A Crosslingual Investigation of Conceptualization in 1335 Languages

Yihong Liu[*◇], **Haotian Ye**[*◇], **Leonie Weissweiler**[*◇], **Philipp Wicke**[*◇]
Renhao Pei[*], **Robert Zangenfeind**[*], **Hinrich Schütze**[*◇]

[*]Center for Information and Language Processing, LMU Munich
[◇]Munich Center for Machine Learning (MCML)
{yihong, yehao, weissweiler, pwicke}@cis.lmu.de

Abstract

Languages differ in how they divide up the world into concepts and words; e.g., in contrast to English, Swahili has a single concept for 'belly' and 'womb'. We investigate these differences in conceptualization across 1,335 languages by aligning concepts in a parallel corpus. To this end, we propose Conceptualizer, a method that creates a bipartite directed alignment graph between source language concepts and sets of target language strings. In a detailed linguistic analysis across all languages for one concept ('bird') and an evaluation on gold standard data for 32 Swadesh concepts, we show that Conceptualizer has good alignment accuracy. We demonstrate the potential of research on conceptualization in NLP with two experiments. (1) We define crosslingual stability of a concept as the degree to which it has 1-1 correspondences across languages, and show that concreteness predicts stability. (2) We represent each language by its conceptualization pattern for 83 concepts, and define a similarity measure on these representations. The resulting measure for the conceptual similarity between two languages is complementary to standard genealogical, typological, and surface similarity measures. For four out of six language families, we can assign languages to their correct family based on conceptual similarity with accuracies between 54% and 87%.[1]

1 Introduction

Languages differ in how they divide up the world into concepts and words. The Swahili word 'tumbo' unites the meanings of the English words 'belly' and 'womb'. Therefore, English forces its speakers to differentiate between the general body region "front part of the human trunk below the ribs" and one particular organ within it (the womb) whereas Swahili does not. Similarly, Yoruba 'irun' refers to both hair and wool. Again, English speakers must make a distinction whereas Yoruba has a single hair concept that includes the meaning *animal hair for clothing*.

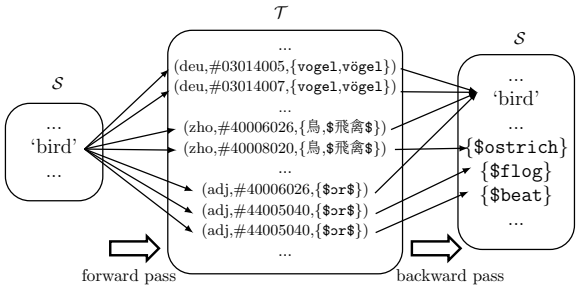


Figure 1: An example of the directed bipartite graph we construct, for the concept 'bird'. Each node in $\mathcal{S}$ is a set of strings. Each node in $\mathcal{T}$ is a triple of language, verse identifier (i.e., a sentence ID) and set of strings identified as correlated with 'bird'. Conceptualizer induces edges from both $\mathcal{S}$ to $\mathcal{T}$ and $\mathcal{T}$ to $\mathcal{S}$ that we then use for analysis and prediction. In the example, we see that the set of strings correlated in Mandarin (zho) also refers to "ostrich" and that the correlated string in Adioukrou (adj) is ambiguous between "bird" and "flog".

While studies have looked at conceptualization within different languages (Ravin and Leacock, 2000; Goddard and Wierzbicka, 2013), we present a crosslingual study that directly compares conceptualization in 1,335 languages. The empirical basis are word and ngram correspondences in the **P**arallel **B**ible **C**orpus (PBC, (Mayer and Cysouw, 2014)). We introduce Conceptualizer, a method that reliably aligns a set of 83 concepts across all PBC languages. The 83 concepts are partly chosen to be well represented in the Bible, and partly from Swadesh 100 (Swadesh, 2017). The alignments are formalized as a bipartite graph between English (the source) and the target languages.

The simple idea underlying Conceptualizer– illustrated in Figure 1 – is that, starting with one of the 83 concepts in English as the focal concept (the search query), we can identify divergent conceptualizations by first searching for target ngrams

[1]We release our code at https://github.com/yihongL1U/conceptualizer

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 12969–13000
July 9-14, 2023

highly associated with the focal concept, and then searching for English ngrams highly correlated with the target ngrams we found. If the English ngrams correspond to the original focal concept, then the conceptualizations do not diverge. In contrast, take the example of divergence described above: we start with the focal concept 'hair', find Yoruba 'irun' and then two English concepts, not one, that are highly associated with 'irun': 'hair' and 'wool'. This indicates that English and Yoruba conceptualizations diverge for 'hair'.

Our main contribution is that we present the first empirical study of crosslingual conceptualization that grounds the semantics of concepts directly in contexts – the sentences of the parallel corpus. This ensures that our work is based on identical (or at least very similar) meanings across all 1,335 languages we investigate. For example, verse Matthew 9:7 has the same meaning in English: "Then the man got up and went home.", in Chinese: "那個人就起來，回家去了。" and in each of the other 1,333 languages. Such a direct grounding in meaning across a large set of languages has not previously been achieved in work on conceptualization in theoretical or computational linguistics.

In addition, we make the following contributions. (i) We propose Conceptualizer, an alignment method specifically designed for concept alignment, that operates on the level of ngrams and ngram sets. (ii) We conduct an evaluation of Conceptualizer for the concept 'bird' in all 1,335 languages. The result is a broad characterization of how the conceptualization of bird varies across the languages of the world. Out of 1,335 languages, Conceptualizer only fails 15 times (due to data sparseness) for 'bird'. (iii) We evaluate Conceptualizer for 32 Swadesh concepts on a subset of 39 languages for which translation resources exist and demonstrate good performance. (iv) Using the ratings provided by Brysbaert et al. (2014), we give evidence that concreteness (i.e., the degree to which a concept refers to a perceptible entity) causes a concept to be more stable across languages: concrete concepts are more likely to have one-to-one mappings than abstract concepts. (v) We propose a new measure of language similarity. Since we have aligned concepts across languages, we can compute measures of how similar the conceptualization of two languages is. We show that this gives good results and is complementary to genealogical, typological and surface similarity measures that are commonly used. For example, Madagascar's Plateau Malagasy is conceptually similar to geographically distant typological relatives like Hawaiian, but also to typologically distant "areal neighbors" like Mwani and Koti. For four out of six language families, based on conceptual similarity, we can assign languages to their correct family with between 54% and 87% accuracy.

2 Related Work

In linguistics, conceptualization has been studied empirically with regards to crosslingual polysemy or colexification (François, 2008; Perrin, 2010; List et al., 2013; Jackson et al., 2019) as well as areal and cultural influences on concept similarity (Gast and Koptjevskaja-Tamm, 2018; Thompson et al., 2020; Georgakopoulos et al., 2022). Most of this work is based on human annotations, such as CLICS (List, 2018; List et al., 2018; Rzymski et al., 2020), a database of colexification. However, the coverage of such resources in terms of concepts included, especially for some low-resource languages, is low. Therefore we explore the use of an unannotated broad-coverage parallel corpus as an alternative. Expanding this work to many languages is important to the extent that we accept some (weak) form of linguistic relativity, i.e., the hypothesis that language structure (including conceptualization) influences cognition and perception (Boroditsky et al., 2003; Deutscher, 2010; Goddard and Wierzbicka, 2013).

Methodologically, our work is closely related to Östling (2016) who explores colexification through PBC. He targets specific colexification pairs and investigates their geographical distribution using word alignments. In comparison, our method allows us to identify alignments beyond the word level and therefore richer associations among concepts are obtained. Our proposed method Conceptualizer is also close to semantic mirrors (Dyvik, 2004), a method to explore semantic relations using translational data. The authors focus on an English-Norwegian lemmatized parallel corpus; in contrast, we investigate 1,335 languages, most of which are low-resource and for many of which lemmatization is not available. In addition, this paper is related to recent work that uses PBC to investigate the typology of tense (Asgari and Schütze, 2017), train massive multilingual embeddings (Dufter et al., 2018), extract multilingual named entities (Severini et al., 2022), find case markers in a multilingual setting

(Weissweiler et al., 2022) and learn language embeddings containing typological features (Östling and Kurfalı, 2023).

Like Conceptualizer, Şenel et al. (2017, 2018) analyzed the semantic similarity of concepts across languages (mainly European ones). But they use pretrained word embeddings (Mikolov et al., 2013; Pennington et al., 2014), which are not available in high enough quality for most of the low-resource languages we cover in this work.

Computational criteria for language similarity have been taken from typology (Ponti et al., 2019; Georgi et al., 2010; Pires et al., 2019; Daumé III, 2009), morphology (Zervanou et al., 2014; Dautriche et al., 2017) and language-model surface similarity (Pires et al., 2019; Wu and Dredze, 2020). We propose a new similarity measure, based on conceptualization, with complementary strengths and weaknesses.

There is a large body of work on statistical and neural word alignment; recent papers with extensive discussion of this subfield include (Ho and Yvon, 2019; Zenkel et al., 2020; Wu et al., 2022). We show below that the standard alignment method Eflomal (Östling and Tiedemann, 2016) does not work well for our problem, i.e., for identifying high-accuracy associations between concepts.

3 Methodology

3.1 Data

We work with the Parallel Bible Corpus (PBC, Mayer and Cysouw (2014)). We use 1,335 Bible translations from PBC, each from a different language as identified by its ISO 639-3 code. For most languages, PBC only covers the New Testament (NT) (≈7,900 verses). For a few hundred, it covers both NT and Hebrew Bible (≈30,000 verses). See §A.1 for details of the PBC corpus.

From Swadesh 100 (Swadesh, 2017), a set of 100 basic universal concepts, we select the 32 concepts that occur with frequency $5 < f \leq 500$ in both NT and Hebrew Bible. We call the resulting set of 32 concepts **Swadesh32**. We also select **Bible51** from the Bible, a set of 51 concepts that are of interest for crosslingual comparison. Notably, we include abstract concepts like 'faith' that are missing from Swadesh32. See §A.2 for concept selection details.

3.2 Conceptualizer

Bipartite graph. We formalize the concept alignment graph as a directed bipartite graph. With

	Notation
$\mathcal{P}$	power set
Σ	alphabet
$\mathcal{G}$	directed bipartite graph
$\mathcal{S}$	the set of source nodes of $\mathcal{G}$
$\mathcal{T}$	the set of target nodes of $\mathcal{G}$
Λ	the set of 1335 languages
l	$l \in \Lambda$, a language
Π	the set of 31,157 Bible verses
$\Pi(l, U)$	set of verses of l containing $u \in U$
V	$V \subseteq \Pi$, a set of verses
v	$v \in \Pi$, a verse
F	focal (English) concept (this is a set), $F \in \mathcal{S}$
s	source (English) string
S	set of source (English) strings
t	target string
T	set of target strings
U	set of strings (source or target)

Table 1: Notation

Σ denoting the alphabet, let $\mathcal{S} \subset \mathcal{P}(\Sigma^*)$ be the set of source nodes, each corresponding to a concept, represented as a set of strings from the source language; e.g., {`$belly$`, `$bellies$`} for 'belly', where `$` denotes the word boundary. In this paper, we always use English as the source language. With Λ denoting the set of languages and Π the set of verses, let $\mathcal{T} \subset \Lambda \times \Pi \times \mathcal{P}(\Sigma^*)$ be the set of target nodes, each corresponding to a triple of target language l, Bible verse v and a set of strings from language l, one of them occurring in v. We represent the concept correspondences as a directed bipartite Graph $\mathcal{G} \subset \mathcal{S} \times \mathcal{T} \cup \mathcal{T} \times \mathcal{S}$, as shown in Figure 1. See §B for method details. Table 1 gives our notation. The reason for our asymmetric design of the graph (concept *types* on the source side, concept *tokens* occurring in context on the target side) is that we want to track how much evidence there is for a concept-concept correspondence. The more edges there are in the graph, the more reliable the correspondence is.

Association for alignment. We can represent a source concept (e.g., {`$belly$`, `$bellies$`}) as the set of verses V in which it occurs. In contrast to standard alignment algorithms, we exhaustively search all strings t of the target language l for high correlation with V. For example, we search for the French string t that has the highest correlation with the verses that contain {`$belly$`, `$bellies$`}; the result is t="ventre". This means that we are not limited to knowing what the relevant (tokenization) units are in advance, which is not possible for all 1,335 languages. We use the χ^2 score $\chi^2(l, t, V)$ as a measure of correlation: we test, for all t, whether the two categorical variables $t \in v$ (short for: t is

Algorithm 1: Forward Pass (FP)

Input: focal concept F, language l

1 $T \leftarrow \emptyset$;
2 **for** $i \leftarrow 1$ ***to*** M **do**
3 $\quad V \leftarrow \{v \in \Pi | (\exists s \in F : s \in v) \wedge (\neg \exists t \in T : t \in v)\}$;
4 $\quad$ **if** $i = 1$ **then** $V_1 \leftarrow V$;
5 $\quad$ **if** *COVERAGE*$(l, T, V_1) \geq \alpha$ **then** break;
6 $\quad t \leftarrow \arg\max_{\{t \in \mathcal{P}(\Sigma^*) : t \notin T\}} \chi^2(l, t, V)$;
7 $\quad T \leftarrow T \cup \{t\}$;
8 **end**
9 **return**
$\{(F, (l, v, T)) | \exists t \in T : t \in v\}$

Algorithm 2: Backward Pass (BP)

Input: focal concept F, language l

1 $S \leftarrow \emptyset$;
2 **for** $i \leftarrow 1$ ***to*** M **do**
3 $\quad V \leftarrow \{v \in \Pi | (\exists (F, (l, v, T)) \in \mathcal{G}) \wedge (\neg \exists s \in S : s \in v)\}$;
4 $\quad$ **if** $i = 1$ **then** $V_1 \leftarrow V$;
5 $\quad$ **if** *COVERAGE*$(eng, S, V_1) \geq \alpha$ **then** break;
6 $\quad s \leftarrow \arg\max_{\{s \in \mathcal{P}(\Sigma^*) : s \notin S\}} \chi^2(\text{eng}, s, V)$;
7 $\quad S \leftarrow S \cup \{s\}$;
8 **end**
9 **return**
$\{((l, v, T), S') | (\exists (F, (l, v, T)) \in \mathcal{G}) \wedge (\exists s \in S' : s \in S, s \in v)\}$

Figure 2: Forward Pass (FP) and Backward Pass (BP) for graph induction

a substring of verse v in language l) and $v \in V$ are independent. We select the t with the highest score.

