Aligning Alignments: Do Colexification and Distributional Similarity Align as Measures of cross-lingual Lexical Alignment?

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

Abstract

The data-driven investigation of the extent to which lexicons of different languages align has mostly fallen into one of two categories: colexification-based and distributional. The two approaches are grounded in distinct methodologies, operate on different assumptions, and are used in diverse ways. This raises two important questions: (a) are there settings in which the predictions of the two approaches can be directly compared? and if so, (b) what is the extent of the similarity and what are its determinants? We offer novel operationalizations for the two approaches in a manner that allows for their direct comparison, and conduct a comprehensive analysis on a diverse set of 16 languages.

Our analysis is carried out at different levels of granularity. At the word-level, the two methods present different results across the board. However, intriguingly, at the level of semantic domains (e.g., kinship, quantity), the two methods show considerable convergence in their predictions. Our findings also indicate that the distributional methods likely capture a more fine-grained alignment than their counterpart colexification-based methods, and may thus be more suited for settings where fewer languages are evaluated.[1]

1 Introduction

To what degree do translation equivalents in different languages – for example, English *red* and French *rouge* – encode the same meaning? This question, in various forms, has long been a topic of interest in the cognitive sciences (Whorf, 1956; Fodor, 1975; Frawley, 1998; Burns, 1994; Snedeker and Gleitman, 2004; Majid et al., 2008; Croft, 2010). Indeed, lexicons are often viewed as reflecting the structure of human cognition; understanding how meaning is expressed across languages helps understand how humans categorize and represent the world.

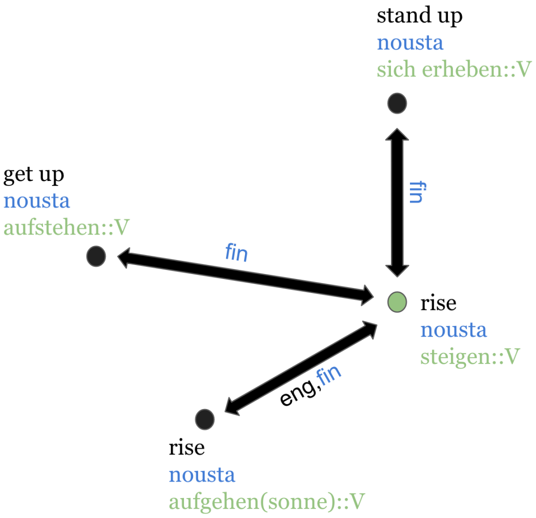

Figure 1: Colexification graph for the target concept “steigen::V” (which corresponds to the word *rise* in English, and *nousta* in Finnish. Each vertex corresponds to a concept that is colexified with the target concept either in English or Finnish. The English lexicalizations of the target concept are in black and the Finnish lexicalizations are in Blue. The concepts themeselves are in Green. Each edge (marked by an arrow) denotes that a colexification exists in English/Finnish (as labeled).

A building block in answering such a question is the ability to evaluate the similarity between words that seemingly express a similar meaning (henceforth, *translation pairs*) across different languages.

Traditionally, in linguistic and cognitive research, comparing the meaning of words across languages involves methodologies and approaches that are less data-driven in nature, prioritizing in-depth, relatively small-scale exploration of meaning, such as descriptive comparisons (Karidi et al., 2024; 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).

[1] Our code and data is available at https://github.com/tai314159/Aligning_Alignments.

Proceedings of the 28th Conference on Computational Natural Language Learning, pages 327–341
November 15-16, 2024 ©2024 Association for Computational Linguistics

The difficulty in defining lexical similarity between concepts, let alone translation equivalents, has motivated a transition from theoretical frameworks to data-driven approaches. Indeed, a significant amount of recent works has focused on using data-driven methods to measure the equivalence of word pairs across different languages (Majid et al., 2014; Youn et al., 2016; Thompson et al., 2018; Jackson et al., 2019; Thompson et al., 2020; Rabinovich et al., 2020; Beinborn and Choenni, 2020; Georgakopoulos et al., 2022). All work on this question inherits an even more fundamental set of questions: how is meaning defined and how is the meaning of words captured? Within this rich body of work, we can identify two main methodological approaches.

The first approach is based on **colexification patterns**, which aims to compare the association between lexical form and senses across languages. Colexification is the case where two or more concepts are lexicalized with a single form in a given language (François, 2008; Rzymski et al., 2020) (see Table 2). For example, both English *right* and German *recht* colexify (i) a sense related to the correctness of a fact and (ii) a sense related to location or direction in space, while the Arabic *yamin* is associated with the spatial sense, but not the correctness sense. According to this approach, the degree to which words or sets of words in a certain domain in different languages align, can be defined as the degree to which the words colexify the same concepts. For example, the English *right* may be said to be more similar to the German *recht* than the Arabic *yamin* (cf. Haspelmath, 2003). Recently, a large-scale cross-lingual database of colexifications has been compiled (CLICS; Rzymski et al., 2020)[2]. This database provides a valuable resource for exploring the relationships between words and concepts across a wide range of languages, and enables the quantitative comparison of colexification patterns in different languages (Youn et al., 2016; Jackson et al., 2019; Xu et al., 2020; Georgakopoulos et al., 2022; Karjus et al., 2021a; Bao et al., 2021).

The second approach is based on **distributional word embeddings** (here, DISTA). This approach was recently proposed as a viable data-driven method for cross-lingual lexical semantic investigations (Thompson et al., 2018, 2020; Beinborn and Choenni, 2020; Rabinovich et al., 2020; Karidi et al., 2024), for improving cross-lingual transfer (Sun et al., 2021) and for investigating multicultural knowledge in LLMs (Havaldar et al., 2023). While all distributional methods use the word embeddings of translation pairs for computing similarity, many different operationalizations of this general approach are possible. See §2.1.

