drift in multilingual representations. *Computational Linguistics*, 46:1–34.

Luisa Bentivogli and Emanuele Pianta. 2023. Looking for lexical gaps.

Brent Berlin and Paul Kay. 1991. *Basic color terms: Their universality and evolution*. University of California Press.

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.

Cecil Brown and Stanley Witkowski. 1983. Polysemy, lexical change and cultural importance. *Man*, 18:72.

Allan Burns. 1994. Review of John A. Lucy, grammatical categories and cognition: A case study of the linguistic relativity hypothesis. *Language in Society - LANG SOC*, 23:445–448.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *arXiv preprint arXiv:2303.17466*.

J. C. Catford. 1978. *A Linguistic Theory of Translation: An Essay in Applied Linguistics*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *The International Conference on Learning Representations (ICLR)*.

William Croft. 2010. Relativity, linguistic variation and language universals. *CogniTextes*, 4.

William Croft. 2022. On two mathematical representations for "semantic maps". *Zeitschrift für Sprachwissenschaft*, 41.

Johannes Dellert, Thora Daneyko, Alla Münch, Alina Ladygina, Armin Buch, Natalie Clarius, Ilja Grigorjew, Mohamed Balabel, Hizniye Isabella Boga, Zalina Baysarova, et al. 2020. Northeuralex: A wide-coverage lexical database of northern eurasia. *Language resources and evaluation*, 54:273–301.

Johannes Dellert, Thora Daneyko, Alla Münch, Alina Ladygina, Armin Buch, Natalie Clarius, Ilja Grigorjew, Mohamed Balabel, Hizniye Boga, Zalina Baysarova, Roland Mühlenbernd, Johannes Wahle, and Gerhard Jäger. 2019. NorthEuraLex: a wide-coverage lexical database of Northern Eurasia. *Language Resources and Evaluation*, 54:1–29.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145, Copenhagen, Denmark. Association for Computational Linguistics.

Jon Andoni Duñabeitia, Davide Crepaldi, Antje Meyer, Boris New, Christos Pliatsikas, Eva Smolka, and Marc Brysbaert. 2018. Multipic: A standardized set of 750 drawings with norms for six European languages. *The Quarterly Journal of Experimental Psychology*, 71:808–816.

Juuso Eronen, Michal Ptaszynski, and Fumito Masui. 2023. Zero-shot cross-lingual transfer language selection using linguistic similarity. *Information Processing & Management*, 60(3):103250.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. pages 55–65.

Jerry Fodor. 1975. *The Language of Thought*. Harvard University Press.

William Frawley. 1998. Review of Anna Wierzbicka, Semantics: primes and universals. *Journal of Linguistics*, 34:227–297.

Aina Garí Soler and Marianna Apidianaki. 2021a. Let's play mono-poly: BERT can reveal words' polysemy level and partitionability into senses. *Transactions of the Association for Computational Linguistics*, 9:825–844.

Aina Garí Soler and Marianna Apidianaki. 2021b. Let's play mono-poly: Bert can reveal words' polysemy level and partitionability into senses. *Transactions of the Association for Computational Linguistics*, 9:825–844.

Thanasis Georgakopoulos, Eitan Grossman, Dmitry Nikolaev, and Stéphane Polis. 2022. Universal and macro-areal patterns in the lexicon: A case-study in the perception-cognition domain. *Linguistic Typology*, 26:439–487.

Wilhelm Glaser. 1992. Picture naming. *Cognition*, 42:61–105.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the conference on empirical methods in natural language processing. Conference on empirical methods in natural language processing*, volume 2016, page 2116. NIH Public Access.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Martin Haspelmath. 2003. The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In Michael Tomasello, editor, *The new psychology of language*, pages 217–248. Erlbaum.

Shreya Havaldar, Sunny Rai, Bhumika Singhal, Langchen Liu Sharath Chandra Guntuku, and Lyle Ungar. 2023. Multilingual language models are not multicultural: A case study in emotion. *arXiv preprint arXiv:2307.01370*.

T. Hermans. 1996. Norms and the determination of translation: a theoretical framework. *Hermans, T. (1996) Norms and the determination of translation: a theoretical framework. In: Alvarez, R. and Vidal, M., (eds.) Translation, Power, Subversion. Topics in Translation . Multilingual Matters, Clevedon, England, pp. 25-51. ISBN 1853593508*.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, et al. 2022. Challenges and strategies in cross-cultural nlp. *arXiv preprint arXiv:2203.10020*.