Termination. For a query string q, e.g., `$hair`, occurring in verses V, we want to find a set U of highly associated ngrams in the target language l that covers all of V. Because of noise, translation errors, nonliteral language etc., this is often impossible. We therefore terminate the search for additional target strings when $\text{COVERAGE}(l, U, V) \geq \alpha$ where we set $\alpha = .9$ and define:

$$\text{COVERAGE}(l, U, V) = \frac{|\Pi(l, U) \cap V|}{|V|}$$

i.e., the fraction of V covered by the strings in U.

Graph induction. Figure 2 shows that Conceptualizer consists of a forward pass (FP, Algorithm 1) that adds edges $e \in \mathcal{S} \times \mathcal{T}$ and a backward pass (BP, Algorithm 2) that adds edges $e \in \mathcal{T} \times \mathcal{S}$ to $\mathcal{G}$. FP and BP are essentially the same. To abstract from the direction, we will use the terms query language and retrieval language. In FP (resp. BP), the query language is the source (resp. target) language and the retrieval language is the target (resp. source) language.

- Let q be the query string from the query language.
- The set R holds retrieval language strings that are highly associated with q. R is initially empty. R is T (a set of target strings) or S (a set of English source strings) in the algorithms.
- In each iteration, we find the retrieval language string r with the highest association to those verses containing the query q that are not yet covered by R.
- We terminate when coverage by R (of verses containing q) exceeds the threshold α.
- We return all edges that go from a query language node that contains q to a retrieval language node that contains a string from R.

The formal description in Figure 2 is slightly more complex because the query q is not a single string but a set. But this extension is straightforward. We now explain the formal description in Figure 2.

We invoke FP and BP for all pairs (focal concept F, target language l) and merge the result with $\mathcal{G}$ for each invocation. Writing PASS for FP or BP:

$$\mathcal{G} \leftarrow \mathcal{G} \cup \text{PASS}(F, l)$$

For the following description of the algorithms, we write $s \in v$ for "string s (in language l) is a substring of (the language l version of) verse v". For brevity, we describe FP (Algorithm 1) [and describe BP (Algorithm 2) in square brackets]. Line 1: T [S] collects target [source] strings. Line 2: M is the maximum number of iterations; we set $M = 5$. Line 3: V is the set of verses that contain a string in F [were linked by an edge from F in FP], but are not yet covered by T [S]. Line 4: We save the result for $i = 1$ (or $T = \emptyset$ [$S = \emptyset$]) in V_1, the base set of verses. Line 5: If the coverage that T [S] has of V_1 exceeds a threshold α, we terminate; we set $\alpha = .9$. Line 6: We find the target string t [source string s] that is most associated with V, ignoring target [source] string candidates already covered. Line 7: t [s] is added to T [S]. Line 9: In FP, we return a set of new edges that start at the focal concept F and end at a target node (l, v, T) whose verse v contains a string t from T. Line 9: In BP, we return a set of new edges that start at a target node (l, v, T) that was connected to F in FP and end at an S' that contains a highly associated source string s (i.e., $s \in S$) in v.

4 Evaluation

4.1 Single concept across all languages

We first evaluate how well our method performs at identifying associated concepts across the highly

1-1	polysemy	ambiguity	failure	total
687	579	54	11	1331
match	**overlap**	**no overlap**	**no translation**	**total**
488	192	457	194	1331

Table 2: Evaluation of Conceptualizer for 'bird'. Top: Linguistic analysis. Bottom: PanLex results.

model	partial	strict	relaxed	FP
Conceptualizer	87.21	84.88	89.69	1.03
Eflomal 0	89.52	87.80	91.23	10.42
Eflomal 1	86.98	84.88	89.18	4.50
Eflomal 0.1	78.68	76.12	81.44	1.07

Table 3: Recall of proposed Swadesh32 translations T on NoRaRe translations N, averaged over concept-language pairs. The score for a concept-language pair is $|T \cap N|/|N|$ (partial), 1 iff $|T \cap N| = |N|$ (strict) and 1 iff $|T \cap N| \geq 1$ (relaxed). FP: average false positives.

diverse set of languages we cover. Since there is no appropriate broad-coverage high-quality resource, this requires an expensive manual analysis by a linguist. We can therefore only perform it for one concept in this paper. We choose the focal concept 'bird', defined as {`$bird`, `$fowl`, `$flying$creature`, `$winged$creature`}. For each language l, we analyze the hits we get for 'bird' in l, primarily by looking at its BP hits in English, i.e., the English strings that are proposed in BP by running Conceptualizer on 'bird'. Defining R as the set of verses in which BP hits occur and B as the set of verses in which 'bird' occurs, we use four evaluation categories. (1) **one-to-one**. $R \approx B$. In detail: $|R - B| < .1|B|$ and $R - B$ does not contain plausible additional hits. (2) **polysemy**. $R \supset B$ and $R - B$ consists of verses with concepts closely related to 'bird', e.g., 'dove', 'fly'. (3) **ambiguity**. $R - B$ contains verses in which neither 'bird' nor closely related meanings occur. However, there is a second "non-bird" meaning of the BP hits; e.g., for Adioukrou the FP hit is "ɔr" and the BP hits correspond to two clusters, a *bird* cluster and a *hitting* cluster. (4) **failure**. $R - B$ or $B - R$ is large and this cannot be attributed to polysemy or (simple) ambiguity. See §C.1.1 for details. Table 2 (top) shows that Conceptualizer found the translation of 'bird' in almost all languages where we count the **ambiguity** case (e.g., Adioukrou "ɔr" meaning both *bird* and *hitting*) as a success. The search failed for 4 languages ($4 = 1335 - 1331$) for which we have no verse that contains 'bird' in English and 11 languages for many of which the number of verses was small. Thus, Conceptualizer requires a large enough parallel corpus for good performance.

We also evaluate on PanLex (Kamholz et al., 2014), `http://panlex.org`. Defining P as the translations from PanLex and T as the FP hits for 'bird', we use the following four categories. (1) PanLex gives **no translation**. $P = \emptyset$. (2) **no overlap**. $P \cap T = \emptyset$. (3) **overlap**. $0 < |P \cap T| < |T|$. (4) **match**. $|P \cap T| = |T|$. See §C.1.2 for details. Table 2 (bottom) shows that for PanLex languages, Conceptualizer performs well on $\approx 60\%$: $(488 + 192)/(488 + 192 + 457)$. In a qualitative analysis, we found four reasons for the 457 **no overlap** cases. (i) A language has a very small corpus in PBC. (Sparseness was also the reason for failure in Table 2, top). (ii) Conceptualizer did find correct translations of 'bird', but they are missing from PanLex. (iii) There is a dialect/variety mismatch Bible vs PanLex (no occurrence of the PanLex translation in our corpus). (iv) PanLex incorrectly translates through an intermediate language. For example, since PanLex has no direct translation of English 'bird' to Chorote Iyowujwa, it goes through Gimi 'nimi' (which means both *bird* and *louse*) and returns Chorote Iyowujwa 'inxla7a'. But 'inxla7a' only means *louse*. Another example is that PanLex translates 'bird' as 'San' instead of the correct (Sampu et al., 2005) 'nghoq' for Achang. Thus, PanLex translations through the intermediate mechanism are unreliable while our FP hit can find the correct translation.

Taking the two evaluations together (manual analysis of BP hits and comparison of FP hits to PanLex translations), we interpret the results as indicating that Conceptualizer reliably finds the correct translation of the focal concept, but can fail in case of data sparseness.

4.2 Swadesh concepts

We next evaluate on Swadesh32 (§3.1). Table 2 indicates that PanLex quality is low for many languages. We therefore use NoRaRe (Tjuka et al., 2022), `http://norare.clld.org`. We use all 582 concept-language pairs for which NoRaRe gives a translation. For a concept-language pair, let T be the proposed translations (from Conceptualizer or Eflomal) and N gold (from NoRaRe). Then we compute recall as $|T \cap N|/|N|$. We match two ngrams if one is a substring of the other;

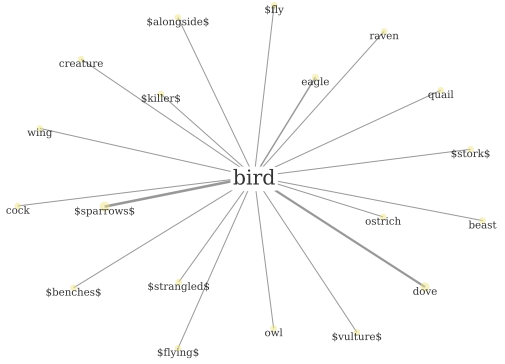


Figure 3: The crosslingual semantic field of 'bird'

e.g., "oiseau" is correct for "oiseaux". For Eflomal (Östling and Tiedemann, 2016), we set T to the set of target language words aligned with one of the focal concept words (e.g., {`$belly$`, `$bellies$`}). Eflomal 0, 1, 0.1 denotes that we only keep translations whose frequency is > 0, > 1 and $> .1\Pi(l, F)$, respectively.

Table 3 shows that Conceptualizer's Swadesh32 translations have high recall (roughly 85% and higher, depending on the measure), with few false positives (1.03). For Eflomal, however, as we restrict matches to high-precision matches (i.e., going from 0 to 1 and to .1), both recall and false positives (FP) drop. Our interpretation is that the alignments obtained by Eflomal are noisy: Eflomal misaligns the focal concept with many irrelevant words. In contrast to Conceptualizer, Eflomal offers no good tradeoff. This validates that we use Conceptualizer instead of standard aligners like Eflomal. Most importantly, the evaluation on NoRaRe shows that Conceptualizer has high recall and produces few false positives, which are prerequisites for further reliable exploration/analysis. See §C.2 for details of the evaluation (including an additional experiment in terms of coverage compared with Eflomal).

4.3 Concept stability

We define the crosslingual semantic field $\mathcal{F}$ of focal concept $F \in \mathcal{S}$ as the second neighborhood of F, the set of nodes at a distance 2 from F:

$$\mathcal{F}(F) = \{S \in \mathcal{S} | \exists c : (F, c) \in \mathcal{G} \wedge (c, S) \in \mathcal{G}\}$$

Figure 3 shows the crosslingual semantic field of 'bird'. The strength of the line connecting 'bird' and S (which contains an English string) indicates the number of languages through which 'bird' can reach S. "eagle", "dove" and "sparrows" have thick lines, indicating that there are many languages for which Conceptualizer connects 'bird' to a target string whose meaning includes a bird species. The size of a node indicates the number of paths from 'bird' to the string the node represents. For example, the size of the 'bird' node (i.e., F) indicates the number of recurrent paths, i.e., $|\{c|(F, c) \in \mathcal{G} \wedge (c, F) \in \mathcal{G}\}|$. The visualization in Figure 3 suggests that 'bird' is *stable* crosslingually: if we go roundtrip from English to a target language l and back, in most cases what we get is 'bird'. This is often not true (as we will see shortly) for a more abstract concept like 'mercy'. The proportion of recurrent paths is small: many paths starting from 'mercy' go to other nodes, such as "pity" and "poor", indicating that it is unstable. See §E for visualizations of all 83 concepts.

Accuracy	Precision	Recall	F_1
0.71	0.65	0.88	0.75

Table 4: Performance of predicting a concept's stability from its concreteness

We define the *stability* $\sigma(F)$ of a focal concept $F \in \mathcal{S}$ as:

$$\sigma(F) = \frac{|\{c|(F, c) \in \mathcal{G} \wedge (c, F) \in \mathcal{G}\}|}{|\{c|(F, c) \in \mathcal{G}\}|}$$

Thus, for a stable concept F (one whose stability is close to 1.0), most paths starting from F are part of a "recurrent" path that eventually returns to F. In contrast, an unstable concept F like 'mercy' has relatively fewer such recurrent paths and a large proportion of its paths go to other concepts.

We hypothesize that one cause of stability is concreteness: concrete concepts are more stable across languages than abstract ones because they are directly grounded in a perceptual reality that is shared across languages. To test this hypothesis, we define a concept to be concrete (resp. abstract) if its concreteness score γ according to (Brysbaert et al., 2014) is $\gamma \geq 3.5$ (resp. $\gamma \leq 2.5$). 69 of our 83 concepts are either abstract or concrete, according to this definition (see Tables 12 and 13 in the Appendix for concreteness and stability measures of all 83 concepts). We define a concept F to be stable iff $\sigma(F) \geq 0.6$. Table 4 shows that when we predict stability based on concreteness (i.e., a concept is predicted to be concrete iff it is stable), accuracy is high: $F_1 = .75$. This is evidence that

k	concepts	ATLA	AUST	INDO	GUIN	OTOM	SINO	all
2	32	.21	.2	.53	.09	.14	.00	.13
	51	.24	.19	.26	.08	.04	.03	.11
	83	.29	.31	.49	.11	.14	.04	.17
4	32	.54	.41	.80	.24	.39	.15	.29
	51	.52	.45	.48	.18	.12	.09	.24
	83	.63	.51	.77	.31	.28	.09	.32
6	32	.63	.49	.85	.30	.43	.16	.33
	51	.64	.57	.57	.20	.13	.13	.30
	83	.74	.60	.83	.40	.37	.12	.37
8	32	.68	.53	**.87**	.34	.51	**.18**	.36
	51	.71	.59	.60	.22	.14	.15	.32
	83	.78	.60	.86	**.42**	.36	**.18**	**.39**
10	32	.73	.56	.84	.34	**.54**	**.18**	.37
	51	.74	**.61**	.61	.21	.09	.12	.32
	83	**.80**	**.61**	.83	.41	.28	.16	.38

Table 5: Accuracy of prediction of typological family based on nearest neighors in Conceptualizer-based representation space. Representations for Swadesh32 (32), Bible51 (51) and All83 (83) concepts. k: number of nearest neighbors. Family abbreviations: see text. **Bold** (underlined): best (second-best) result per column.

our hypothesis is correct: concreteness is an important contributor to stability. See §5.1 for further analysis of the stability of concepts.