Both approaches have had a substantial impact on the computational cognitive science literature (Youn et al., 2016; Jackson et al., 2019; Thompson et al., 2020). These approaches seek to reveal an abstract structure that underlies the relation between words and their meanings (e.g, languages from different language families might have the same structure of kinship terms). However, while both are data-driven and aim to capture similar phenomena, they rely on different data and methodologies, and in fact likely capture different aspects of linguistic meaning. Colexification-based approaches set out to quantify similarity in lexicographical resources, while distributional embeddings use any signal that can be reliably extracted from the data. For example, DISTA may not represent rare senses, while colexification does not take frequency into account at all. They are also applied differently: colexification-based approaches often constructs intricate cross-lingual networks to explore meaning universality (Youn et al., 2016; Jackson et al., 2019), while distributional alignment methods operate at the word level and can then be extended to larger word sets (Thompson et al., 2020).

In this work we seek to empirically compare the predictions of these two approaches. However, given the divergence in methodologies and underlying assumptions adopted by these various approaches, it is not clear if it is sound, or even possible, to compare them. Moreover, obtaining a meaningful signal from colexification data typically requires aggregating information across thousands of languages (Youn et al., 2016; Jackson et al., 2019) and is rarely used for analysis at the word-pair level; instead, its strength lies in the analysis of intricate networks. Therefore, working with a substantially smaller set of languages or even comparing a single language pair at a time, as is often the case in multilingual NLP research, requires adapting the approaches so they will yield compa-

[2] Another valuable resource for lexical semantics is Babelnet (Navigli and Ponzetto, 2012). In this work we choose CLICS over BabelNet because BabelNet's fine-grained sense distinctions, such as separating "apple" as a fruit from "apple" as a tree, introduce excessive noise, whereas CLICS provides more manageable colexifications for our purposes.

rable predictions. We ask whether these distinct approaches converge at *interface settings* – settings in which the two approaches offer coherent similarity measures that can be compared. We show that such cases of convergence exist (§5) and, in these cases, ask whether – and when – the different approaches yield similar predictions. This is, to the best of our knowledge, the first time that these questions have been tackled within NLP.[3]

Analysis at various levels of granularity reveals that at the word-level, the two methodologies yield different results across the board. However, at the domain-level[4], the trends presented by the two methods show substantially higher correlation. In general, there is an overall greater similarity across different distributional methods than between the two families of approaches, in terms of their predictions and the factors that influence them (§5). Moreover, while distributional methods are correlated in their alignment predictions with external similarity measures (§5.4), the colexification approach is not. This suggests that the distributional approach captures more fine-grained aspects of meaning and is better suited for either delicate analysis of the results or when using a smaller set of languages. Also, the domain-level might be a more robust level to report alignment than the word-level. Additonally, we find that rate of lexical change is a significant predictor for cross-lingual alignment, across all methodologies. We discuss the implications of these results in §7.

To recap, we (i) operationalize distribution-based and colexification-based approaches so as to enable a direct empirical comparison between them, (ii) perform in-depth comparison of different operationalizations of the two approaches, (iii) study the ramifications of different design choices that they incorporate.

2 cross-lingual Lexicon Alignment

Concept	Languages
CLAW, FINGERNAIL	Japanese, Finnish, Estonian
SNOW, ICE	Hindi
DUST, ASH	German, French, Dutch Polish, Finnish
MONTH, MOON	Japanese, Korean Estonian, Turkish
DREAM, SLEEP (STATE)	Spanish, Polish, Finnish Italian
BABY, CHILD	French, Dutch, Hindi Polish
NEPHEW, NIECE	Italian

Figure 2: Colexifications. Examples of concepts from the CLICS dataset and their colexifications. Each colexification indicates the languages in which these concepts colexify, drawn from 16 languages used in this paper.

Much research on cross-lingual alignment between lexicons has sought to uncover whether certain concepts, notably in domains perceived as basic to the human experience, such as space, time, color, quantity, and family relations, are universal, on the one hand, or culturally- or historically-contingent, on the other hand (Fodor, 1975; Brown and Witkowski, 1983; Burns, 1994; Frawley, 1998; Evans and Levinson, 2009; Wierzbicka, 2010; Åke Viberg, 1983; Majid et al., 2014). Alignment can either be defined with respect to individual words (i.e, word-level alignment) or with respect to domains (i.e, domain-level alignment). For example, we might expect the word *Sunday* in English not to align well with the Hebrew multiword expression denoting the same day of the week *yom rishon*, as the latter does not bear any of the religious connotations of *Sunday* in English. The degree of their alignment is a **word-level alignment**. One can also compare the extent to which the concepts of time align more generally, in which case we might expect Hebrew and English to be relatively similar, given that Hebrew, spoken in Israel, *prima facie* has a Western conception of time, with, for example, a division of the year into twelve months, a division of the week into seven days, and so on. This is termed **domain-level alignment**.

2.1 Distribution-based Alignment

Distribution-based alignment measures leverage NLP tools to evaluate cross-lingual similarity (Artetxe et al., 2018; Conneau et al., 2017; Vulić et al., 2021; Rabinovich et al., 2020; Thompson et al., 2020; Karidi et al., 2024). Traditionally, these assessments have been performed using *global methods*, which align whole language spaces simultaneously and then assess their similarity using downstream tasks, such as Bilingual Lexicon Induction (Artetxe et al., 2018; Conneau et al., 2017).

[3] Recently, (Liu et al., 2023a) used co-occurrences to discover colexification patterns . However, their focus was primarily on reconstructing the colexifications from textual data, rather than analyzing colexification as a measure of cross-lingual semantic similarity and comparing it against methodologies that are based on word embeddings.