Joshua Jackson, Joseph Watts, Teague Henry, Johann-Mattis List, Robert Forkel, Peter Mucha, Simon Greenhill, Russell Gray, and Kristen Lindquist. 2019. Emotion semantics show both cultural variation and universal structure. *Science*, 366.

Mathilde Josserand, Emma Meeussen, Asifa Majid, and Dan Dediu. 2021. Environment and culture shape both the colour lexicon and the genetics of colour perception. *Scientific Reports*, 11:19095.

Hadi Khalilia, Gábor Bella, Abed Alhakim Freihat, Shandy Darma, and Fausto Giunchiglia. 2023. Lexical diversity in kinship across languages and dialects. *Frontiers in Psychology*, 14:1229697.

Temuulen Khishigsuren, Gábor Bella, Khuyagbaatar Batsuren, Abed Freihat, Nandu Nair, Amarsanaa Ganbold, Hadi Khalilia, Yamini Chandrashekar, and Fausto Giunchiglia. 2023. Using linguistic typology to enrich multilingual lexicons: the case of lexical gaps in kinship.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models. *arXiv preprint arXiv:2402.10946*.

Asifa Majid, James Boster, and Melissa Bowerman. 2008. The cross-linguistic categorization of everyday events: A study of cutting and breaking. *Cognition*, 109:235–250.

Asifa Majid, Fiona Jordan, and Michael Dunn. 2014. Semantic systems in closely related languages. *Language Sciences*, 49.

Raja Marjieh, Pol van Rijn, Ilia Sucholutsky, Theodore Sumers, Harin Lee, Thomas Griffiths, and Nori Jacoby. 2022. Words are all you need? capturing human sensory similarity with textual descriptors. *arXiv preprint arXiv:2206.04105*.

Cornelia Moldovan, Rosa Sanchez-Casas, Josep Demestre, and Pilar Ferré. 2012. Interference effects as a function of semantic similarity in the translation recognition task in bilinguals of catalan and spanish. *PSICOLOGICA*, 33:77–110.

George Murdock. 1970. Kin term patterns and their distribution. *Ethnology*, 9.

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225.

Dmitry Nikolaev, Ofir Arviv, Taelin Karidi, Neta Kenneth, Veronika Mitnik, Lilja Maria Saeboe, and Omri Abend. 2020. Fine-grained analysis of cross-linguistic syntactic divergences. *arXiv preprint arXiv:2005.03436*.

Mark Pagel, Quentin Atkinson, and Andrew Meade. 2007. Frequency of word-use predicts rates of lexical evolution throughout indo-european history. *Nature*, 449:717–20.

Ross Purves, Philipp Striedl, Inhye Kong, and Asifa Majid. 2023. Conceptualizing landscapes through language: The role of native language and expertise in the representation of waterbody related terms. *Topics in cognitive science*, 15.

Xiaoying Qi. 2017. Reconstructing the concept of face in cultural sociology: in goffman's footsteps, following the chinese case. *The Journal of Chinese Sociology*, 4(1):19.

Ella Rabinovich, Yang Xu, and Suzanne Stevenson. 2020. The typology of polysemy: A multilingual distributional framework. *(Annual Meeting of the Cognitive Science Society (CogSci)*.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2024. Normad: A benchmark for measuring the cultural adaptability of large language models. *arXiv preprint arXiv:2404.12464*.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Christoph Rzymski, Tiago Tresoldi, Simon Greenhill, Mei-Shin Wu, Nathanael Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, Timotheus Bodt, Abbie Hantgan, Gereon Kaiping, Sophie Chang, Yunfan Lai, Natalia Morozova, Heini Arjava, Nataliia Hübler, Ezequiel Koile, Steve Pepper, Mariann Proos, Briana Epps, and Johann-Mattis List. 2020. The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific Data*, 7.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. *arXiv preprint arXiv:1902.09492*.

Satoshi Shioiri, Rumi Tokunaga, and Ichiro Kuriki. 2023. Cross-cultural comparison of lexical partitioning of color space. In *Issues in Japanese Psycholinguistics from Comparative Perspectives: Volume 1: Cross-Linguistic Studies*, pages 41–61. de Gruyter.

Jesse Snedeker and Lila Gleitman. 2004. *Weaving a Lexicon*. MIT Press.