4.4 Language similarity

We now propose and evaluate a new measure of similarity between languages, *conceptual similarity*, based on conceptualization. Since we have aligned concepts across languages, we can compute measures of how similar the conceptualization of two languages is. For example, in contrast to Western European languages, Chinese, Korean, and Japanese have one concept that means both *mouth* and *entrance*. Our measure aggregates such patterns over many concepts and predicts higher similarity between the three East Asian languages and lower similarity to Western European languages.

To compute conceptual similarity, we represent a language l as the concatenation of 83 vectors $\vec{v}(l, F_j)$, each capturing how it represents one of our 83 concepts:

$$\vec{v}(l) = [\vec{v}(l, F_1); \vec{v}(l, F_2); \ldots; \vec{v}(l, F_{83})]$$

where $[;]$ is vector concatenation. We define $\vec{v}'(l, F_j)$ as a 100-dimensional vector and set

$$\vec{v}'(l, F_j)_i = |\{c | (F_j, c) \in \mathcal{G} \wedge (c, \{e_i\}) \in \mathcal{G}|$$

i.e., the number of paths from F_j to the English ngram e_i; here we only consider nodes $c = (l', v, T)$ for which $l' = l$, i.e., only nodes that belong to language l. For example, 'mouth' connects with Chinese nodes containing "口" in FP. BP connects these nodes not only to 'mouth', but also to "entrance". Our convention is that the first dimension $\vec{v}'(l, F_j)_1$ always represents the value of the focal concept F_j. To define the other dimensions, we sort all associated English ngrams e_k according to the number of languages in which they are associated with F_j and select the top 99[2]; these are then the dimensions 2-100 of $\vec{v}'(l, F_j)$. We compute the final vector $\vec{v}(l, F_j)$ by normalizing $\vec{v}'(l, F_j)$ by $\sum_k \vec{v}'(l, F_j)_k$.

$\vec{v}(l, F_j)$ captures which concepts related to F_j are clustered in l and thereby indicates l's similarity to other languages. For example, for the focal concept 'mouth', the $\vec{v}(l, F_j)$ for Chinese, Japanese and Korean are more similar, but they are less similar to $\vec{v}(l, F_j)$ for Western European languages.

We can now define the *conceptual similarity* between two languages l_1 and l_2 as the cosine similarity between their vectors:

$$\text{c-sim}(l_1, l_2) = \cos(\vec{v}(l_1), \vec{v}(l_2))$$

We evaluate on Glottolog 4.7 (Hammarström et al., 2022). We select the six language families that have more than 50 members in the PBC: Atlantic-Congo (ATLA), Austronesian(AUST), Indo-European (INDO), Nuclear Trans New Guinea (GUIN), Otomanguean (OTOM) and Sino-Tibetan (SINO). We then evaluate conceptual similarity on a binary classification task: *Is the majority of language l's k nearest neighbors in the same family as l?* In addition to representations based on all 83 focal concepts (referred to as **All83**), we also analogously create representations based just on Swadesh32 and Bible51.

Table 5 shows that for two "dense" families (i.e., most members have close relatives), our results are good (up to .8 for ATLA, .87 for INDO). For AUST, GUIN and OTOM, about half of the predictions are correct for the best k. SINO performance is bad, indicating that SINO languages are conceptually more distant from each other. The difference between Swadesh32 and Bible51 performance is large in some cases, especially for INDO and OTOM. We hypothesize that the conceptualization for more abstract concepts in Bible51 is more variable than for more concrete concepts in Swadesh32.

[2]For some focal concepts that are less divergent, e.g., 'bird', we obtain fewer than 99 dimensions

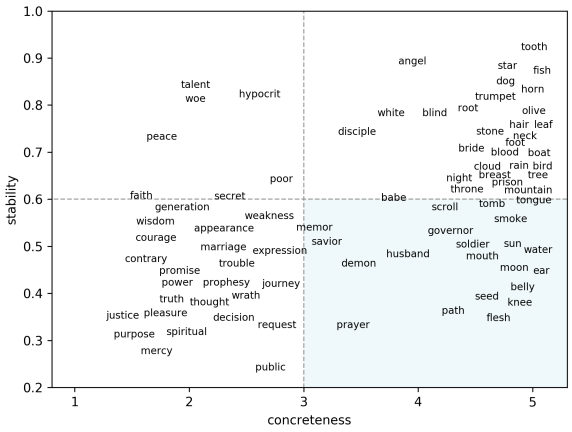


Figure 4: *stability* measure vs. *concreteness* scores.

types	concept	associated
metonymy	'ear'	*hear*, *listen*
	'mouth'	*word*, *speak*
	'sun'	*day*, *east*
	'moon'	*month*
	'belly'	*womb*, *birth*
	'soldier'	*army*, *military*
metaphor	'seed'	*son*, *offspring*
	'mouth'	*entrance*
other	'water'	*waters*, *sea*, *drink*
	'knee'	*obeisance*, *worship*

Table 6: Selected examples of focal concepts that are unstable due to different types of meaning extensions: **metonymy**, **metaphor**, and **other**.

The main result of this evaluation is that the language representations (i.e., the $\vec{v}(l)$) derived by Conceptualizer are a good basis for assessing the similarity of languages. §5.2 discusses in detail the complementarity of conceptual similarity to other similarity measures.

5 Analysis

5.1 Concept stability

Figure 4 and Table 4 support our hypothesis that concreteness contributes to stability, as most concrete concepts tend to be stable and most abstract concepts tend to be less stable. However, recall is higher than precision in Table 4, indicating that stable concepts are mostly concrete but not vice versa, as plenty of concepts are located in the lower right in Figure 4. In our analysis, we find that some concrete concepts are unstable because meaning is extended by semantic processes like metonymy and metaphor. We now give examples of our analysis of this phenomenon in Table 6. See §E for the visual semantic fields of all 83 concepts.

Table 6 shows (i) concrete concepts whose instability is plausibly caused by metaphor, metonymy, and other semantic extensions and (ii) crosslingually associated English concepts that indicate instability. For example, metonymy, where a concept is represented by the name of something associated with it, results in instability frequently. This is in line with crosslinguistic studies that show that metonymy seems to be a universal language phenomenon (Khishigsuren et al., 2022).

Processes of semantic extension are central to cognitive linguistics for explaining historical changes, language use and framing in discourse (Evans, 2006; Lakoff and Johnson, 2008). Since these processes are pervasive in communication (Sopory and Dillard, 2002), in the lexicon (Khishigsuren et al., 2022) and in historical language change (Xu et al., 2017), they are also a plausible cause for crosslingual instability of concepts.

5.2 Language similarity

Figure 5 shows a visualization of conceptual similarity using t-SNE (Van der Maaten and Hinton, 2008). Each point in the figure represents a language. Color indicates family on the left (for the six families from §4.4) and area on the right (for all languages, area extracted from Glottolog 4.7).

The right figure (area) shows that conceptual similarity is correlated with geographic proximity: except for Australia, all regions are concentrated in a few subclusters, with Eurasia being most clustered and Africa least. For example, there are two clusters (top left, top center) that cover most South American languages. This is consistent with findings that close languages often influence each other (Sapir, 1912). Of course, the main reason is the correlation between conceptual and typological relatedness (see below) and between typological and geographic proximities. But there are also indications of geography-induced conceptual similarity, e.g., most Eurasian languages have close African neighbors. We present examples at the end of §5.2.

That geographic proximity is not sufficient for conceptual similarity is demonstrated by Papunesia: it is distributed over most of the figure, indicating that the underlying conceptualizations are quite different. This is consistent with the well-known fact that Papunesia exhibits enormous linguistic diversity in a very small geographic area (Foley, 1986). Of course, there is still subcluster structure: most Papunesian languages have near Papunesian

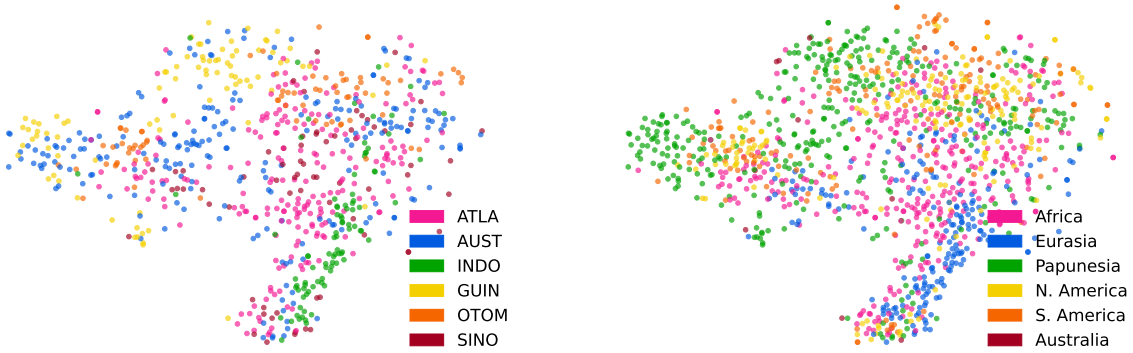


Figure 5: t-SNE of languages represented as Conceptualizer-based vectors of the Swadesh32 concepts. Colors indicate family (left) and area (right).

neighbors in conceptual space.

Turning to the left figure (family), we see that there is agreement of conceptual similarity with typological similarity. Some families form a relatively clear cluster, in particular, Indo-European, suggesting that Indo-European languages are similar in conceptualization. This explains Indo-European's high accuracy in Table 5.

But there is also complementarity between conceptual similarity and typological similarity. Sino-Tibetan is spread out across the entire figure, explaining its low accuracy in Table 5. For Chinese and Tibetan, we find their conceptualizations to be quite different, in particular, for body parts (e.g., `mouth` and `neck`). See §F for examples.

Conversely, we now present three examples of typologically distant languages that are still conceptually close. The first example is **Tagalog**. While Indo-European languages mostly occur in a relatively tight region of the figure, that region also contains many non-Indo-European languages. We hypothesize that Indo-European languages have influenced the conceptualization of other languages worldwide due to their widespread use, partly as part of colonization. One example is that the Tagalog words "dila" and "wika" mean both *tongue* and *language*, a conceptualization similar to Spanish ("lengua"), a language Tagalog was in close contact with for centuries. Standard Malay is typologically related to Tagalog, but its word for *tongue* "lidah", does not include the meaning *language*. This may contribute to Tagalog being conceptually more similar to Spanish on our measure than other Austronesian languages. **Plateau Malagasy**, an Austronesian language spoken in Madagascar, is conceptually similar to both far-away Austronesian languages like Hawaiian (reflecting its typology) as well as to geographically close, but typologically dissimilar Atlantic-Congo languages like Mwani and Koti. **Masana** is an Afro-Asiatic language spoken in Nigeria. It is conceptually close to the Atlantic-Congo languages Yoruba, Igbo and Twi, also spoken in and around Nigeria. Geographic proximity seems to boost conceptual similarity in these three cases. We leave further investigation of the hypothesis that Conceptualizer-based representations reveal historical interlanguage influences to future work.

6 Conclusion & Future Work

We propose Conceptualizer, a method that automatically aligns source-language concepts and target-language strings by creating a directed bipartite graph. We investigate the structure of such alignments for 83 focal concepts. Our extensive manual evaluation demonstrates good performance of Conceptualizer. We introduce the notion of crosslingual stability of a concept and show, using Conceptualizer, that concrete concepts are more stable across languages than abstract concepts. We also define conceptual similarity, a new measure of language similarity based on Conceptualizer representations. In our experiments, conceptual similarity gives results that partially agree with established measures like typological and areal similarity, but are complementary in that they isolate a single clearly defined dimension of similarity: the degree to which the conceptualization of two languages is similar.

In the future, we would like to improve the efficiency of Conceptualizer and, extending our work on a sample of 83 in this paper, apply it to all concepts that occur in PBC.

Limitations

The Conceptualizer we propose consists of two core steps, i.e., forward pass and backward pass. The forward pass identifies the most associated target-language strings for a focal concept. However, due to possible data sparsity of PBC in some low-resource languages and some cases of verse-level misalignment, χ^2 scores of the real translations can be indistinguishable compared with some other rare words that also occur in the same verses. Under such rare cases, Conceptualizer will not work well enough. In addition, the genre of PBC is limited to religion and therefore the diversity of the concepts across languages is largely influenced. Nevertheless, PBC, as far as we know, provides texts in the largest number of low-resource languages. PBC is thus a good fit for our goal.

In this work, we select 83 concepts, including the Swadesh32 and Bible51, representing a wide range of interesting crosslingual concepts. The runtime for computing the results for one concept in all languages is around 10 hours on average. The relatively long runtime, however, can prevent us from exploring more interesting concepts.

We find that the concreteness of a focal concept can be a contributor to the stability measure. As we use English as the source language for representing the focal concepts, we naturally resort to concreteness scores from English language ratings only. In addition, the analysis is carried out from an English perspective. Nevertheless, as we want to compare different languages, we have to use a unified source language. Theoretically, we can use any language as the source language and represent the concepts in that language. We therefore plan to use other languages, e.g., Chinese, or some low-resource languages, as the source language in future research.

Ethics Statement & Risks

In this work, we investigate the differences in conceptualization across 1,335 languages by aligning concepts in a parallel corpus. To this end, we propose Conceptualizer, a method that creates a directed bipartite alignment graph between source language concepts and sets of target language strings. The corpus we used, i.e., PBC, contains translations of the Bible in different languages (one language can have multiple editions). As far as we know, the corpus does not include any information that can be used to attribute to specific individuals. Therefore, we do not foresee any risks or potential ethical problems.

Acknowledgments

We would like to acknowledge Verena Blaschke for her valuable suggestions. We would also like to thank the reviewers for their positive and constructive feedback. This work was funded by the European Research Council (grant #740516).

References

Ehsaneddin Asgari and Hinrich Schütze. 2017. Past, present, future: A computational investigation of the typology of tense in 1000 languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 113–124, Copenhagen, Denmark. Association for Computational Linguistics.