[4] A semantic domain is a way of grouping words together based on common aspects of meaning or function.

This approach typically includes techniques like linear transformations or joint model training across multiple languages (Pires et al., 2019; Gonen et al., 2020). [5] However, for identifying patterns of divergence and convergence in the usage of specific words and domains, this approach is suboptimal, as globally optimal alignment (one that minimizes the distance between the image of one language in the space of another language) may completely distort the alignment of specific words or subsets, in the interest of improving the alignment of other, larger word sets (Karidi et al., 2024).

On the other hand, *local methods* take a more granular approach, comparing the similarity of individual word meanings one at a time.

Intuitively, 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. However, the current method falls short in considering the intricate semantic relations within the groups of neighbors. To address this drawback, metrics for historical semantic change (Hamilton et al., 2016) have been adopted (Thompson et al., 2020; Beinborn and Choenni, 2020; Karidi et al., 2024). This is done by comparing the vectors of distances between a word and its neighbors across languages. Our computational approach is fully adapted from (Karidi et al., 2024).

2.2 Colexification-based Methods

The most extensive resource on colexification is the CLICS database (Rzymski et al., 2020). It provides information on colexification patterns for a wide range of concepts (a notion of a word sense; see §4), such as individual terms in domains like basic colors, body parts, and kinship, as well as more complex conceptual domains like emotion, time, and space, across 3156 languages. Each concept is linked to a set of words in different languages that are used to express that concept.

Colexification patterns are frequently used by cognitive scientists to estimate word similarity, working under the assumption that colexification of two concepts reflects similarity between them (François, 2008; Xu et al., 2020; Harvill et al., 2022). For example, the word *ka-um* in Tagalog can be linked to the concepts FATHER and ELDER BROTHER. This colexification is taken to reflect the cultural concept of the importance and authority of older male relatives in Tagalog society. Youn et al. (2016) analyzed a subset of 22 basic concepts from the Swadesh list, and showed that they exhibit patterns of meaning universality across languages. Jackson et al. (2019) conducted the first large-scale analysis using colexification patterns to assess cultural variability in people's conceptualization of emotions. However, the hypothesis that colexification and semantic similarity are tightly related is still missing direct empirical validation at scale (Natale et al., 2021).

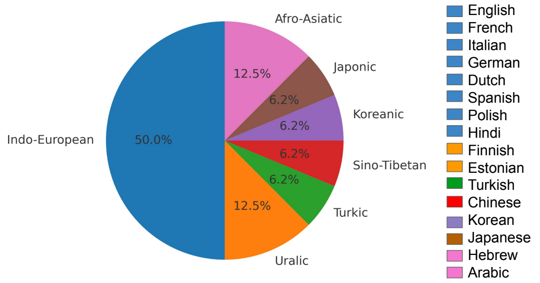

Figure 3: Distribution of languages by family. The 16 languages used in our analysis, color-coded by their language family. Each segment represents the proportion of languages within their respective families.

Recently, colexification has also been utilized in NLP to study cross-lingual transfer (Liu et al., 2023a,b; Chen et al., 2023).

3 Experimental Setup

We briefly describe our experimental setup, with full details in Appendix §A.

Data & Languages. We perform our analysis on a diverse set of 16 **languages**, spanning 7 different language families from many geographical areas across Eurasia (see Figure 3): English, French, Italian, German, Dutch, Spanish, Polish , Finnish, Estonian, Turkish, Chinese, Korean, Japanese , Hebrew, Hindi and Arabic.

The lexicon used in our analysis consists of 1,016 **concepts** sourced from NorthEuraLex (NEL) (Dellert et al., 2020), a comprehensive linguistic resource containing these concepts with their word

[5] 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).

forms in 107 different languages.

We map the concepts in NEL to **domains**, using Concepticon.[6] There are 20 domains (e.g, animals, kinship; full list is in Appendix §A), each containing $22-136$ concepts.

Models & Settings. For static word embeddings we use fastText[7] 300-dimension word embeddings, trained on Wikipedia using the skip-gram model (Bojanowski et al., 2017). For contextualised word embeddings (CWE) we use mBERT[8] (*bert-base-multilingual-uncased* model) 768-dimension vectors for the 16 languages. To extract sentences to use with contextualised models, we use the Leipzig corpus.[9] We replicate our experiments with other architectures and datasets (see Appendix D).

4 Alignment Metrics

We now turn to presenting the metrics we use in the paper. Each metric either follows the distribution-based Alignment (DISTA) or the Colexification-based Alignment (COLEXA) approach. For DISTA we follow the metrics and notations outlined in (Karidi et al., 2024).

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), and take it to mean a word sense defined independently of any specific language. Let Ω be a set of languages. 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). 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.

4.1 Colexification-based Alignment

We operationalize the notion of colexification-based alignment (COLEXA) to establish a common ground that facilitates a valid empirical comparison between DISTA and COLEXA. We experiment with a lexical alignment method that is based on colexification data (Rzymski et al., 2020). This method measures the alignment of a single concept across multiple languages. We furthermore extend it to measure the alignment of an entire domain across multiple languages. We note that different works that used COLEXA have used different methodologies, since there is no standard methodology for them. We therefore define measure that in our view captures the core statistics used by these papers.