Mahesh Srinivasan and Hugh Rabagliati. 2015. How concepts and conventions structure the lexicon: Cross-linguistic evidence from polysemy. *Lingua*, 157:124–152.

B. Thompson, S. G. Roberts, and G Lupyan. 2018. Quantifying semantic similarity across languages. *(Annual Meeting of the Cognitive Science Society (CogSci)*.

Bill Thompson, Seán Roberts, and Gary Lupyan. 2020. Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour*, 4:1–10.

Annika Tjuka, Robert Forkel, and Johann-Mattis List. 2024. Universal and cultural factors shape body part vocabularies. *Scientific Reports*, 14(1):10486.

Natasha Tokowicz, Judith Kroll, Annette Groot, and Janet van Hell. 2002. Number-of-translation norms for dutch–english translation pairs: A new tool for examining language production. *Behavior research methods, instruments, & computers : a journal of the Psychonomic Society, Inc*, 34:435–51.

Mor Ventura, Eyal Ben-David, Anna Korhonen, and Roi Reichart. 2023. Navigating cultural chasms: Exploring and unlocking the cultural pov of text-to-image models. *arXiv preprint arXiv:2310.01929*.

Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. 2021. LexFit: Lexical fine-tuning of pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5269–5283, Online. Association for Computational Linguistics.

Ivan Vulić, Edoardo Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics.

Benjamin Lee Whorf. 1956. *Thought and Reality: Selected Writing*, first edition. MIT Press.

Anna Wierzbicka. 1972. Semantic primitives. *Frankfurter anthropologische Blätter*, 11:1–16.

Anna Wierzbicka. 2008. Why there are no 'colour universals' in language and thought. *Journal of the Royal Anthropological Institute,*, pages 407–425.

Hyejin Youn, Logan Sutton, Eric Smith, Cristopher Moore, Jon Wilkins, Ian Maddieson, William Croft, and Tanmoy Bhattacharya. 2016. On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences*, 113.

Li Zhou, Taelin Karidi, Nicolas Garneau, Yong Cao, Wanlong Liu, Wenyu Chen, and Daniel Hershcovich. 2024. Does mapo tofu contain coffee? probing llms for food-related cultural knowledge. *arXiv preprint arXiv:2404.06833*.

A Experimental Setup

In this section we provide further details regarding our experimental setup.

Languages. We perform our analysis on a diverse set of 16 languages, spanning 7 different top-level language families from many geographical areas across Eurasia: English (eng), French (fra), Italian (ita), German (deu), Dutch (nld), Spanish (spa), Polish (pol), Finnish (fin), Estonian (est), Turkish (tur), Chinese (chn), Korean (kor), Japanese (jap), Hebrew (heb), Hindi (hin) and Arabic (arb).

NorthEuraLex (**NEL**) is a lexical resource compiled from dictionaries and other linguistic resources available for individual languages in Northern Eurasia (Dellert et al., 2020). NEL comprises a list of 1016 distinct *concepts*[23] together with their word forms in 107 languages (Table 1). Rare cases where a concept does not have any realization in a given language are excluded for that language.

Semantic Domains. We map the concepts in NEL to domains, using Concepticon.[24] There are 20 domains, each containing $22 - 136$ concepts (number of concepts is written next to each domain): animals (47), agriculture and vegetation (23), time (68), quantity (40), kinship (26), basic actions and technology (140), clothing and grooming (27), cognition (30), emotions and values (54), food and drink (42), modern world (28), motion (70), possession (26), sense perception (50), social and political relations (30), spatial relations (85), speech and language (25), the body (94), the house (20) and the physical world (75).

[23]The concept in NEL are given in German.

[24]https://concepticon.clld.org/

Word Embeddings. In the main paper, for static word embeddings we use fastText[25] 300-dimension word embeddings, trained on Wikipedia using the skip-gram model (Bojanowski et al., 2017). For contextualised word embeddings (CWE) we use mBERT[26] (*bert-base-multilingual-uncased* model) 768-dimension vectors for the 16 languages. To extract sentences for SNC-CLOUD, we use the Leipzig corpus.[27] Due to lack of space we choose to focus on this setup after experimentation with other models architectures as trends are consistent across the board. We also run our experiments also on XLM-RoBERTa-base [28] for SNC-CLOUD and SNC-AVE and on 300-dim word2vec multilingual embeddings [29] for SNC-STATIC. Furthermore, we perform all computations for SNC-CLOUD and SNC-AVE with a different dataset; the Wikipedia section in the Leipzig Corpus, for the latest year available in each language [30]. The trends were highly similar to what we report in the paper.[31].