Lera Boroditsky, Lauren A Schmidt, and Webb Phillips. 2003. Sex, syntax, and semantics. *Language in mind: Advances in the study of language and thought*, 22:61–79.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911.

Hal Daumé III. 2009. Non-parametric Bayesian areal linguistics. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 593–601, Boulder, Colorado. Association for Computational Linguistics.

Isabelle Dautriche, Kyle Mahowald, Edward Gibson, and Steven T Piantadosi. 2017. Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive science*, 41(8):2149–2169.

Guy Deutscher. 2010. *Through the language glass: Why the world looks different in other languages*. Metropolitan books.

Philipp Dufter, Mengjie Zhao, Martin Schmitt, Alexander Fraser, and Hinrich Schütze. 2018. Embedding learning through multilingual concept induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1520–1530, Melbourne, Australia. Association for Computational Linguistics.

Helge Dyvik. 2004. *Translations as semantic mirrors: from parallel corpus to wordnet*, pages 309 – 326. Brill, Leiden, The Netherlands.

Vyvyan Evans. 2006. *Cognitive linguistics*. Edinburgh University Press.

William A Foley. 1986. *The papuan languages of New Guinea*. Cambridge University Press.

Alexandre François. 2008. Semantic maps and the typology of colexification. *From polysemy to semantic change: Towards a typology of lexical semantic associations*, page 163.

Volker Gast and Maria Koptjevskaja-Tamm. 2018. The areal factor in lexical typology. *Trends in Linguistics Studies and Monographs*, 3.

Thanasis Georgakopoulos, Eitan Grossman, Dmitry Nikolaev, and Stéphane Polis. 2022. Universal and macro-areal patterns in the lexicon. *Linguistic Typology*, 26(2):439–487.

Ryan Georgi, Fei Xia, and William Lewis. 2010. Comparing language similarity across genetic and typologically-based groupings. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 385–393, Beijing, China. Coling 2010 Organizing Committee.

Cliff Goddard and Anna Wierzbicka. 2013. *Words and Meanings: Lexical Semantics Across Domains, Languages, and Cultures*. Oxford University Press.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2022. glottolog/glottolog: Glottolog database 4.7.

Anh Khoa Ngo Ho and François Yvon. 2019. Neural baselines for word alignment. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.

Joshua Conrad Jackson, Joseph Watts, Teague R Henry, Johann-Mattis List, Robert Forkel, Peter J Mucha, Simon J Greenhill, Russell D Gray, and Kristen A Lindquist. 2019. Emotion semantics show both cultural variation and universal structure. *Science*, 366(6472):1517–1522.

David Kamholz, Jonathan Pool, and Susan Colowick. 2014. PanLex: Building a resource for panlingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3145–3150, Reykjavik, Iceland. European Language Resources Association (ELRA).

Temuulen Khishigsuren, Gábor Bella, Thomas Brochhagen, Daariimaa Marav, Fausto Giunchiglia, and Khuyagbaatar Batsuren. 2022. Metonymy as a universal cognitive phenomenon: Evidence from multilingual lexicons. In *Proceedings of the 44th Annual Conference of the Cognitive Science Society*. Cognitive Science Society.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

Johann-Mattis List. 2018. Data underlying clics version 1.0.

Johann-Mattis List, Simon J Greenhill, Cormac Anderson, Thomas Mayer, Tiago Tresoldi, and Robert Forkel. 2018. Clics2: An improved database of cross-linguistic colexifications assembling lexical data with the help of cross-linguistic data formats. *Linguistic Typology*, 22(2):277–306.

Johann-Mattis List, Anselm Terhalle, and Matthias Urban. 2013. Using network approaches to enhance the analysis of cross-linguistic polysemies. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Short Papers*, pages 347–353, Potsdam, Germany. Association for Computational Linguistics.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163, Reykjavik, Iceland. European Language Resources Association (ELRA).

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Robert Östling. 2016. *The Lexical Typology of Semantic Shifts*, chapter Studying colexification through massively parallell corpora. De Gruyter.

Robert Östling and Murathan Kurfalı. 2023. Language embeddings sometimes contain typological generalizations. *arXiv preprint arXiv:2301.08115*.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with markov chain monte carlo. *The Prague Bulletin of Mathematical Linguistics*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Loïc-Michel Perrin. 2010. Polysemous qualities and universal networks, invariance and diversity. *Linguistic Discovery*, 8:1–22.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing. *Computational Linguistics*, 45(3):559–601.

Yael Ravin and Claudia Leacock. 2000. Polysemy: an overview. *Polysemy: Theoretical and computational approaches*, pages 1–29.

Christoph Rzymski, Tiago Tresoldi, Simon J Greenhill, Mei-Shin Wu, Nathanael E Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, Timotheus A Bodt, Abbie Hantgan, Gereon A Kaiping, et al. 2020. The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific data*, 7(1):1–12.

Nasaw Sampu, Wilai Jaseng, Thocha Jana, and Douglas Inglis. 2005. *A preliminary Ngochang-Kachin-English Lexicon*. Payap University, Chiang Mai.

Edward Sapir. 1912. Language and environment. *American anthropologist*, 14(2):226–242.

Lütfi Kerem Şenel, İhsan Utlu, Veysel Yücesoy, Aykut Koç, and Tolga Çukur. 2018. Semantic structure and interpretability of word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1769–1779.

Lütfi Kerem Şenel, Veysel Yücesoy, Aykut Koç, and Tolga Çukur. 2017. Measuring cross-lingual semantic similarity across european languages. In *2017 40th international conference on telecommunications and signal processing (TSP)*, pages 359–363. IEEE.

Silvia Severini, Ayyoob ImaniGooghari, Philipp Dufter, and Hinrich Schütze. 2022. Towards a broad coverage named entity resource: A data-efficient approach for many diverse languages. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3923–3933, Marseille, France. European Language Resources Association.

Pradeep Sopory and James Price Dillard. 2002. The persuasive effects of metaphor: A meta-analysis. *Human communication research*, 28(3):382–419.

Morris Swadesh. 2017. *The origin and diversification of language*. Routledge.

Bill Thompson, Seán G Roberts, and Gary Lupyan. 2020. Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour*, 4(10):1029–1038.

Annika Tjuka, Robert Forkel, and Johann-Mattis List. 2022. Linking norms, ratings, and relations of words and concepts across multiple language varieties. *Behavior research methods*, 54(2):864–884.

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

Leonie Weissweiler, Valentin Hofmann, Masoud Jalili Sabet, and Hinrich Schuetze. 2022. CaMEL: Case Marker Extraction without Labels. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5506–5516, Dublin, Ireland. Association for Computational Linguistics.

Di Wu, Liang Ding, Shuo Yang, and Mingyang Li. 2022. MirrorAlign: A super lightweight unsupervised word alignment model via cross-lingual contrastive learning. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 83–91, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? *arXiv preprint arXiv:2005.09093*.

Yang Xu, Barbara C Malt, and Mahesh Srinivasan. 2017. Evolution of word meanings through metaphorical mapping: Systematicity over the past millennium. *Cognitive psychology*, 96:41–53.

Thomas Zenkel, Joern Wuebker, and John DeNero. 2020. End-to-end neural word alignment outperforms GIZA++. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1605–1617, Online. Association for Computational Linguistics.

Kalliopi Zervanou, Elias Iosif, and Alexandros Potamianos. 2014. Word semantic similarity for morphologically rich languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

A Details of data

A.1 Parallel Bible Corpus

We work with the Parallel Bible Corpus (PBC, (Mayer and Cysouw, 2014)), which contains 1,775 editions of the Bible in 1,335 unique languages (we regard dialects with their own ISO 639-3 codes[3] as different languages). As far as we know, there is no explicit licence for PBC dataset. For each language, we only use one of its available editions. We use the *New World* edition for each language, if available, and the edition with the largest number of verses otherwise. Different from previous work (Asgari and Schütze, 2017; Dufter et al., 2018; Weissweiler et al., 2022) which only used verses that are available in all languages, we use all parallel verses between English and any other target languages. This means the number of parallel verses between English and other languages can be different. In general, we have parallel verses with a number greater than 30,000 (Hebrew + New Testament) for high-resource languages (141 languages), e.g. French, German and Chinese, while around 7,900 (New Testament only) for most of the other languages (1038 languages).

[3] https://iso639-3.sil.org/

A.2 Concept selection

We have 83 focal concepts in total which are listed in Tables 12 and 13. We classify the focal concepts into Swadesh32 and Bible51, for which we explain the selection of concepts in detail as follows.

The *Swadesh 100* (Swadesh, 2017) list offers 100 English words that represent universal and basic concepts. The words include nouns, adjectives, and verbs. We limit our selection to nouns to facilitate the comparison between concepts. Because we choose Bible editions as our resource, many concepts in the list can have very low frequencies of occurrences or even do not occur at all. On the contrary, some concepts in the list can occur many times but are less interesting to us, e.g., 'I', 'you', 'we', 'this' etc. We therefore only keep the concepts in the list which occur equal to or more than 5 times in the New Testament (only New Testament is available for most low-resource languages so the concept has to appear in the New Testament) and less than or equal to 500 times in Hebrew + New Testaments. There are 32 concepts that fulfill the criterion and we refer them to as **Swadesh32**, which are shown in Table 12.

For Bible concepts, we first obtain the distinct types of strings that have a length between 4 and 15 characters from all words in the English Bible. The ngrams can contain but cannot go beyond the word boundaries. For example, from a part of sentence `$bird$fly$` (substituting whitespaces with `$`), we can obtain ngrams such as `$bird`, `$bird$` and `fly$`, but `$bird$f` is not possible because it contains `$` in the middle. After this, we randomly select 10 languages from the available languages of PBC plus Chinese and German (12 languages in total). For each of the selected languages, we compute the coverage of the identified most associated target-language string obtained by performing a forward pass (setting the max number of iterations M to 1) (Algorithm 1) by regarding each English string as a focal concept. If the coverage of a string is larger than 0.5 for more than five languages, then we keep it otherwise we filter it out. This results in around 1000 strings. Then we filter out those that represent named entities. Finally, we manually check the list and select the strings that represent nouns and are not in the Swadesh list and are more or less specific to the Bible. This finally results in a set of 51 concepts (**Bible51**) which are shown in Table 13.

B Additional details of Conceptualizer

B.1 Focal concepts & strings

The bipartite graph $\mathcal{G}$ we construct contains source nodes set $\mathcal{S}$ and target nodes set $\mathcal{T}$. Each node $s \in \mathcal{S}$ is a concept and is represented by a set of strings. We restrict the length of the strings between 1 and 8 for any language except the source language English (we restrict the length larger than 2 and the strings cannot go beyond word boundaries) for efficiency. To differentiate the nodes in $\mathcal{S}$, we refer to the set of our chosen 83 focal concepts as $\mathcal{S}_F \subset \mathcal{S}$. Each focal concept F in $\mathcal{S}_F$ is a set which can contain multiple strings, e.g., 'belly' concept: {`$belly$`, `$bellies$`}. In contrast, other concepts in $\mathcal{S}$ are sets which contain only a single string, e.g., {`$sparrows$`}. In the backward pass for a focal concept F, if a string s being identified belongs to F, then we create an edge that ends at F instead of s.

B.2 String candidates

The Conceptualizer consists of (1) a forward pass: for building edges $(F, c) \in \mathcal{S} \times \mathcal{T}$ and (2) a backward pass for building edges $(c, f) \in \mathcal{T} \times \mathcal{S}$. In the forward pass, for example, an edge $(F, c) \in \mathcal{S} \times \mathcal{T}$ is constructed if a target node from the target language l: $c =< l, v, T >$'s verse v contains a string t that is highly associated with F. As the search space of target-language strings is extremely large for each l, we therefore restrict the search space to the set of strings which occur in the verses whose corresponding English verses contain F. Formally, let $\Pi(\text{eng}, F)$ be the verses where the focal concept F occurs and $\mathcal{P}(\Sigma^*)(l, v)$ be the strings that fulfill our string selection conditions in verse v for language l. We will then only consider the strings in $\mathbb{T} = \bigcup_v \mathcal{P}(\Sigma^*)(l, v) | v \in \Pi(\text{eng}, F)$ to be candidates in language l that are possibly associated with F. Similarly, in the backward pass, we also restrict the search space to be the set of English strings in the verses whose corresponding target-language verses contained the identified target-language string set T in the forward pass. Formally, let $\Pi(l, T)$ be the verses where the T occurs and $\mathcal{P}(\Sigma^*)(\text{eng}, v)$ be the strings that fulfill our string selection conditions in verse v for English. We will then only consider English strings in $\mathbb{S} = \bigcup_v \mathcal{P}(\Sigma^*)(\text{eng}, v) | v \in \Pi(l, T)$. Furthermore, we will only consider the strings $t \in \mathbb{T}$ which occur more than $\Pi(\text{eng}, F)/10$ in the target-language verses of $\Pi(\text{eng}, F)$ and $s \in \mathbb{S}$ which occur more

	in $\Pi(l_2, U)$	in $\neg\Pi(l_2, U)$
$I(t)$	n_{00}	n_{01}
$\neg I(t)$	n_{10}	n_{11}

Table 7: The contingency table for a given string t from language l_1. $I(t)$ (resp. $\neg I(t)$) denotes the string occurs (resp. does not occur) in the corresponding parallel verses for l_1. For example, n_{00} denotes the number of verses in $\Pi(l_2, U)$ in which t occurs.

than 2 times in the English verses of $\Pi(l, T)$.