Concept-Level Colexification-based Alignment. For every concept $c \in \mathcal{C}$ and language L_i $(i = 1, 2)$, let ${Z_c}^{(i)}$ the inverse image of $r_{L_i}(c)$:

$$Z_c^{(i)} = \{c' \in \mathcal{C} | r_{L_i}(c) = r_{L_i}(c')\}$$

We define:

$$\vartheta(c)_{L_1,L_2} = \frac{1}{2}\left(\frac{|{Z_c}^{(1)} \cap {Z_c}^{(2)}|}{|{Z_c}^{(1)}|} + \frac{|{Z_c}^{(2)} \cap {Z_c}^{(1)}|}{|{Z_c}^{(2)}|}\right)$$

Intuitively, this is a measure of the joint colexifications of the concept. For example, in Figure 1,the concept *steigen::V* is colexified with *aufgehen(sonne)::V* in English, and lexicalized as the word form *rise*, while in Finnish, an additional two concepts (*aufstehen::V* and *sich erheben::V*) are colexified (lexicalized as the word form *nousta*).

Domain-Level Colexification Based Alignment. Given the scarcity of colexifications that occur at the level of individual concepts (as many concepts are not colexified with any other concept), it is reasonable to extend the concept-level measure to quantify the alignment of a semantic domain across languages. For this we aggregate the concept-level alignment.This is done by aggregating ϑ over the concepts in $\mathcal{D}$.[10]

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

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

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

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

[10] We note that both concept-level and domain-level measures obtain values in $[0, 1]$, where a value of 1 is obtained in the case of identity in the colexifications in the domain and 0 is obtained where there are no joint colexifications.

4.2 Distribution-based Alignment

In this section, we first present the computational framework we adopt in this paper, namely Semantic Neighborhood Comparison; a standard approach for comparing embeddings in different spaces, used for both computational historical linguistics and lexical similarity tasks (Hamilton et al., 2016; Thompson et al., 2020; Beinborn and Choenni, 2020), that has recently been facilitated as an NLP task and extended to architectures beyond static representations (Karidi et al., 2024). We present several variants of this approach, including one based on contextualized word embeddings.[11]

Semantic Neighborhood Comparison (SNC). 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 $\{n_1^{(1)}, ..., n_k^{(1)}\}$ ($k = 100$ in our experiments[12]; see §3). We then translate the nearest neighbors to L_2 (§3 for translation retrieval method), by taking their translation pairs, and denote the resulting vectors with $\{n_1^{(2)}, ..., n_k^{(2)}\} \in \ell_2$. We define the unidirectional metric as

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

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

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

We refer to this alignment strategy as DISTA-STATIC.

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

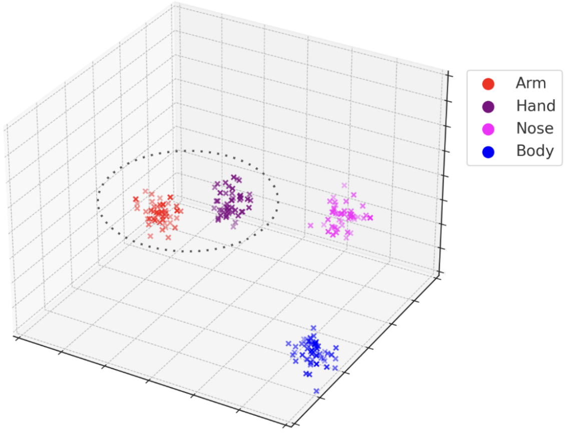

Figure 4: Illustration of nearest neighbors in the contextualized space. t-SNE plot in 2D of point clouds for the words: *arm*, *hand*, *nose*, and *body*. The nearest neighbor of *hand* is *arm*, as they have the minimal distance among all pairs of points from distinct clouds.

DISTA-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.[15] We then proceed with the SNC process, as described with DISTA-STATIC.

DISTA-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 (see Figure 4). 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 [16].

[11] In a subsequent paper (Karidi et al., 2024), we present the variants of the standard approach, for contextualised word embeddings, and perform extensive evaluation on them. Here, we choose two variants (DISTA-AVE and DISTA-CLOUD) to use in our analysis.

[12] We experimented with other values of k and selected the one that overall correlated the most with human-judgment based evaluations (see §5.4).

[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 approach of averaging over layers as described in (Karidi et al., 2024), consistent with the method used in (Vulić et al., 2020).

[16] We experiment with various pooling strategies and computational methods for building the contextualised spaces. For example, we experiment with pooling from different layers or combination of layers, similarly to DISTA-AVE. We also experiment with several definitions for the point-cloud distance, and several processing steps for generating the point-cloud itself, such as averaging the vectors within the point-cloud or clustering the set into clusters using a Gaussian Mixture

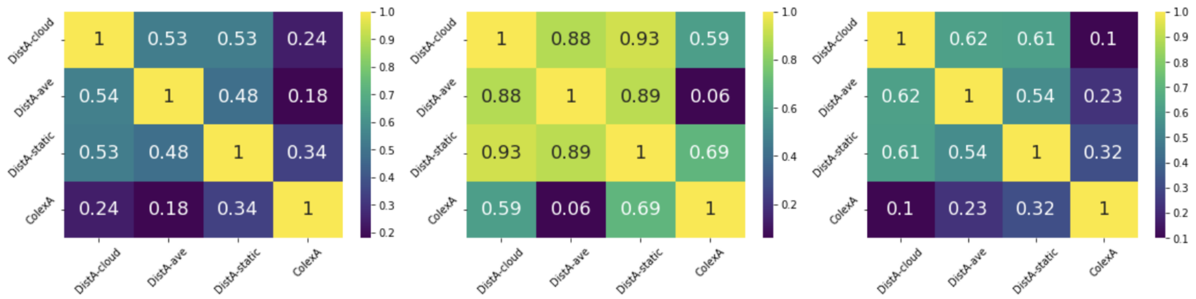

Figure 5: Correlation between DISTA and COLEXA. Correlation (Pearson) is computed for various aggregation methods: (a) concept-level, (b) domain-level and (c) language-level. All correlation values are significant with $p < 0.05$.