Hyperparameters. For our distributional based alignments SNC and NO (§4.2), we set $k = 100$. We experimented with other values of k ($k = 10, 50, 1000$) and selected the one that overall correlated the most with our validation and showed more robust results in terms of correlation with other features (such as lexical ans enviromental features). This is also the hyperparameter chosen in the original work of (Thompson et al., 2020) for the SNC-STATIC methods, based on similar reasons.

Lexical and Language Features. We report results while controlling for a variety of lexical features and features of the languages compared. Geographic distance between languages is computed as the geodesic distance (distance in an ellipsoid) between their latitude and longitude coordinates (taken from Glottolog[32]). Cultural distance is computed as the proportion of common cultural traits from a set of 92 non-linguistic cultural traits for 16 societies representing the languages in our analysis,

[25]https://fasttext.cc/docs/en/unsupervised-tutorial.html

[26]https://huggingface.co/bert-base-multilingual-uncased

[27]https://corpora.uni-leipzig.de/en?corpusId=deu_news_2021

[28]https://huggingface.co/xlm-roberta-base

[29]https://github.com/Kyubyong/wordvectors

[30]https://wortschatz.uni-leipzig.de/en

[31]See Appendix G for experiments on other architectures than the ones presented in the main paper.

[32]https://glottolog.org/

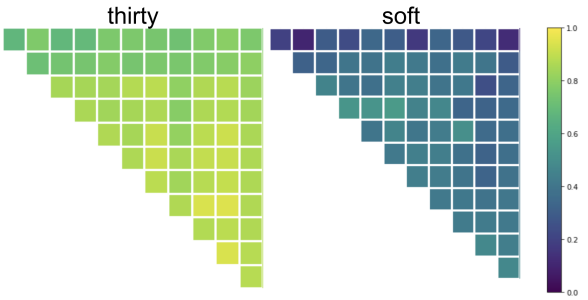

Figure 5: Alignment of individual concepts under SNC-CLOUD. Left: alignment for the concept "thirty". Right: alignment for the concept "soft". Computed for a sample of 12 languages. Each square represent alignment for a language pair. More aligned pairs are closer to 1. The alignment scores are computed for: Turkish, Hebrew, German, Finnish, English, Dutch, Spanish, Russian, French, Hindi, Italian, and Arabic.

taken from D-PLACE[33] (Thompson et al., 2020). We use the *wordfreq* library[34] for word frequencies. We then compute the log-transformed frequency (to reduce the impact of outliers and extreme values).

Rate of Lexical Change. Realizations of some concepts, such as *tail*, evolve rapidly, while others, such as *two* evolve at a much slower rate. This phenomenon is referred to as the *rate of (lexical) change*. We use lexical change rates derived from (Pagel et al., 2007), available for Russian, Greek, English and Spanish.

B Word-level and Domain-level Analysis

We hereby describe in full details how we perform the analysis at the word-level and the domain-level. Let $\mathcal{M}$ be the set of alignment metrics. We denote the raw data as follows:

$$\mu(m, L_p, L_j) \quad \forall\, m \in \mathcal{M}, L_p \times L_j \in \Omega^2$$

For a pair of languages L_p, L_j and a metric m, $\mu(m, L_p, L_j) \in \mathbb{R}^{|C|}$ is a vector whose i-th coordinate is the alignment value of concept c_i under metric m between L_p and L_j.

Throughout the following section we use Pearson's r (with a two-tailed test for significance) for computing correlation, unless stated otherwise.

B.1 Word-level Correlations.

The most straightforward level of comparison between metrics is their word-level correlation. Let $\binom{\Omega}{2}$ be the set of all language pairs (without repetitions), and denote its size with $l=\binom{|\Omega|}{2}$. For $m \in \mathcal{M}$, define $\hat{\mu}(m) \in \mathbb{R}^{l|C|}$ the concatenation of $\mu(m, L_p, L_j)$ for all language pairs. Word-level correlation is the Pearson correlation between $\hat{\mu}(m)$, for $m \in \mathcal{M}$ (See Figure 7 and Figure 7).