B.3 Measuring association

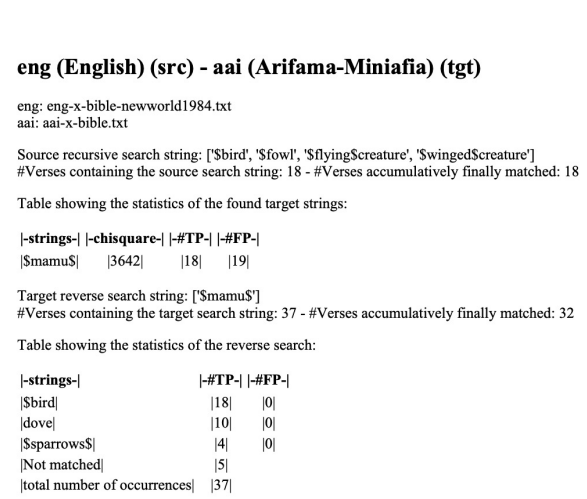
Given a set of strings U in a language l_2, we want to find a string in language l_1 that is associated with U. To this end, we use χ^2 score to measure the degree of the association. Specifically, we divide the verse set into two subsets: verses containing U and verses not containing U in the Bible of l_2, i.e., $\Pi(l_2, U)$ and $\neg\Pi(l_2, U)$. We then build a contingency table for each string candidate $t \in \mathcal{P}(\Sigma^*)$ and t comes from language l_1, as shown in Table 7. After that, we compute the χ^2 score for each string: the higher the score, the more associated the string in l_1 is with the set of strings U. We then choose the string that has the highest χ^2 score as a hit in language l_1 for a set of strings U in language l_2.

B.4 Adding edges

For efficiency and stability reasons, in the actual implementation of the backward pass, Conceptualizer does not add edges from a target node to all the strings in S that fulfill the criterion, i.e., the set of the identified associated source language strings, as shown in Algorithm 2 (line 9). Instead, we only add one edge only from each target node. This means that in each iteration of the backward pass, we will add new edges starting from the involved target nodes to a single s: $\{((l, v, T), s) | (F, (l, v, T)) \in \mathcal{G} \wedge s \in v \wedge v \in V\}$. In this way, each target node that was previously connected with the focal concept F can only be connected to one source node only. By doing this, we find that some undesirable associations can be avoided.

B.5 Hyperparameters

We have two hyperparameters in Conceptualizer, i.e., (1) the maximum number of iterations of searching associated strings for a focal concept in each language: M and (2) the threshold α for the minimum coverage of the set of identified associated string U. We set $M = 5$ and $\alpha = .9$ as



eng (English) (src) - aai (Arifama-Miniafia) (tgt)

eng: eng-x-bible-newworld1984.txt
aai: aai-x-bible.txt

Source recursive search string: ['$bird', '$fowl', '$flying$creature', '$winged$creature']
#Verses containing the source search string: 18 - #Verses accumulatively finally matched: 18

Table showing the statistics of the found target strings:

|-strings-|	**|-chisquare-|**	**|-#TP-|**	**|-#FP-|**
|$mamu$|	|3642|	|18|	|19|

Target reverse search string: ['$mamu$']
#Verses containing the target search string: 37 - #Verses accumulatively finally matched: 32

Table showing the statistics of the reverse search:

|-strings-|	**|-#TP-|**	**|-#FP-|**
|$bird|	|18|	|0|
|dove|	|10|	|0|
|$sparrows$|	|4|	|0|
|Not matched|	|5|	
|total number of occurrences|	|37|	

Figure 6: Part of the document for annotation for one target language, i.e., aai (Arifama-Miniafia) The true positive and false positive verses are not shown here.

default values for all involved computations. Based on preliminary experiments on 'bird' concept, we found that the number of associated strings usually will not go beyond 5. Moreover, when M is large, we might have an efficiency problem (more iterations for each language) when computing other focal concepts. Therefore, we set $M = 5$ to reduce the runtime of Conceptualizer while not sacrificing the accuracy too much. As for the coverage threshold α, we conduct preliminary experiments with .85, .9 and .95 respectively on 'bird' concept. We found when coverage is small (<.9), the search stops when there are still possibly unidentified associated strings in the rest verses. If the coverage is too large (>.9), we found that some less related strings can be identified at the later stage of iterations for some languages. We should note that PBC can have verse-level misalignment problems, which means for some parallel verses 'bird' occurs in English but the target-language verse can be unrelated to 'bird' at all. Moreover, as we remove the parallel verses that we have covered in each iteration, the verses uncovered become smaller and smaller in each iteration. χ^2 scores computed on later iterations can be not significant and multiple strings can have the same highest χ^2 scores if they only appear in the uncovered verses with the same number of occurrences. Therefore, to ensure that Conceptualizer finds enough strings while guaranteeing the quality of the associations between them with the search string, we set the coverage threshold $\alpha = .9$.

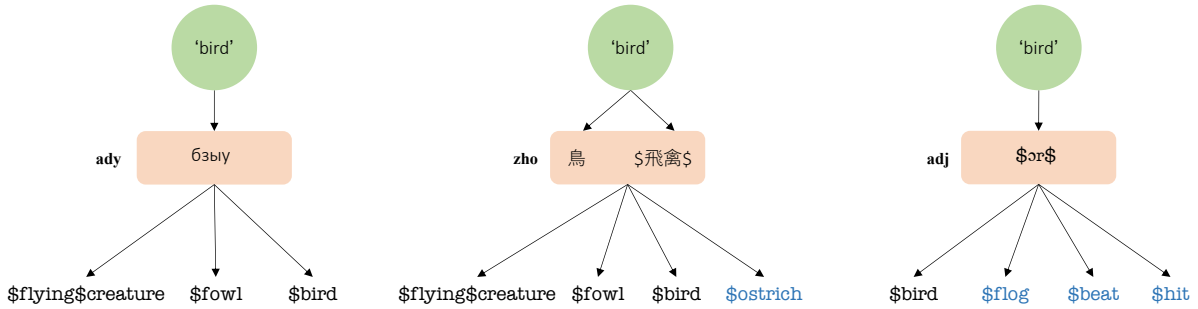


Figure 7: Examples of languages identified as **one-to-one**, **polysemy** and **ambiguity** respectively. The strings that do not belong to the set of strings representing the focal concept 'bird' are marked with a different color. For Chinese, we have a BP hit of 'ostrich', a hyponym of 'bird'. For Adioukrou, we do not have BP hits that are closely related to 'bird' but we have hits that are related to another cluster of meanings, i.e., *hit*.

C Details of Evaluation

C.1 Single concept across all languages

In the English version of the Bible, we find the 'bird' concept is often expressed by the following words/phrases: "bird(s)", "fowl(s)", "flying creature(s)" and "winged creature(s)". Therefore, we use the following strings: `$bird`, `$fowl`, `$flying$creature` and `$winged$creature` to represent the 'bird' concept. We agree that this set of strings might not be optimal for other large-resource parallel datasets (if there are any). However, for PBC dataset, this set of strings could empirically cover the 'bird' concept. We perform our Conceptualizer which includes the forward pass and backward pass for this 'bird' concept for all the languages in PBC. We perform three types of evaluations as follows based on the FP and BP hits:

C.1.1 'bird' conceptualization in all languages

We provide a pdf document in which the statistics of each string identified in both forward pass and backward pass are shown. In addition, for each target-language string, a) two randomly sampled **True Positive** parallel verses, i.e., target-language verses that contain the identified strings and the parallel English verses contain 'bird'; b) two randomly sampled **False Positive** parallel verses, i.e., target-language verses that contain the strings and the parallel English verses that do not contain 'bird', are shown. We also show three randomly sampled **False Negative** parallel verses, i.e., target-language verses that do not contain the strings and the parallel English verses that contain 'bird' concept. An example of part of the document for one language is shown in Figure 6. By checking the general patterns demonstrated in the document, we define four evaluation categories: **one-to-one**, **polysemy**, **ambiguity** and **failure**. Noticeably, our category **polysemy** and **ambiguity** do not directly correspond to the definition in linguistics, but reflect general patterns of the conceptualization. The classification of these two categories is based on the pattern of strings identified in the backward pass. More specifically, we classify the conceptualization pattern of a language as **polysemy**, if it shows one of the following patterns: (a) *hyponymy*, where we found strings such as `dove` and `sparrows`, which are hyponyms of 'bird'. (b) *meronymy*, where we found strings like `$wings`, which is a meronym of 'bird'. (c) *other related words*, where we found strings such as `$fly` and `$chirp`, which are apparently related to 'bird', but do not fit into any well-defined lexical semantic relation. The conceptualization pattern is classified as **ambiguity**, if the strings we found in the backward pass are not semantically related to 'bird' at all (such as `$new$` and `$kid$`), but nevertheless are deemed as highly associated with 'bird' by our algorithm. These cases are generally caused by having homonyms of 'bird' in the target language. In case the linguist annotator cannot be sure of the classification, consultation has been made with other experts to resolve these issues and find the most common agreement.

l	*P*	*T*	category
ify	ke:keq, qemayuq	`$sisit$`	**no overlap**
ind	cewek, burung	`$burung, terbang$`	**overlap**
akb	unggas	`$unggas$`	**match**
afr	voël, vliegtuig	`voël`	**match**

Table 8: Examples of languages that are classified into categories **no overlap**, **overlap** and **match** respectively for PanLex annotation.

l	P	T	factor
uig	qush	куш (qush)	**script**
mua	ʒùː	juu	**transcription**
lip	ɔklɔbɛ	baklɔbɛ	**prefix**
mse	layra	layagi	**suffix**
sbl	ma‘nok	manokmanok	**reduplication**

Table 9: Example of PanLex translations that have different forms caused by non-lexical factors

model	Coverage(g)	Coverage(a)	trans. per l
Conceptualizer	.93	.95	1.6
Eflomal 0	.94	.96	7.0
Eflomal 1	.81	.82	2.9
Eflomal 0.1	.70	.79	2.1

Table 10: Coverage & the number of translations proposed per language on average of Conceptualizer and eflomal (Östling and Tiedemann, 2016) on ‘bird’. Coverage (g) (resp. (a)) denotes the coverage computed globally for all languages (resp. computed for each language separately and then averaged over all languages). Elfomal 1 (resp. 0.1) denotes filtering the translations proposed whose frequency <= 1 (resp. <= 10% of the number of verses containing ‘bird’).

C.1.2 Translations compared with PanLex

For each language, the linguist annotator also checks the translation of “bird” provided by PanLex (Kamholz et al., 2014)[4]. The translations are available in 1,137 languages out of the 1,331 languages in PBC where we found translations of the focal concept ‘bird’. We define the following four categories for the PanLex evaluation where P are the tranlations from PanLex and T are the FP hits (target-language strings).

- **no translation**: $P = \emptyset$, i.e., PanLex gives no translation.
- **no overlap**: $T \cap P = \emptyset$, i.e., none of the FP hits is found in PanLex translations.
- **overlap**: $0 < |T \cap P| < |T|$, i.e., some but not all of the FP hits are found in PanLex translations.
- **match**: $|T \cap P| = |T|$, i.e., all the FP hits can be found in PanLex translations. Note that we do not require all the translations in PanLex to be present in our set of target strings, since PanLex often gives a very long list of translations and our goal is to use PanLex translations to confirm the strings we identified.

We show examples for each category (except for **no translation**) in Table 8. When deciding whether a translation from PanLex matches an FP hit, the linguist annotator does not only look for an exact match of strings but also takes the differences in scripts, transcriptions, and morphological forms into consideration. For languages with multiple writing systems, words are naturally transliterated into a unified script for them to be comparable. It has also been observed, that the same word can sometimes be transcribed differently in different sources, especially for low-resource languages that have no standard writing system. Therefore, “ʒùː” in PanLex and “juu” in our FP hit will still be considered as a match. Furthermore, it is also possible that the PanLex translation uses a morphologically different form compared to our FP hit, such as dictionary form versus inflected form. Possible morphological processes such as affixation, reduplication, vowel mutation, vowel ellipsis, and metathesis have been taken into consideration. We show some examples of these factors in Table 9.

With careful examination, if the linguist annotator concludes that Conceptualizer actually has found the same lexeme as the PanLex translation, and the difference between PanLex and our FP hit is merely attributed to the above-mentioned factors, they will still be considered as matches.

C.1.3 Translations compared with Eflomal

We also compare the coverage and the average number of translations proposed per language of Conceptualizer and eflomal (Östling and Tiedemann, 2016), a statistical word alignment tool, in Table 10. We collect words that are aligned with one of the strings representing ‘bird’ in all verses where ‘bird’ occurs for eflomal baselines. The coverage means the fraction of verses containing ‘bird’ covered by the set of proposed translations. Global coverage denotes that we compute the coverage directly in all verses regardless of language while average coverage denotes that we first compute the coverage for each language and then average over all languages. We notice that, eflomal, without filtering any translations, obtains the highest global and average coverage, which can be regarded as the upper bound. However, the number of translations per language on average is so high: 7.0. After filtering some proposed translations by their frequencies (1 and 0.1), we observe a sudden drop in the coverage. This indicates that (1) eflomal can propose many wrong alignments and (2) some correct alignments have very small frequencies. Because of its word-level alignment nature, eflomal cannot take the possible morphological changes of

[4]https://translate.panlex.org/

the words into consideration at all. On the contrary, Conceptualizer only proposes 1.6 translations on average while keeping the coverage very close to the upper bound, suggesting that Conceptualizer can identify the strings (ngrams) that are most associated with the concept and alleviate the possible problems caused by, e.g., morphological changes, in many languages.

C.2 Swadesh concepts

We resort to NoRaRe (Tjuka et al., 2022)[5] to find the available translations of the 32 Swadesh concepts in 39 languages (NoRaRe covers). For each concept and each language, we store a triple indexed by the concept-language pair:

$$< \text{concept}, \text{language}, \text{translation(s)} >$$

We finally obtain 582 triples for evaluation (NoRaRe does not provide translations of a concept in all covered languages). We use T to represent the set of translations proposed by Conceptualizer and N for NoRaRe translations (as ground-truth translations) in the triple. When judging whether a translation in N matches a translation in T generated by Conceptualizer, we do the match leniently to allow for morphological changes. Specifically, if a translation in N is a substring of a translation in T generated by Conceptualizer or the other way around, we regard it as a successful match. This is because N often provide the dictionary forms of the nouns but T are generated automatically based on the actual Bible verses where the nouns can change their suffixes or prefixes quite often depending on their roles in the verses.