	COLEXA	DIST-STATIC	DIST-CLOUD
Top 3	fingernail sweep cover	March August January	thirty fifty twelve
Bottom 3	recognize endure wool	rise groan set	corner soft round

Table 1: Most and least aligned words. Word-level alignment, averaged across languages.

5 Comparing COLEXA and DISTA

The main goal of this paper is to asses the feasibility of applying two key types of alignment metrics, colexification-based (COLEXA §4.1) and distribution-based (DISTA §4.2) , within an interface setting, allowing for a direct empirical comparison of their outcomes. To this end, we initially establish the metrics in a manner that allows for a technically viable comparison (§2). We examine the convergence of their empirical findings as well as compare different metrics within the same category, that represent different operationalizations of a similar approach and data to account for.

5.1 Word-level Comparison

We start with the most straightforward level of comparison between metrics, which is their word-level correlation.[17]. Table 1 shows examples of the most and least aligned words.

Figure 5 shows that: (1) COLEXA is in low correlation with all of the DISTA methods (highest correlation is achieved between COLEXA and DISTA-STATIC, $r = 0.34$); and (2) DISTA methods are moderately correlated among themselves ($r \approx 0.5$).

Another natural question to ask is whether COLEXA and DISTA make similar predictions in terms of what concepts are more or less aligned across languages on average. That is, we investigate the correlation between COLEXA and DISTA over the set of concepts $\mathcal{C}$, where we average the score over all language pairs.

To conclude, by directly examining the statistical relation between the scores, we find that although there are similarities in the trends presented by the two methodologies, they yield different results across the board.

5.2 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 DISTA-STATIC metric, that more structured domains tend to be better aligned across languages. To examine the alignment at the domain level, we aggregate the word-level alignment over each domain (without aggregating over languages; see Figure 5). Strikingly, as opposed to the concept-level comparison, here the similarity between the DISTA methods is very high, reaching $r = 0.93$ (between DISTA-CLOUD and DISTA-STATIC). In addition, the correlaton between COLEXA and DISTA highly increases (reaching $r = 0.65$ with DISTA-STATIC). The differentiation both amongst the DISTA methods themselves and between DISTA and COLEXA has become less distinct. This finding encourages the formulation of conclusions at the domain level, as it presents to be more stable.

Model. They all yield similar trends, and are not reported due to space limitations.

[17] A metric and a language pair give rise to a vector of alignment scores. Full details on how we compute the correlations at the word, domain, and language levels can be found in Appendix §B.

	ColexA	DistA-static	DistA-cloud
Top 3	Quantity The House Social Politics	Quantity Time Kinship	Quantity Kinship Time
Bottom 3	Basic actions Sense perception Motion	Basic actions Motion The house	Agriculture Spatial relations The house

Table 2: 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".

Most and Least Aligned Domains. For DistA, the most aligned domains are Quantity, Time and Kinship (Figure 6, for DistA-cloud)[18], 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, in the Kinship domain). This robust effect exhibited in DistA is partially preserved with ColexA; Quantity the most aligned domain, whereas Time is the 4th aligned. However, Kinship is the 7th most aligned (out of the 20 domains). Table 2 presents a few examples of the differences.

5.3 Factors Influencing Alignment

We turn to analyse whether similar factors influence the alignment results for DistA and ColexA (full analysis is available in Appendix §C). Examining both lexical features, such as frequency, concreteness, and rate of change, alongside environmental features, such as cultural and geographical distance, we find that at the word level, the correlation between alignment measures and these features ranges from none to weak. However, at the domain level, an interesting finding emerges: the **rate of lexical change** is a strong predictor for both DistA and ColexA. Specifically, we observe a correlation of approximately $r \approx -0.6$ for DistA and $r = -0.81$ for ColexA. This interesting result means that words that undergo faster lexical change are less aligned across languages. This aligns with findings that polysemy plays a significant role in the rate of lexical change (Brown and Witkowski, 1983; Thompson et al., 2020), and corresponds with observations that the rate of change is negatively correlated with prototypicality (how representative a word is of its category) (Dubossarsky et al., 2017).

5.4 Comparing Against A Reference Point

Unlike many NLP tasks, when comparing the meanings of translation equivalents across languages, there is no ground truth to reference against. Instead, datasets and tasks from cognitive science literature, such as similarity in picture naming or translation norms, can serve as converging evidence for validating different measures.

This comparison has several caveats: first, it applies to a limited set of languages and stimuli; second, it is not clear that this measure captures the same notion of similarity we aim to quantify using metrics for cross-lingual lexical similarity. We hereby detail these measures and use them as a reference point for comparison.

Multipic. 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.

We filter the pictures in the Multipic dataset to only include pictures with concepts from NEL, which 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 DistA and ColexA metrics for the corresponding concepts. Results show that while DistA-ave and DistA-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. TransSim is a dataset of 562 Dutch-English translation pairs together with a human similarity rating between each pair. 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 DistA-static ($r = 0.59$, $p < 0.05$) and DistA-ave ($r = 0.51$, $p < 0.05$) rank highest.

[18] This trend persists for all DistA methods and various k values.

However, COLEXA is only weakly correlated with TransSim ($r \approx 0.1, p < 0.05$).

To conclude, when comparing both DISTA and COLEXA to norm-based measures, we find that DISTA shows a moderate correlation with some measures, whereas COLEXA does not. This distinction suggests that DISTA may be more suitable for detailed analysis of cross-lingual similarity as it is better aligned with human judgements, while COLEXA might be better suited for coarse-grained analysis. However, since these external measures apply only to a subset of languages and concepts, this limitation should be considered. Therefore, we defer a more comprehensive multi-approach comparison to future work.

