Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2884–2892
Marseille, 11–16 May 2020

The Johns Hopkins University Bible Corpus: 1600+ Tongues for Typological Exploration

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and **David Yarowsky**
Center for Language and Speech Processing
Johns Hopkins University
(arya, rewicks, dlewis77, amueller, wswu, oadams, gnicola2, yarowsky)@jhu.edu,
post@cs.jhu.edu

Abstract

We present findings from the creation of a massively parallel corpus in over 1600 languages, the Johns Hopkins University Bible Corpus (JHUBC). The corpus consists of over 4000 unique translations of the Christian Bible and counting. Our data is derived from scraping several online resources and merging them with existing corpora, combining them under a common scheme that is verse-parallel across all translations. We detail our effort to scrape, clean, align, and utilize this ripe multilingual dataset. The corpus captures the great typological variety of the world's languages. We catalog this by showing highly similar proportions of representation of Ethnologue's typological features in our corpus. We also give an example application: projecting pronoun features like clusivity across alignments to richly annotate languages which do not mark the distinction.

Keywords: low-resource NLP, parallel corpus, typology, function words, Bible

1. Introduction

> I thank my God, I speak with tongues more than ye all.
>
> (1 Cor. 14:18; King James Version)

The high water mark for "low-resource" in NLP has risen in recent years. Guzmán et al. (2019) present "low-resource" translation datasets of half a million parallel sentences, and McCarthy et al. (2019) evaluate on the 600k-sentence WMT Romanian–English dataset as a low-resource benchmark. By contrast, this work is concerned with mere thousands of sentences within each language.

The Bible is a short but multi-parallel text organized in a consistent structure across languages. It has several dozen named constituent books, each divided into chapters. The chapters are collections of verses, which roughly map onto sentences. The book, chapter, and verse pinpoint the particular sequence of text. While certain books may not be considered 'canon' and translators may disagree on which verses to include, the structure is sufficiently parallel for several applications. Further, the Bible provides high coverage of core vocabulary.

Our corpus currently spans 1611 diverse written languages, with constituents of more than 90 language families. These languages display staggering breadth in typological features and their combinations. Common typological features (such as SVO, SOV, or VOS word order) and rarer features like inclusive/exclusive pronoun marking are present. Many of the languages in this corpus display features that are not present in major European languages or other large corpora, which makes this resource beneficial to further the research of machine translation and morphological analysis.

The collection, cleaning, and alignment of new languages and translations is an ongoing project. As new versions of the Bible are scraped and prepared, our corpus continues to grow.

Family	JHU	ETHN	JHU %	ETHN %
Niger–Congo	313	1542	19.43	20.63
Austronesian	277	1257	17.19	16.82
Trans-New Guinea	133	482	8.26	6.45
Sino-Tibetan	101	455	6.27	6.09
Indo-European	91	448	5.65	5.99
Otomanguean	83	178	5.15	2.38
Afro-Asiatic	67	377	4.16	5.04
Nilo-Saharan	52	206	3.23	2.76
Creole	27	93	1.68	1.24
Quechuan	27	44	1.68	0.59
Uto-Aztecan	26	61	1.61	0.82
Mayan	25	31	1.55	0.41
Maipurean	24	56	1.49	0.75
Turkic	20	41	1.24	0.55
Australian	19	381	1.18	5.10
Tucanoan	16	25	0.99	0.323
Tupian	15	76	0.93	1.02
Austro-Asiatic	14	167	0.87	2.23
Language isolate	14	88	0.87	1.18
Algic	12	42	0.74	0.56

Table 1: The top 20 largest language families in the JHUBC corpus. JHU percent denotes the percent of the languages in this corpus that are in each language family (normalized by 1611). Ethnologue percent denotes the percent of all Ethnologue languages that are a member of this family (normalized by 7474).

2. Related Work

As of January 2019, the entire Bible (approximately 40k verses) has been translated into 692 of the world's languages. 1,547 more languages have at least the New Testament (approximately 8k verses), and another 1,123 languages have at least a portion translated into them.[1] The religious scripture represents the most broadly translated literary work in human history (Mayer and Cysouw, 2014), and new translations continue to be produced.

Naturally, we are not the first authors to recognize the Bible as a ripe resource for language exploration and processing. Other parallel Bible corpora have been created, from Resnik et al. (1999)'s 13 parallel languages to Christodouloupoulos and Steedman (2015)'s 100. Although Mayer and Cysouw (2014) presented a corpus of 847 Bibles, the resource is no longer available. Asgari and Schütze (2017) suggest that the corpus grew to 1556 Bibles in 1169 languages before its disappearance. We provide 2.6 times as many translations, in 38% more languages.

Beyond this, multi-parallel corpora are few but popular (Koehn, 2005; Tiedemann, 2012; Duh, 2018; Qi et al., 2018, *inter alia*), including Agić and Vulić (2019), who present 100,000 sentences parallel across 300 languages. None have the linguistic breadth of the Bible.