We are especially interested in the triples in which our identified strings T do not match the ground-truth translations R, i.e., $T \cap R = \emptyset$. We sampled 10 such triples (we provide N and T for each triple) in Table 11. We notice that there are cases where the ground-truth translations N use different versions of transliterations. For example, `$andjing` vs. "anjing", `$daoen` vs. "daun" and `$boelan` vs. "bulan" in Malay (msa). Moreover, there can be multiple equivalent translations for a concept, but N just lists one of them which is not used (or not identified) in PBC, e.g., "颈" is a simpler but more formal translation of 'neck' but the N only lists "脖子" in Chinese (zho); the concept 'path' can also be translated to "rout" and "sentier" in French (fra) but only "chemin" is given by the N. Therefore, we see that this evaluation compared with NoRaRe can actually underestimate the performance of our method.

concept	l	N	T
'dog'	msa	anjing	`$andjing`
'seed'	cym	hedyn	`$had$`, `$heu`
'leaf'	msa	daun	`$daoen`
'horn'	est	ruupor	`sarve`
'mouth'	tur	ağiz	`$ağzı`, `$ağız`
'neck'	zho	脖子	颈
'moon'	msa	bulan	`$boelan`
'water'	cym	dŵr	`dwfr$`, `dyfr`
'rain'	msa	hujan	`$hoedjan`
'path'	fra	chemin	`$la$rout`, `$sentier`

Table 11: 10 randomly selected examples of triples for which we obtain a score of 0 in the "partial" setting in Table 3.

D Infrastructure & environment

We ran all our computational experiments on a CPU server with 48 cores and 1024 GB of memory. We used Python 3.6[6] throughout our implementation of Conceptualizer and for visualizations. Specifically, for fundamental scientific computing (e.g., computing χ^2 scores), we used NumPy[7], SciPy[8] and scikit-learn[9] packages. For visualization, we used NetworkX[10](mainly for the crosslingual semantic fields) and Matplotlib[11] packages.

[5] https://norare.clld.org/

[6] https://www.python.org/

[7] https://numpy.org/

[8] https://scipy.org/

[9] https://scikit-learn.org/stable/

[10] https://networkx.org/

[11] https://matplotlib.org/

Concepts	Concreteness	Stability
‘fish’	5.0	0.86
‘bird’	5.0	0.68
‘dog’	4.85	0.85
‘tree’	5.0	0.64
‘seed’	4.71	0.38
‘leaf’	5.0	0.74
‘root’	4.34	0.78
‘flesh’	4.59	0.36
‘blood’	4.86	0.69
‘horn’	5.0	0.82
‘hair’	4.97	0.77
‘ear’	5.0	0.46
‘mouth’	4.74	0.49
‘tooth’	4.89	0.91
‘tongue’	4.93	0.61
‘foot’	4.9	0.7
‘knee’	5.0	0.38
‘belly’	4.8	0.4
‘neck’	5.0	0.72
‘breast’	4.89	0.65
‘sun’	4.83	0.49
‘moon’	4.9	0.48
‘star’	4.69	0.87
‘water’	5.0	0.48
‘rain’	4.97	0.68
‘stone’	4.72	0.71
‘cloud’	4.54	0.68
‘smoke’	4.96	0.57
‘path’	4.41	0.35
‘mountain’	4.96	0.64
‘white’	3.89	0.77
‘night’	4.52	0.68

Table 12: The concreteness and stability measure of our considered focal concepts: Swadesh32. Concreteness is from (Brysbaert et al., 2014) while stability is computed using the statistics obtained by Conceptualizer. If the concreteness is NA, it means the concept is not included in the resource.

Concepts	Concreteness	Stability
‘babe’	3.67	0.59
‘hypocrit’	2.43	0.81
‘soldier’	4.72	0.49
‘scroll’	4.11	0.57
‘demon’	3.32	0.45
‘boat’	4.93	0.71
‘olive’	4.9	0.8
‘prayer’	3.28	0.32
‘mercy’	1.57	0.29
‘trumpet’	4.86	0.83
‘angel’	3.82	0.88
‘prison’	4.68	0.62
‘savior’	3.04	0.49
‘tomb’	4.73	0.61
‘husband’	4.11	0.47
‘bride’	4.63	0.69
‘talent’	2.19	0.83
‘peace’	1.62	0.72
‘secret’	2.19	0.57
‘faith’	1.63	0.59
‘woe’	1.96	0.8
‘throne’	4.64	0.62
‘wisdom’	1.53	0.54
‘disciple’	3.29	0.73
‘obeisance’	NA	0.37
‘truth’	1.96	0.4
‘memor’	2.83	0.53
‘governor’	4.07	0.52
‘poor’	2.7	0.63
‘blind’	4.03	0.77
‘spiritual’	1.79	0.33
‘justice’	1.45	0.34
‘courage’	1.52	0.53
‘purpose’	1.52	0.3
‘generation’	1.96	0.56
‘contrary’	1.56	0.46
‘prophesy’	2.11	0.41
‘decision’	2.19	0.36
‘request’	2.59	0.32
‘weakness’	2.59	0.55
‘journey’	2.57	0.39
‘public’	2.57	0.23
‘appearance’	2.57	0.55
‘expression’	2.54	0.51
‘marriage’	2.51	0.51
‘wrath’	2.42	0.4
‘trouble’	2.25	0.45
‘promise’	2.09	0.46
‘power’	2.04	0.41
‘pleasure’	2.04	0.35
‘thought’	1.97	0.39

Table 13: The concreteness and stability measure of our considered focal concepts: Bible51. Concreteness is from (Brysbaert et al., 2014) while stability is computed using the statistics obtained by Conceptualizer. If the concreteness is NA, it means the concept is not included in the resource.

E Crosslingual semantic fields

F Further analysis regarding language similarity

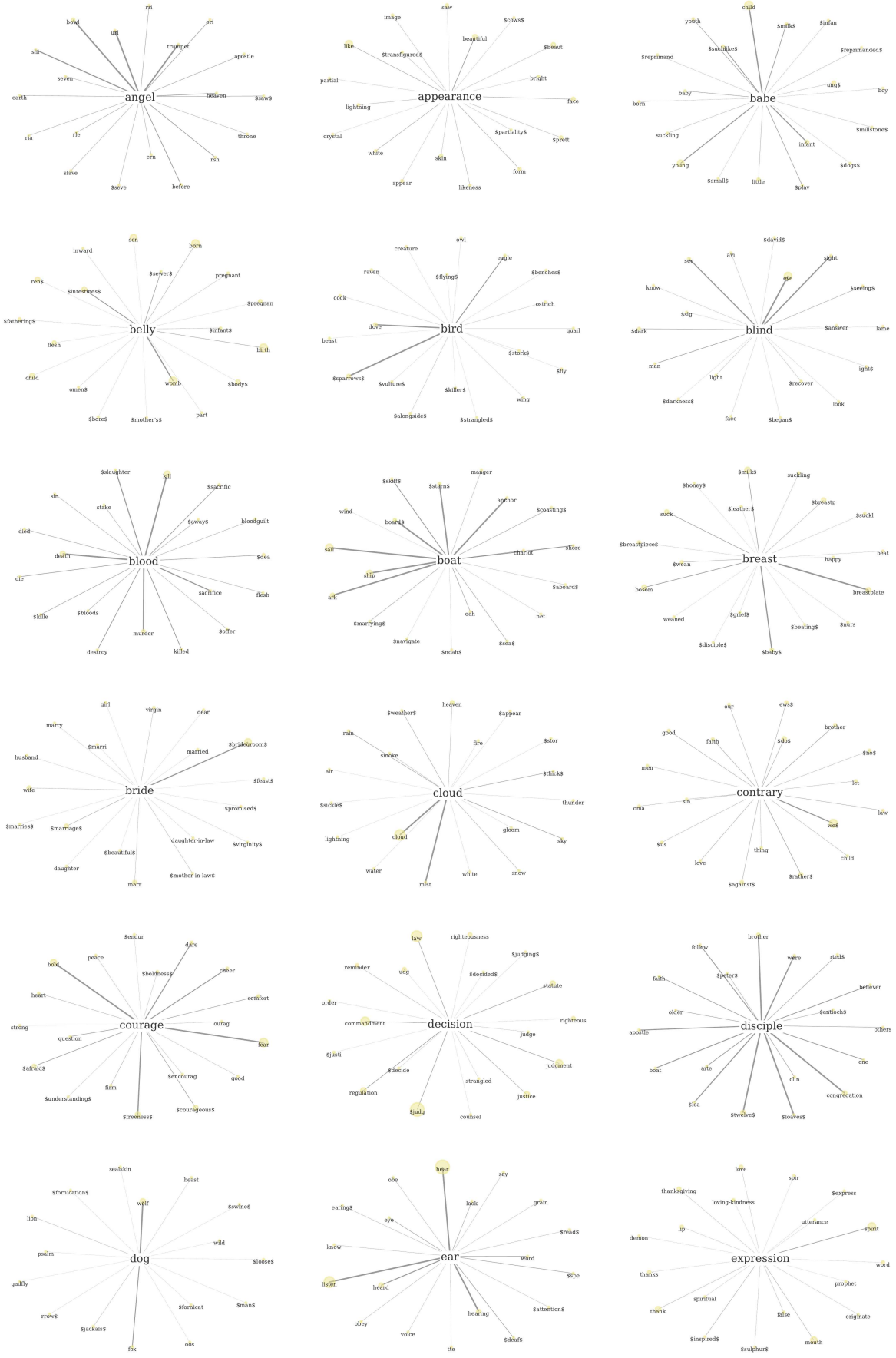


Figure 8: Visualization of semantic field (1).

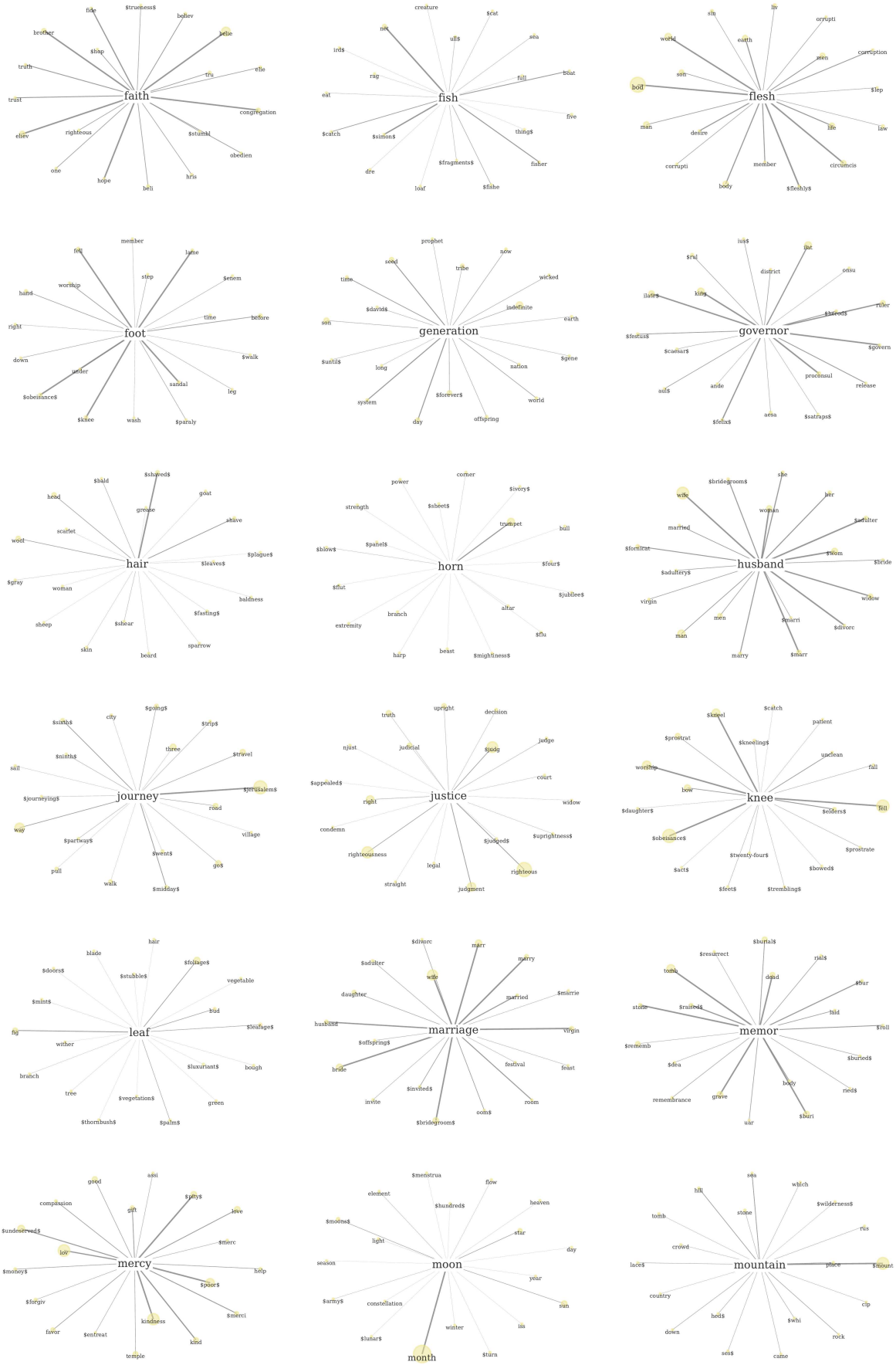


Figure 9: Visualization of semantic field (2).