6 Qualitative Analysis

To further understand the nature of alignment and convergence of the various approaches, we manually examine data from four randomly-selected languages pairs (English-German, German-Arabic, Arabic-Hebrew and Spanish-Hindi); specifically, for each method and language pair we take the top/bottom aligned 100 words, together with their 10 nearest neighbors in each language (for COLEXA we consider colexifications instead of neighbors). Even within the most aligned domains, there is variability in the order of aligned words (e.g., in DISTA-CLOUD numbers such as *seven* and *fifty* are the most aligned, whereas in DISTA-STATIC it is months, such as *March*). However, words in highly aligned domains tend to greatly overlap in their neighbors, and somewhat preserve their order of distances.

It is difficult to draw conclusions at the word-level just by looking at the raw data (this is also reflected in our empirical analysis in disagreement between the methodologies, §5.1). This is especially true for COLEXA or for the least aligned words. We do find, however, that certain words exhibit highly consistent colexification patterns across languages. For instance, the word *fingernail* frequently colexifies with the word *nail*. Based on this analysis, we hypothesize that words that colexify conceptualy similar senses (e.g., *fingernail* and *nail/hand* and *claw*/etc.) tend to have more universal colexification patterns and in turn more aligned (this echoes the finding that conceptual similarity shape colexification (Karjus et al., 2021b)), and that this is also reflected by high alignment in DISTA as this type of polysemy is less prone to affect the dissimilarity of neighbors across languages. Conversely, when two distinct senses are colexified (e.g., *bank* in English colexifies a sense of *financial institution* and a sense of *terrain*), the neighbors are likely a mix of words relating to each sense, leading to lower distributional alignment.

7 Discussion

Distribution-based and colexification-based approaches both capture a data-driven notion of similarity between the lexicons of different languages. However, they rely on different methodologies and assumptions about the data that should be accounted for, and are commonly applied in distinct ways. This raises the question of whether they are comparable, and if so – whether their predictions converge.

We find that despite the inherent differences between the methods, when viewed at the level of domains, the two appraoches show similar trends. We also find that the rate of lexical change is a strong predictor for alignment, words that change less have more stable meaning across languages. In contrast to COLEXA, DISTA is significantly correlated with extrinsic measures for meaning alignment across languages. A possible explanation is that COLEXA captures coarser aspects of meaning or that it is more suitable for scenarios which require aggregation across a more extensive range of languages. We still find this resource highly valuable, especially for investigations of high-level patterns of lexical similarity (e.g., variation in emotion concepts over the worlds languages (Jackson et al., 2019)), since it is less prone to noise stemming from the training data than DISTA. However, for a more fine-grained analysis or when less languages are available, we encourage the use of DISTA.

In this paper we lay the ground for a direct comparison of DISTA and COLEXA. Our findings call for a more nuanced discussion of lexical alignment, and also underscore the importance of taking into account multiple approaches for similarity when drawing empirical conclusions about lexical similarity. Different approaches and settings may well lead to different conclusions, which highlights the importance of justifying the technical approach taken in each paper.

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.

Hongchang Bao, Bradley Hauer, and Grzegorz Kondrak. 2021. On universal colexifications. In *Proceedings of the 11th Global Wordnet Conference*, pages 1–7, University of South Africa (UNISA). Global Wordnet Association.

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.

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.

Yiyi Chen, Russa Biswas, and Johannes Bjerva. 2023. Colex2Lang: Language embeddings from semantic typology. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 673–684, Tórshavn, Faroe Islands. University of Tartu Library.

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.

Nicholas Evans and Stephen Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *The Behavioral and Brain Sciences*, 32:429–48; discussion 448.

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

Alexandre François. 2008. Semantic maps and the typology of colexification. In Martine Vanhove, editor, *From Polysemy to Semantic change: Towards a Typology of Lexical Semantic Associations*, pages 163–215.

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

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.

Hila Gonen, Shauli Ravfogel, Yanai Elazar, and Yoav Goldberg. 2020. It's not greek to mbert: inducing word-level translations from multilingual bert. *arXiv preprint arXiv:2010.08275*.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. 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.

John Harvill, Roxana Girju, and Mark Hasegawa-Johnson. 2022. Syn2Vec: Synset colexification graphs for lexical semantic similarity. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5259–5270, Seattle, United States. 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*.

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.

Taelin Karidi, Eitan Grossman, and Omri Abend. 2024. Locally measuring cross-lingual lexical alignment: A domain and word level perspective. In *Empirical Methods in Natural Language Processing Findings (EMNLP 2024)*.

Andres Karjus, Richard Blythe, Simon Kirby, Tianyu Wang, and Kenny Smith. 2021a. Conceptual similarity and communicative need shape colexification: An experimental study. *Cognitive Science*, 45.

Andres Karjus, Richard A Blythe, Simon Kirby, Tianyu Wang, and Kenny Smith. 2021b. Conceptual similarity and communicative need shape colexification: An experimental study. *Cognitive Science*, 45(9):e13035.

Yihong Liu, Haotian Ye, Leonie Weissweiler, and Hinrich Schuetze. 2023a. Transfer learning for low-resource languages based on multilingual colexification graphs. *arxiv*.

Yihong Liu, Haotian Ye, Leonie Weissweiler, Philipp Wicke, Renhao Pei, Robert Zangenfeind, and Hinrich Schuetze. 2023b. A crosslingual investigation of conceptualization in 1335 languages. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics.

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.

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.

Anna Natale, Max Pellert, and David Garcia. 2021. Colexification networks encode affective meaning. *Affective Science*, 2.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193:217–250.