[33] https://d-place.org/

[34] https://pypi.org/project/wordfreq

	SNC-STATIC	SNC-STATIC	SNC-CLOUD
Top 3	Quantity Time Kinship	Quantity Time Kinship	Quantity Kinship Time
Bottom 3	Possession Basic Actions Motion	Basic actions Motion The house	Agriculture Spatial relations The house

Table 4: Most and least aligned domains for various metrics. Alignment computed by aggregating over languages and over domains. "Basic actions." refers to "Basic actions and technology" and "Agriculture" refers to "Agriculture and vegetation"

B.2 Domain-level Correlations.

Alignment metrics between languages are often used to compare the degree of alignment across different domains. For example, Thompson et al. (2020) argue, based on findings with SNC-STATIC , that more structured domains tend to be better aligned across languages. To examine the alignment at the domain level, for every measure $m \in \mathcal{M}$, we aggregate the word=level alignment over each domain (without aggregating over languages). We get $\hat{\mu}(m) \in \mathbb{R}^{lm}$ (m is the number of semantic domains).

C Controlling for Lexical and Enviromental Features.

To further examine the influence of lexical and environmental features on the alignment methods, we perform partial correlation tests to control for the various features, and multiple regression analysis to account for the overall variance that is explained by them. We compute the partial correlation[35] between SNC methods, while controlling for the lexical and environmental features.

We find that at the word-level measures are still moderately correlated with $r \approx 0.4$. At the domain-level, the methods are still highly correlated with one another ($r \approx 0.9$). We use multiple linear regression to compute the adjusted R-

[35] For the partial correlation computations we use the *pingouin* package https://pingouin-stats.org/build/html/generated/pingouin.partial_corr.html

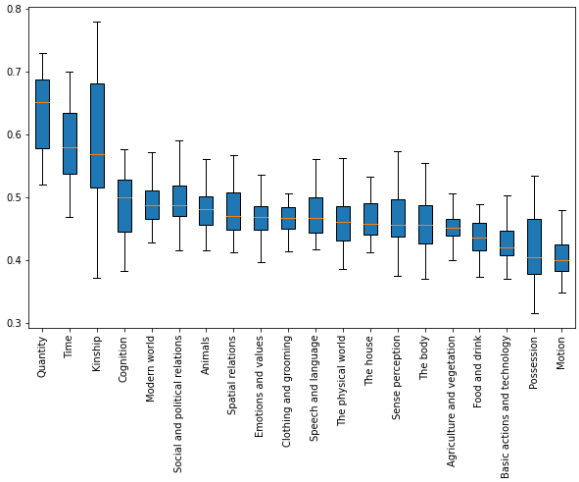

Figure 6: Alignment of domains under SNC-CLOUD. The domains are ranked according to the mean value of the alignment. Each box represents the distribution of alignment values (per language pair), for a specific domain (concepts-level alignment is aggregated within each domain). The centre line is the median, the box limits are the upper and lower quartiles, and the whiskers represents the $1.5\times$ interquartile range. Alignment computed using XLM-RoBERTa-base.

squared value, with the environmental and lexical features as response variables. While the features explain $\approx$ 20% of the variance for SNC. However, when aggregating over domains, the features explain up to 44% of the variance for SNC. This suggests that the analysis is more suitable at the domain-level.

D Comparing SNC to Norm-based Approaches

In order to capture cross-lingual lexical similarity, cognitive scientists have used behavioral stimuli, e.g., sets of pictures that are named by speakers of different languages (Thompson et al., 2020). Where the same pictures are named consistently in two languages with a given pair of words, these words are interpreted as semantically similar (Glaser, 1992; Marjieh et al., 2022). Another common paradigm is the use of translation norms (Hermans, 1996). In this section we discuss two such datasets, and empirically compare how well they align with SNC. We note that these measures do not consider as "ground truth" but rather as converging evidence for the validity of the metrics, as it is not trivial that they measure the same (or even higly similar) things. We use two manual external resources: name-agreement in picture naming (Multipic; Duñabeitia et al., 2018) and English-Dutch translation norms (TransSim; Tokowicz et al., 2002). MultiPic is a standardized set of 750 drawings of concrete objects with name agreement norms for six European languages (English, Spanish, Netherlands Dutch, German, French and Italian). For each picture and language, the norm is an information statistic that reflects the level of agreement across participants. TransSim is a dataset of 562 Dutch-English translation pairs together with a human similarity rating between each pair.