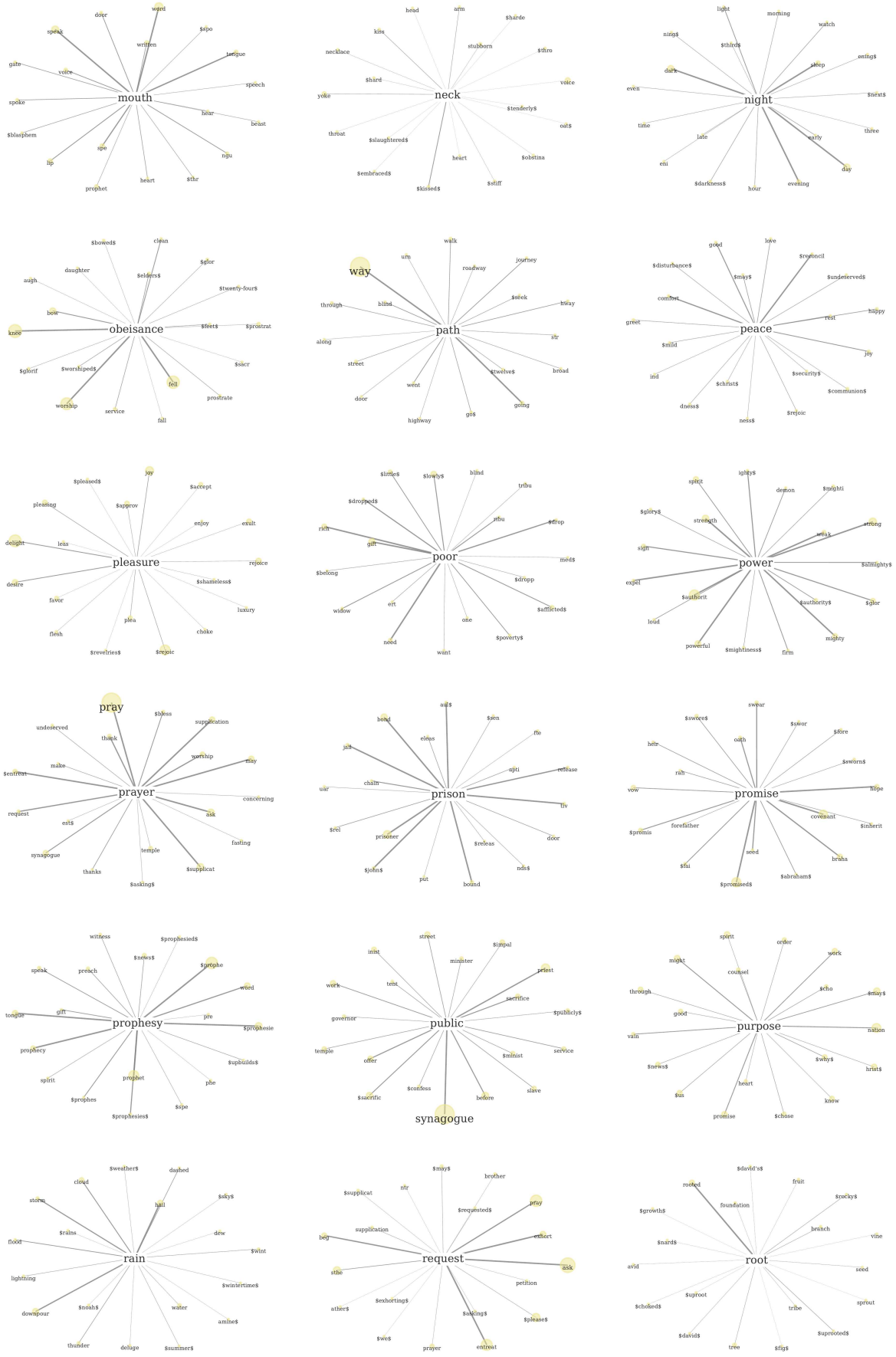


Figure 10: Visualization of semantic field (3).

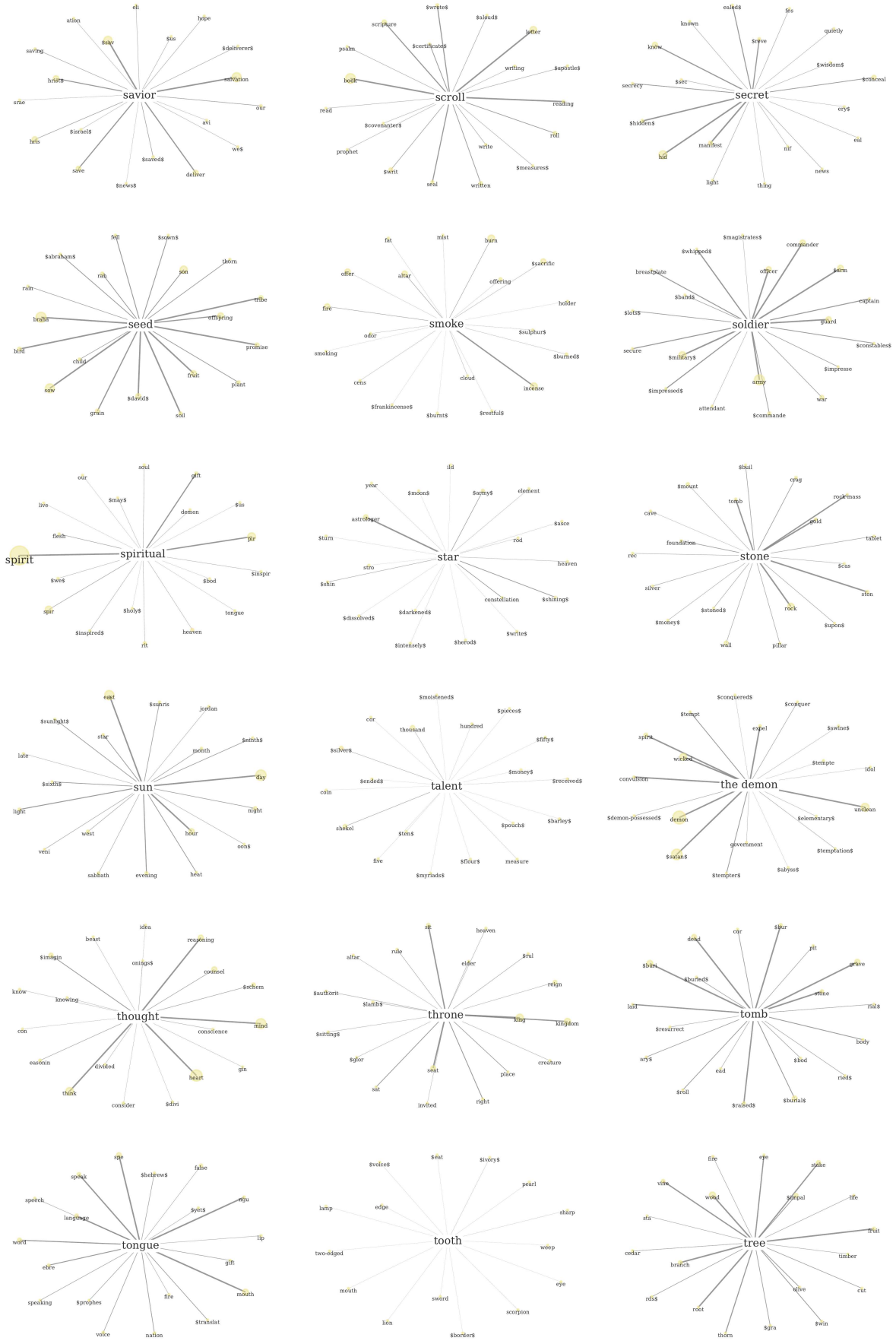


Figure 11: Visualization of semantic field (4).

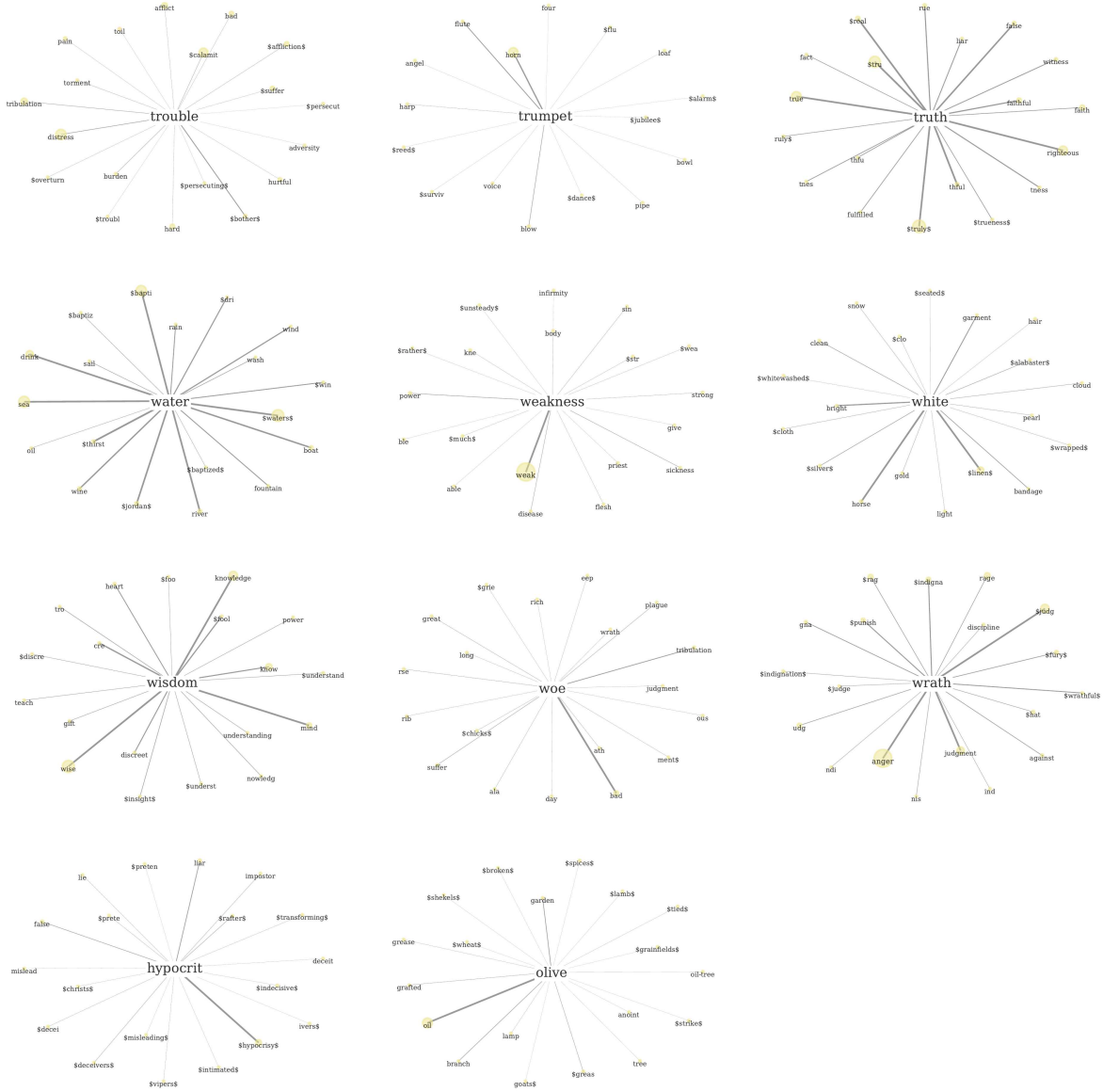
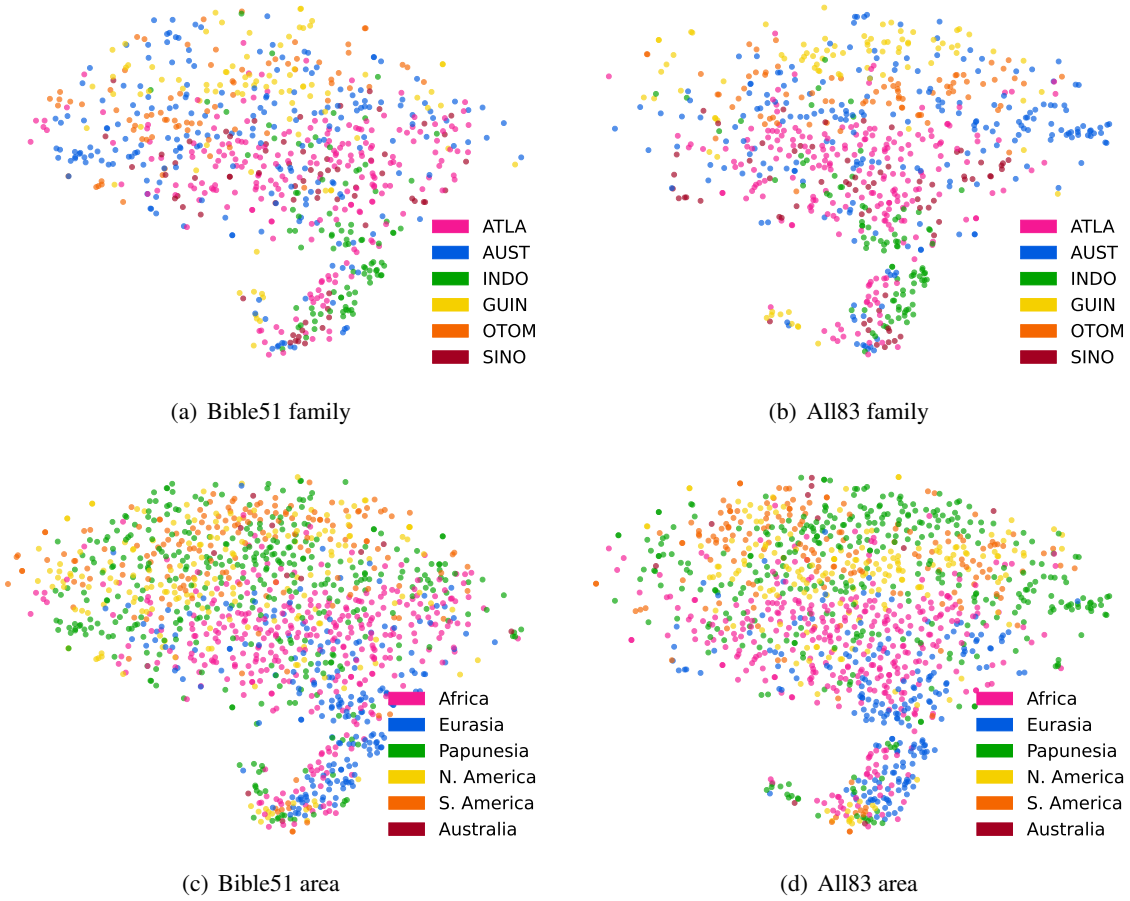


Figure 12: Visualization of semantic field (5).

setting	global	atla	aust	indo	guin	otom	sino
Swadesh32	1280	222	215	91	87	76	68
Bible51	1264	219	210	90	85	76	67
All83	1263	219	210	90	85	76	67

Table 14: Numbers of languages in the 6 families with over 50 languages. The numbers vary across the 3 concatenation settings as we only use languages for which all the considered concepts are available. Thus, **Swadesh32** has the most languages, and the number of languages decreases as we increase the number of concepts. We find that while both the Austronesian and Sino-Tibetan families are spread out, the binary classification result of Austronesian languages is much better since the large geographical span is compensated by its higher number of languages.