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.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

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.

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

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.

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

Jimin Sun, Hwijeen Ahn, Chan Young Park, Yulia Tsvetkov, and David R. Mortensen. 2021. Cross-cultural similarity features for cross-lingual transfer learning of pragmatically motivated tasks. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2403–2414, Online. Association for Computational Linguistics.

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.

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.

Åke Viberg. 1983. The verbs of perception: a typological study. 21(1):123–162.

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. 2010. Lexical universals of kinship and social cognition. *Behavioral and Brain Sciences*, 33:403 – 404.

Yang Xu, Khang Duong, Barbara Malt, Serena Jiang, and Mahesh Srinivasan. 2020. Conceptual relations predict colexification across languages. *Cognition*, 201.

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.

A 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. NEL comprises a list of 1016 distinct *concepts* together with their word forms in 107 languages (Table 5). Rare cases where a concept does not have a realization in a given language are excluded for that language.

Semantic Domains. We map the concepts in NEL to domains, using Concepticon.[19] There are 20 domains, each containing $22 - 136$ concepts: animals, Agriculture and vegetation, time, quantity, kinship, basic actions and technology, clothing and grooming, cognition, emotions and values, food and drink, modern world, motion, posession, sense perception, social and political relations, spatial relations, speech and language, the body, the house and the physical world.

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[20]). 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, taken from D-PLACE[21] (Thompson et al., 2020). We use the *wordfreq* library[22] for word frequencies. We then compute the log-transformed frequency (to reduce the impact of outliers and extreme values). 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.

Word Embeddings. For static word embeddings we use fastText[23] 300-dimension word embeddings, trained on Wikipedia using the skip-gram model (Bojanowski et al., 2017). For contextualised word embeddings (CWE) we use mBERT[24] (*bert-base-multilingual-uncased* model) 768-dimension vectors for the 16 languages. To extract sentences for DISTA-CLOUD, we use the Leipzig corpus.[25] We additionally conduct our experiments using XLM-RoBERTa-base [26] for DISTA-CLOUD and DISTA-AVE and on 300-dim word2vec multilingual embeddings [27] for DISTA-STATIC. Moreover, we run all of the computations for DISTA-CLOUD and DISTA-AVE with a differ-

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

[20] https://glottolog.org/

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

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

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

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

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

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

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

ent dataset; the Wikipedia section in the Leipzig Corpus, for the latest year available in each language [28]. The trends closely match those described in the paper.[29].

Hyperparameters. For our distributional based alignments (§4.2), we set $k = 100$. We experimented with other values of k and selected the one that overall correlated the most with human-judgment based evaluations (see §5.4).

B Word, Domain and Language Level Alignment

We describe here our method for computing correlations at three levels of granularity: word-level, domain-level, and language-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, $\mu(m, L_p, L_j) \in \mathbb{R}^{|\mathcal{C}|}$ is a vector whose i-th coordinate is the alignment value of concept c_i under metric m between L_p and L_j.

We use Pearson's r (with a two-tailed test for significance) for computing correlation, unless stated otherwise.

Word-level Correlation. The most direct 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 $\hat{\mu}(m) \in \mathbb{R}^{l|\mathcal{C}|}$ 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 5).

Domain-level Correlation. 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 DISTA-STATIC , that more structured domains, such as Quantity and Time, 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).

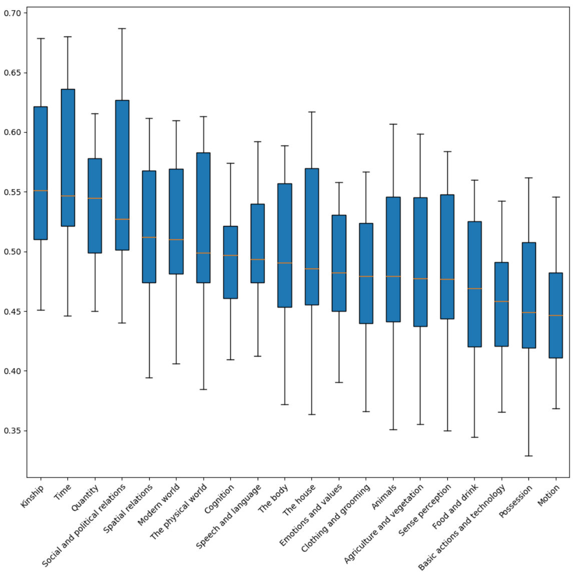

Figure 6: Alignment of domains under DISTA-AVE. The domains are ranked according to the mean value of the alignment. Each box represents the distribution of alignment values (per language pair), for a specific domain (concepts-level alignment is aggregated within each domain). The centre line is the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

Language-level Correlation. Another natural question to ask is whether COLEXA and DISTA make similar predictions in terms of what concepts are more or less aligned across languages on average. That is, we investigate the correlation between COLEXA and DISTA over the set of concepts $\mathcal{C}$, where we average the score over all language pairs. Formally, for each alignment measure $m_i \in \mathcal{M}$: $(\hat{\mu}(m_i))_j = \frac{1}{l}\sum_{(L_j,L_p)\in P} \mu(m_i, L_j, L_p)$ (we average over languages, not over concepts). Results are similar in this setting (Figure 5).

C Factors Influencing the Alignment

We examine factors influencing alignment and control for various features — lexical features like frequency, concreteness, and rate of lexical change, as well as environmental features such as geographical and cultural distance — and compare their effects on different alignment methodologies (see Section 5.3)[30]. Full results are presented in Table 3.

Correlation With Lexical Features. At the word-level ($\mu(m_i) \in \mathbb{R}^{|\mathcal{C}|l}$), there is no correlation

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

[29] See Appendix §D for experiments on other architectures than the ones presented in the main paper.