Multipic. We filter the pictures in the Multipic dataset to only include pictures with concepts from NEL. This results in a total of 194 pictures. We compute the correlation between the agreement scores (average agreement score over all languages) for these pictures and the different SNC. We get that while SNC-AVE and SNC-STATIC are moderately correlated with Multipic ($r \approx 0.3$, $p < 0.05$), the other methods are weakly to not correlated with the dataset.

TransSim. We again filter the dataset to include word pairs that are covered by NEL, resulting in 187 Dutch–English translation similarity judgment scores. We compute the correlation between English-Dutch translation similarity judgements and the alignment metrics for English-Dutch, aggregated by domain (domain-level). A relatively high correlation is presented, where SNC-STATIC ($r = 0.59$, $p < 0.05$) and SNC-AVE ($r = 0.51$, $p < 0.05$) rank highest.

To conclude, we find that overall SNC are moderately correlated with the human-evaluated datasets, which may reflect a relative similarity in the notion of alignment captured by these datasets and the distributional metrics. We defer a more elaborate multi-approach comparison to future work.

E Qualitative Analysis

We conduct a small scale qualitative analysis on four language pairs (English-German, English-Arabic, English-Russian, and English-Hebrew).[36] Additionally, we examine word-level results averaged over all languages. Specifically, for each SNC method and language pair, we selected the top and bottom 100 aligned words along with their 10

[36]The qualitative analysis was conducted by one of the authors.

		SNC-CLOUD	SNC-AVE	SNC-STATIC	NO
Self-Sim	C	$-0.1^\star$	$-0.3^\star$	$-0.42^\star$	0
	D	$-0.41^\star$	$-0.36^\star$	$-0.35^\star$	-0.37*
GMM-Senses	C	$0.16^\star$	$0.22^\star$	$0.31^\star$	$0.21^\star$
	D	$0.58^\star$	$0.47^\star$	$0.52^\star$	$0.7^\star$

Table 5: Correlation with Polysemy Measures. Columns represent the two polysemy measures: Self-Similarity (denoted by **Self-Sim**) and the average number of gmm clusters (denoted by **GMM-Senses**). word-level correlations are denoted by **C** and domain-level correlation by **D**. Neighbors Overlap metric is represented by NO. significant correlation with $p < 0.05$ are marked by $\star$.

nearest neighbors in each language. The analysis appears in Section §7.

F The Impact of Polysemy on Cross-Lingual Alignment

Polysemy is a linguistic phenomenon where a word has multiple related meanings. For example, the word "bank" can refer to a financial institution or the side of a river. In linguistics, much research had been conducted on the relation between polysemy and other lexical phenomenon such as rate of change and frequency, including in NLP. We can differentiate between two types of polysemy: one where a word has multiple related senses, such as "book" (a physical object or an act of reserving), and another where the senses are distinct, such as "bark" (the sound a dog makes or the outer covering of a tree).

To manually quantify the polysemy of a word, one approach is to count the number of entries it has in a dictionary. However, this method is not easily applicable to many languages. Alternatively, BabelNet (Navigli and Ponzetto, 2010) can be used to identify word senses. Yet, this introduces challenges, as BabelNet often includes many highly similar senses (that we do not want to take into account), adding noise to our alignment process. While we intend to address these challenges in future work, space constraints and the specific focus of our paper have led us to adopt alternative measures of polysemy for this preliminary inquiry, based on the word representations themselves. Results are presented in Table 5.

F.1 Self-Similarity

The first measure that we consider is Self-Similarity, introduced by (Ethayarajh, 2019). This measure have shown to highly correlate with polysemy and was used as an alternative measure for the degree of polysemy of words (Garí Soler and Apidianaki, 2021b). For a word w, that is, the average of the pairwise cosine similarities of the representations of its contextualised representations in corpus $\mathcal{A}$. Defined as:

$$SelfSim = \frac{1}{|\mathcal{A}^2|-|\mathcal{A}|}\Sigma_{i\in\mathcal{A}}\Sigma_{j\in\mathcal{A},j\neq i}cos(x_{wi}, x_{wj}), \quad (1)$$

where x_{w_i} is the contextualised representation of word w, taken from its i-th instance in corpus $\mathcal{A}$ (here, $|\mathcal{A}|\leq 1000$). For a pair of languages L_1 and L_2 we calculate the average Self-Similarity score within each language. This average serves as the alignment measure based on Self-Similarity for the language pair.