(a) Bible51 family (b) All83 family

(c) Bible51 area (d) All83 area

Figure 13: t-SNE plot of the languages using **Swadesh32** and **All83** concatenation of concepts. The colors indicate different language families in the upper subfigure and different areas in the lower subfigure.

#neighbors	global	Papunesia	Africa	Eurasia	North America	South America	Australia
k=1	0.51	0.51	0.52	0.64	0.52	0.35	0.00
k=2	0.28	0.27	0.31	0.47	0.27	0.09	0.00
k=3	0.45	0.44	0.51	0.62	0.43	0.22	0.00
k=4	0.53	0.52	0.60	0.69	0.51	0.25	0.00
k=5	0.55	0.54	0.61	0.72	0.57	0.28	0.09
k=6	0.56	0.55	0.62	0.75	0.57	0.26	0.09
k=7	0.56	0.55	0.63	0.74	0.56	0.24	0.09
k=8	0.57	0.56	0.64	0.76	0.56	0.27	0.00
k=9	0.57	0.54	0.66	0.77	0.59	0.26	0.00
k=10	0.58	0.55	0.65	0.79	0.62	0.22	0.00

Table 15: The result of binary classification (area) results using different numbers of nearest neighbors (1 to 10) of **Swadesh32**. The global column shows the results considering all languages. The rest columns denote the results only considering languages in that region.

#neighbors	global	Papunesia	Africa	Eurasia	North America	South America	Australia
k=1	0.49	0.54	0.57	0.63	0.28	0.30	0.00
k=2	0.24	0.25	0.33	0.42	0.06	0.08	0.00
k=3	0.41	0.43	0.59	0.62	0.13	0.09	0.00
k=4	0.48	0.52	0.66	0.70	0.20	0.14	0.00
k=5	0.51	0.55	0.70	0.72	0.18	0.18	0.00
k=6	0.50	0.53	0.72	0.70	0.18	0.15	0.00
k=7	0.50	0.53	0.71	0.70	0.18	0.16	0.00
k=8	0.49	0.51	0.71	0.69	0.14	0.15	0.00
k=9	0.48	0.48	0.73	0.71	0.12	0.12	0.00
k=10	0.49	0.50	0.74	0.71	0.14	0.12	0.00

Table 16: The result of binary classification (area) results using different numbers of nearest neighbors (1 to 10) of **Bible51**. The global column shows the results considering all languages. The rest columns denote the results only considering languages in that region.

#neighbors	global	Papunesia	Africa	Eurasia	North America	South America	Australia
k=1	0.54	0.59	0.65	0.68	0.38	0.25	0.00
k=2	0.33	0.35	0.41	0.53	0.16	0.09	0.00
k=3	0.51	0.51	0.69	0.71	0.27	0.17	0.00
k=4	0.56	0.57	0.75	0.75	0.32	0.21	0.00
k=5	0.58	0.59	0.77	0.78	0.34	0.22	0.00
k=6	0.58	0.60	0.77	0.79	0.35	0.17	0.00
k=7	0.58	0.60	0.79	0.81	0.34	0.15	0.00
k=8	0.57	0.59	0.77	0.81	0.33	0.16	0.00
k=9	0.58	0.59	0.78	0.81	0.32	0.18	0.00
k=10	0.58	0.60	0.79	0.80	0.32	0.17	0.00

Table 17: The result of binary classification (area) results using different numbers of nearest neighbors (1 to 10) of **All83**. The global column shows the results considering all languages. The rest columns denote the results only considering languages in that region.

concept	lang.	ngrams
'mouth'	zho	`$mouth$`, `$mouths$`, `$entrance`, `alat`, `$ford`
	bod	`$mouth$`, `$mouths$`, `$prais`
'neck'	zho	`$neck$`, `$necks$`, `$stiff-necked`
	bod	`$neck$`, `$necks$`, `$obstina`, `$stubbornness$`
'tree'	zho	`$tree$`, `$trees$`, `$vine$`, `-tree$`, `$boughs$`
	bod	`$tree$`, `$trees$`, `$chariot`, `wood`, `$cedar`, `igs$`, `$timber`
'horn'	zho	`$horn$`, `$horns$`, `$corner`
	bod	`$horn$`, `$horns$`, `$trumpet`, `$fathering$`

Table 18: Selected examples from the comparison of **Swadesh32** concepts of Mandarin Chinese (zho) and Tibetan (bod). Differences can be observed for several concepts, especially those related to body parts.

concept	lang.	ngrams
	arb	$fish$, $fishes$
	pes	$fishes$, $fish$, $fish
‘fish’	tur	$fish$, $fishes$, $fish, $honey
	msa	$fishes$, $fish$
	ind	$fish$, $fishes$
	arb	$stars$, $star$
	pes	$stars$, $star$
‘star’	tur	$stars$, $star$
	msa	$stars$, $star$
	ind	$stars$, $star$
	arb	$blood$, $bloodguilt$
	pes	$blood$
‘blood’	tur	$blood$, $bloods, $bloodguilt$
	msa	$blood$
	ind	$blood$, $blood
	arb	$tongue$, $tongues$
	pes	$tongue$, $tongues$
‘tongue’	tur	$tongue$, $tongues$, $request, $language
	msa	$tongues$, $tongue$, ebrew$
	ind	$tongue$, $tongues$, $language
	arb	$birds$, $bird$, $flying$, $fowls$
	pes	$birds$, $bird$
‘bird’	tur	$birds$, $bird$, $fowl
	msa	$birds$, $bird$, dove, $sparrows$, $eagle
	ind	$birds$, $bird$, $eagle, $turtledove, $ostrich, $raven, $bird

Table 19: Selected examples from the comparison of **Swadesh32** concepts of several languages influenced by Islam. arb: Standard Arabic, pes: Western Farsi, tur: Turkish, msa: Standard Malay, ind: Standard Indonesian.

concept	lang.	ngrams
‘blood’	eng	$blood$
	spa	$blood$
	ell	$blood$
	rus	$blood$, $blood
	tgl	$blood$
	swh	$blood$, $blood
	hye	$blood$, $blood
‘tongue’	eng	$tongue$, $tongues$
	spa	$tongue$, $tongues$, $language
	ell	$tongue$, $tongues$, $language
	rus	$tongue$, $tongues$, $language
	tgl	$tongue$, $tongues$, $language
	swh	$tongue$, $tongues$, $language
	hye	$tongue$, $tongues$, $language
‘bird’	eng	$birds$, $bird$
	spa	$birds$, $bird$, $fowls$
	ell	$birds$, $bird$, $bird
	rus	$birds$, $bird$, $fowl, $bird
	tgl	$birds$, $bird$, $fowl
	swh	$birds$, $bird$, $fowl
	hye	$birds$, $bird$, $flying$, $fowl

Table 20: Selected examples from the comparison of **Swadesh32** concepts of several languages influenced by Christianity. eng: English, spa: Spanish, ell: Modern Greek, rus: Russian, tgl: Tagalog, swh: Swahili, hye: Eastern Armenian.

concept	lang.	ngrams
‘blood’	eng	$blood$
	deu	$blood$
	fra	$blood$
	jpn	$blood$, $blood
	kor	$blood$, $escap, $skin$, $airs$, $pip, $flut
	zho	$blood$, $blood
‘tongue’	eng	$tongue$, $tongues$
	deu	$tongue$, $tongues$
	fra	$tongue$, $tongues$, $language
	jpn	$tongue$, $tongues$
	kor	$tongue$, $tongues$, $language
	zho	$tongue$, $tongues$, $language
‘mouth’	eng	$mouth$, $mouths$
	deu	$mouth$, $mouths$
	fra	$mouth$, $mouths$
	jpn	$mouth$, $mouths$, $entrance, $kiss, $whistl, $doorkeeper, $contention
	kor	$mouth$, $mouths$, $entrance, $lip, $clothe, $kiss, $overla
	zho	$mouth$, $mouths$, $entrance, alat, $ford

Table 21: Selected examples from the comparison of **Swadesh32** concepts of languages possibly influenced by western and Chinese languages. eng: English, deu: German, fra: French, jpn: Japanese, kor: Korean, zho: Mandarin Chinese.

concept	lang.	ngrams
'bird'	spa	$birds$, $bird$, $fowls$
	tgl	$birds$, $bird$, $fowl
	ceb	$birds$, $bird$, $fowl
	hil	$birds$, $bird$, $sparrows$
'ear'	spa	ear, $ears$, $heard$
	tgl	$ears$, ear, $listen$, $hearing$
	ceb	$ears$, ear, $hear$, $hearing$
	hil	ear, $ears$, $hear
'tongue'	spa	$tongue$, $tongues$, $language
	tgl	$tongue$, $tongues$, $language
	ceb	$tongue$, $tongues$, $language
	hil	$tongue$, $tongues$

Table 22: Selected examples from the comparison of **Swadesh32** concepts of several Philippine languages influenced by Spanish. spa: Spanish, tgl: Tagalog, ceb: Cebuano, hil: Hiligaynon.

concept	lang.	ngrams
'tree'	yor	$tree$, $trees$, wood, $stake, $frankincense$, $thornbush, $palm-tree$
	ibo	$tree$, $trees$, $pole, wood, $impal, $stake, $panel
	mcn	$tree$, $trees$, wood, $stake, $impale, $cedar, $timber
	twi	$tree$, $trees$, $wood, $panel$, $pole, $figs$, $timber
'hair'	yor	$hair$, $hairs$, $wool$
	ibo	$hair$, $hairs$, $wool, $shear, $beard
	mcn	$hair$, $hairs$, $wool$, $shave, $baldness$, $shear, goat
	twi	$hair$, $hairs$, $beard, $shave, $head$, $wool
'mouth'	yor	$mouth$, $mouths$, $entrance, $kiss, $palate$, $marvel, $suckling
	ibo	$mouth$, $mouths$, $gate, $entrance, $lip, curse, $precious$
	mcn	$mouth$, $mouths$, $lips$, fulfill, $denie, $disown, $entrance
	twi	$mouth$, $mouths$, $gat, $collect, $lip, $entrance, $registered$

Table 23: Selected examples from the comparison of **Swadesh32** concepts of four African languages. yor: Yoruba, ibo: Igbo, mcn: Masana, twi: Twi.

concept	lang.	ngrams
'bird'	cak	$tree$, $trees$, wood, $pole, $cedar, $figs$, $palm-tree$
	kjb	$tree$, $trees$, $cedar, $panel, $wood, $figs$, $pole
	tzj	$tree$, $trees$
'seed'	cak	$seed$, $seeds$, braham, $vine, $sow, fruit, $harvest
	kjb	$seed$, $seeds$, braham, $garden, fruits$, $vine, $harvest
	tzj	$seed$, $seeds$, $sow, $harvest$, fruit
'knee'	cak	$knees$, $bow, $worship, $trembl, fell$
	kjb	$knees$, $knee$, $obeisance$, $worship, $bow, $fell$
	tzj	$knees$, $worship, $obeisance$, $fell$

Table 24: Selected examples from the comparison of **Swadesh32** concepts of three Mayan languages. cak: Kaqchikel, kjb: Q'anjob'al, tzj: Tz'utujil.

concept	target lang.	trans. in target lang. (eng)
‘mouth’	jpn	口 (mouth, opening, entrance)
	kor	구(口) (entrance, gate, mouth)
	zho	口(mouth, gate, entrance), 嘴(mouth, lips)
	fra	bouche (mouth)
‘tongue’	spa	lengua (tongue, language)
	tgl	dilà (tongue, language), wikà (tongue, language)
	ceb	dila (tongue), pinulongan (tongue, language)
	hil	dilà (tongue), dilâ (tongue)
	msa	lidah (tongue), oojoo leeda (tongue)

Table 25: We use online dictionaries such as PanLex and Google Translate to look up two example concepts in the target languages and verify the associated meanings of their translations in English. We show English translations considering all used sources. For example, we obtain different translations for Tagalog “dilà” and “wikà” depending on the dictionary source and find that they can possibly mean both “tongue” and “language”. The target language translations are consistent with our findings: the three East Asian languages (jpn, kor, zho) share a common conceptualization of `mouth` as `entrance`, which is missing for French (fra); similar to Spanish (spa), some Philippine languages (tgl, ceb) conceptualize `tongue` as `language`, whereas another Austronesian language, Standard Malay (msa), does not.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Yes, in 'Limitations' section.

☑ A2. Did you discuss any potential risks of your work?
Yes, in 'Ethics Statement & Risks' section.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Yes, in Abstract and in Section 1.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Yes, in Section 3, Section 4, Section A in the appendix and Section C in the appendix.

☑ B1. Did you cite the creators of artifacts you used?
Yes, in Section 3, Section 4, Section A in the appendix and Section C in the appendix.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Yes, Section A. And as far as we know, there is no explicit license for PBC dataset.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Yes, in 'Ethics Statement & Risks' section.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Yes, in 'Ethics Statement & Risks' section.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Yes, in Section 3, Section 4, Section A in the appendix and Section C in the appendix.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Yes, in Section 3, Section 4, Section A in the appendix and Section C in the appendix.

C ☑ Did you run computational experiments?

Yes, in Section 3, Section 4.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No, we do not use neural networks so there are no "parameters" in our model. Nevertheless, we mention the runtime in 'Limitation' section and infrastructure in Section D in the appendix.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes, in Section 3, Section 4 and Section B in the appendix.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Yes, in Section 4 and Section C in the appendix

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Yes, in Section D.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Yes, our human annotator is one of the coauthors with a linguistic background and manually evaluates the results of our method and classifies each language into a category. Details are in Section 4 and Section C in the appendix.

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Yes, we introduce the criterion in Section 4 and Section C in the appendix.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. No, not relevant in our case.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. No, not relevant in our case.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. No, not relevant in our case.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. No, not relevant in our case.