[30] We follow the analysis done in (Karidi et al., 2024) and extend it to other methodologies.

between both DistA and ColexA with respect to frequency and concreteness. There is a weak-moderate negative correlation with rate of lexical change (strongest for DistA-static, $r = -0.32$). When aggregating over domains ($\mu(m_i) \in \mathbb{R}^{lm}$) concreteness is still not correlated with any of the alignment methods; however, the correlation goes up for frequency (albeit still weakly) and jumps for rate of change ($r \approx -0.6$ for DistA and $r = -0.81$ for ColexA). This interesting result means that words that undergo faster lexical change are less aligned across languages.

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 DistA, with cultural distance playing a more prominent role. However, ColexA metrics only present a weak correlation with environmental methods. These results indicate yet another point of divergence between ColexA and DistA.

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[31] between DistA and ColexA, while controlling for the lexical and environmental features.

We find that at the concept-level the two measures are still moderately correlated with $r \approx 0.4$. At the domain-level, DistA methods are still highly correlated with one another ($r \approx 0.9$), with a moderate correlation between DistA and ColexA ($r \approx 0.5$). We use multiple linear regression to compute the adjusted R-squared value, with the environmental and lexical features as response variables. While the features explain $\approx 20\%$ of the variance for DistA, they only explain a negligible amount of the variance for ColexA. However, when aggregating over domains, the features explain up to 44% of the variance for DistA, and 69% for ColexA. This suggests that the analysis is more suitable at the domain-level.

		DistA	DistA	DistA	CA
		cloud	ave	static	
CLT	C	0.14*	0.1*	0.25*	-0.04
	D	0.2*	0.49*	0.13*	0.13*
GEO	C	0.03*	0.09*	0.22*	-0.02
	D	0.16*	0.41*	0.05	0.05
frequency	C	0.04*	0.06	0.06	0.01
	D	0.33*	0.18*	0	0
concreteness	C	0.03	0	0	0.02
	D	0.18*	0.06	0.1*	0.15*
rate-change	C	-0.32*	-0.22*	-0.25*	-0.14*
	D	-0.57*	-0.62*	-0.62*	-0.81*

Table 3: Correlation with lexical and enviromental features. Columns represent the features (CA represents ColexA, CLT denotes cultural distance and GEO denotes geographical distance) and subcolumns represents concept-level aggregation (C) vs. domain-level aggregation (D). significant correlation with $p < 0.05$ are marked by $\star$.

		DistA	DistA	DistA	CA
		cloud	ave	static	
CLT	C	0.1*	0.08	0.27*	0
	D	0.23*	0.31*	0.11*	0.11*
GEO	C	0.1	0.08*	0.15*	0
	D	0.2*	0.39*	0.1*	-0.03
frequency	C	0	-0.04	0.01	0
	D	0.35*	0.15*	0	0.01
concreteness	C	0	0	0	0
	D	0.15*	0.1*	0.15*	0.1*
rate-change	C	-0.25*	-0.27*	-0.3*	-0.1*
	D	-0.55*	-0.48*	-0.65*	-0.73*

Table 4: Correlation with lexical and enviromental features (other architectures). Columns represent the features (CA represents ColexA, 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 $\star$.

D Other Architectures

In the main paper, we conduct our analysis using the following models and data: for static word embeddings, we use fastText[32] 300-dimension word embeddings, trained on Wikipedia using the skip-gram model (Bojanowski et al., 2017). For

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

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

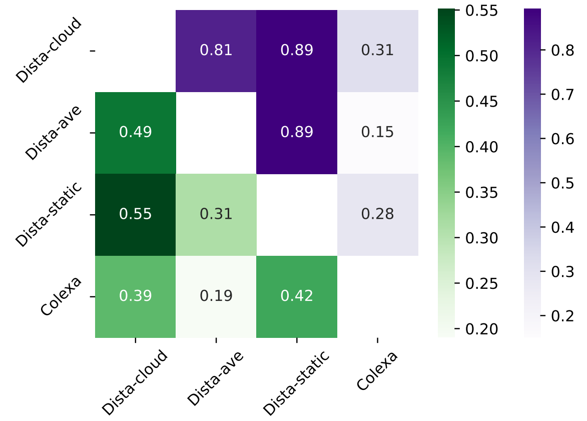

Figure 7: Correlation between DISTA and COLEXA (other architectures). Pearson correlation is computed for different aggregation methods. The **upper** matrix represents concept-level correlations, while the **bottom** matrix represents domain-level correlations. All correlation values are significant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

contextualised word embeddings (CWE) we use mBERT[33] (*bert-base-multilingual-uncased* model) 768-dimension vectors for the 16 languages.To extract sentences for DISTA-CLOUD, we use the Leipzig corpus.[34]

We additionally conduct our experiments using XLM-RoBERTa-base [35] for DISTA-CLOUD and DISTA-AVE and on 300-dim word2vec multilingual embeddings [36] for DISTA-STATIC.

Moreover, we run all of the computations for DISTA-CLOUD and DISTA-AVE with a different dataset; the Wikipedia section in the Leipzig Corpus, for the latest year available in each language [37]. The trends closely match those described in the paper (see Figure 7 and Table 4).[38].

ENGLISH FORM	CONCEPT	DOMAIN
mother	mutter::N	Kinship
mind	verstand::N	Cognition
go	gehen::V	Motion

Table 5: Concepts and their domains. Examples of concepts, labled according to the NEL dataset (§3). Each concept belongs to a semantic domain (“Domain” column). The “English Form” column contains the lexicalization of each concept in English.

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

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

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

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

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

[38] See Appendix §D for experiments on other architectures than the ones presented in the main paper.