3. Constructing a Corpus

Our use of the Christian Bible is motivated by several factors. First, it has been translated into more languages than any other text. Second, its division into cross-lingually consistent chapters and verses ascribes a natural parallel structure, necessary for learning effective machine translation. Third, the data is easily accessible online. Finally, the text covers a vast amount of languages' **core vocabulary**. Resnik et al. (1999) find high overlap between the English Bible and both the Longman Dictionary of Contemporary English (Summers and Gadsby, 1995) and the Brown Corpus (Francis and Kucera, 1964). Wu et al. (2020) similarly find that knowledge of how to express a cross-lingually consistent set of core concepts provides 68% token coverage in the English Bible (before accounting for inflectional variants).

The Bibles we release represent an aggregation and normalization of prior work (Mayer and Cysouw, 2014; Asgari and Schütze, 2017; Black, 2019) and independent web scraping.[2] Our corpus currently contains 4272 Bibles from 1611 languages, including 27 English translations.

3.1. Acquiring and Expanding the CMU Wilderness Corpus

The CMU Wilderness Corpus (Black, 2019) is distributed as a web scraping script, drawing content from the website `bible.is`. Where possible, we scraped directly from `bible.is`, which includes verse IDs in a structured format for all verses. Pleasantly, it also often includes the Old Testament, which is absent in all of the data of the CMU Wilderness Corpus. In some cases, the text was rendered unobtainable by changes to the website since Black (2019)'s initial scrape. In these cases, we backed off to text we scraped via the provided scripts before the website changes. These scrapes include verse numbers embedded in the plaintext rather than presented in a structured format. Since numbers which are *not* verse IDs can also be found in the text, the verse identification is ambiguous. We treat the verse alignment problem within a chapter as an instance of the longest common subsequence problem. Here, in a chapter with a pre-known number of verses m, the sequence of verse numbers $A = [1, 2, \dots, m]$ is matched to the sequence B of n numbers extracted from the chapter text. The **longest common subsequence** is going to give the best explanation of the numbers B seen in text, "explaining" them either as a verse ID or a number seen in text. It can be solved efficiently using dynamic programming in $O(m \times n)$ time (Wagner and Fischer, 1974).

3.2. Preparation

Verse alignment Our re-release of the Mayer and Cysouw (2014) Bibles and the web-scraped Bibles (Asgari and Schütze, 2017) is already verse-aligned. The CMU Wilderness Bibles (Black, 2019) are verse-aligned using the dynamic programming approach explained in §3.1.

Normalization We normalize all characters to their canonical (Unicode NFKC) form.[3] Cross-references, footnote markers, and explanatory parentheticals are stripped. We replace archaisms in the King James Version of the English Bible ('thou' forms; '-est' and '-eth' verb forms) with their modern equivalents.

Tokenization We preserve the tokenization of the Mayer and Cysouw (2014)–derived Bibles.[4] For all others, we apply the spaCy tokenizer.[5]

Deduplication We use a simple but effective heuristic to remove duplicate Bibles: we compute pairwise edit distances on a common sample of verses for all files, ignoring non-letter characters (to mitigate differences in tokenization). Then we eliminate Bibles above a threshold of similarity. This reduces 4998 acquired Bibles to our 4272.

Availability The corpus is provided as a collection of plain text files, one verse per line. Verse alignment is maintained by including blank lines for verses excluded in certain translations. A master list that maps line numbers to verse numbers is also provided, using the fixed-width book, chapter, and verse ID system of Mayer and Cysouw (2014).

Due to copyright strictures for many translations of Scripture, the dataset is available by contacting the authors.

4. Corpus Features and Statistics

86% of the Bibles contain only the New Testament (nearly 8000 verses); the remainder is the Old Testament, which contains over 31,000 verses.

Table 2 stratifies the type–token ratios (Ure, 1971) of our Bibles by language family. Figure 1 shows the presence of each book of the Bible, sorted by language.

[1] `http://web.archive.org/web/20200304154817/https://www.unitedbiblesocieties.org/key-facts-bible-access/`

[2] The Mayer and Cysouw (2014) corpus is no longer available.

[3] `https://unicode.org/reports/tr15/`

[4] The exception is Chinese; we provide a character-segmented replacement for the punctuation-separated original. Domain mismatch may explain poor segmentation performance in an early experiment with the Stanford Segmenter (Chang et al., 2008).