Table 5 shows that at the word-level and the domain-level Self-Similarity is significantly (negatively) weakly-moderately correlated with SNC methods, reaching $r = -0.42$ for SNC-STATIC. The negative correlations means that the more polysemous the word is, the less it is aligned across languages.

F.2 Sense Clusters

We define another measure for polysemy, that is based on clusters of the word embeddings. In the computation process of SNC-CLOUD (Section §4.2), for a word w we have ≤ 1000 contextualised representations, based on an external corpus (Appendix §A), we denote this point-cloud by $\mathcal{O}$. We perform Gaussian-Mixture Models (GMM) clustering on $\mathcal{O}$ and choose the optimal number of clusters using the Elbow Method.[37]

We define the **degree of polysemy** for a word w to be the number of clusters we calculated. For a pair of languages L_1 and L_2 we calculate the average degree of polysemy score within each language. This average serves as the alignment measure based on the degree if polysemy for the language pair.

Table 5 shows (the degree of polysemy is denoted by **GMM-sense**) that the methods are significantly correlated with this measure, at the domain level reaching $r = 0.7$ for NO and $r = 0.58$ for SNC-CLOUD.

Intuitively, as discussed in the main paper, different patterns of polysemy can result in varied

[37]To ensure robustness, we repeat this process 10 times and select the number of clusters that appears most frequently across these iterations.

		SNC-CLOUD	SNC-AVE	SNC-STATIC	NO
CLT	C	0.1*	0.08	0.27*	0
	D	0.23*	0.31*	0.11*	0.08*
GEO	C	0.1	0.08*	0.15*	-0.01
	D	0.2*	0.39*	0.1*	0.17*
frequency	C	0	-0.04	0.01	0
	D	0.35*	0.15*	0	0.58*
concreteness	C	0	0	0	0.1
	D	0.15*	0.1*	0.15*	0.08
rate-change	C	-0.25*	-0.27*	-0.3*	-0.11
	D	-0.55*	-0.48*	-0.65*	-0.39*

Table 6: Correlation with lexical and enviromental features (other architectures). Columns represent the features (CLT denotes cultural distance and GEO denotes geographical distance) and subcolumns represents concept-level aggregation (C) vs. domain-level aggregation (D). NO represents Neighbors Overlap metric. significant correlation with $p < 0.05$ are marked by $\star$.

nearest neighbors across languages. For instance, in English, the word *bank* might have a mix of neighbors related to both *river bank* and *financial institution*. However, in Hebrew, where the sense of a river bank does not exist, neighbors are likely all related to the financial institution sense of *bank*. This mismatch can lead to greater divergence in alignment between English and Hebrew.

Although this measure has been shown to reflect the degree of polysemy of words (Garí Soler and Apidianaki, 2021a), the correlations we compute still require further control.

G Other Architectures.

We also run our experiments on XLM-RoBERTa-base [38] for SNC-CLOUD and SNC-AVE and on 300-dim word2vec multilingual embeddings [39] for SNC-STATIC. Moreover, we run all of the computations for SNC-CLOUD and SNC-AVE with a different dataset; the Wikipedia section in the Leipzig Corpus, for the latest year available in each language [40]. The trends were highly similar to what we report in the main paper and are presented in Table 6, Figure 6 and Figure 7.

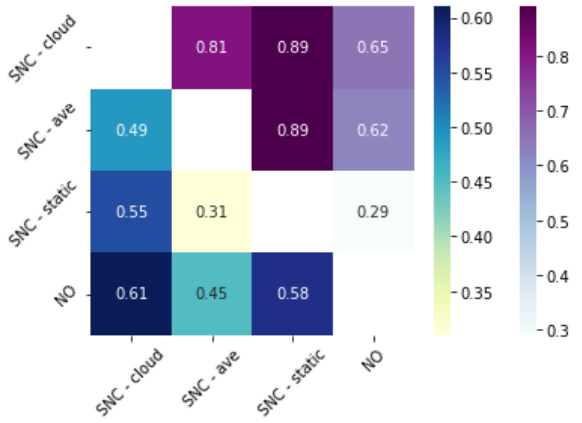

Figure 7: Correlation between the various metrics (other architectures). Pearson correlation is computed for different aggregation methods. The **upper** matrix represents concept-level correlations, while the **bottom** matrix represents domain-level correlations. All correlation values are significant with $p < 0.05$.

[38] https://huggingface.co/xlm-roberta-base
[39] https://github.com/Kyubyong/wordvectors
[40] https://wortschatz.uni-leipzig.de/en