[5] `https://spacy.io/`

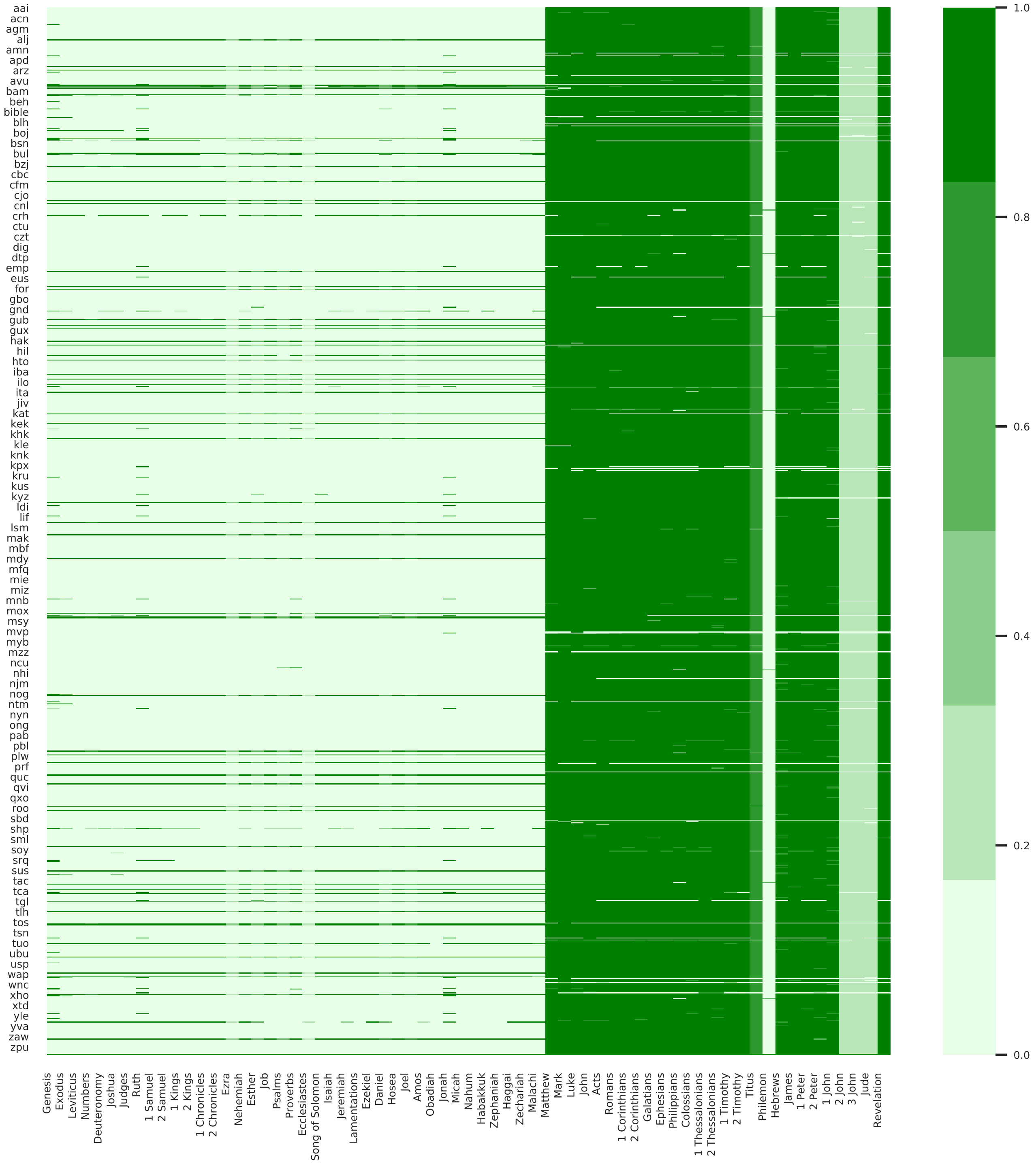


Figure 1: Heat map of the 66 Bible books' presence by language. (Twenty non-canon books which appear in only a handful of languages are omitted.) Nearly all languages have a complete New Testament, and several also have a complete Old Testament.

5. Bibles as a Low-Resource Asset

Exploiting the Bible, Agić et al. (2015) learn POS taggers for 100 languages and evaluate on 25 languages with test sets. Parallel Bibles also aid a variety of cross-lingual tasks, e.g., dependency parsing (Schlichtkrull and Søgaard, 2017), sentence embedding (Levy et al., 2017), verbal morphology induction (Yarowsky et al., 2001), and multilingual optical character recognition (Kanungo et al., 2005). None of these employ the fact that multiple interpretations can be used in tandem.

By contrast, Xia and Yarowsky (2017) leverage 27 English translations of the Bible. They use alignment and consensus to transfer dependency parses across languages. Nicolai and Yarowsky (2019) build on this, using projection with the same 27 English translations to develop morphosyntactic analyzers for low-resource languages. Nicolai et al. (2020)

Family	TTR
Tai-Kadai	0.621
Arai (Left May)	0.590
Khoe-Kwadi	0.535
South-Central Papuan	0.439
Kiowa-Tanoan	0.427
Eskimo-Aleut	0.426
Aymaran	0.370
Tungusic	0.350
Dravidian	0.263
Eyak-Athabaskan	0.261
Algic	0.261
Mixed language	0.260
Piawi	0.249
Iroquoian	0.244
Harákmbut	0.228
Pauwasi	0.214
Quechuan	0.208
Unclassified	0.207
Maipurean	0.206
East Geelvink Bay	0.202
Yuat	0.201
Senagi	0.198
Tequistlatecan	0.197
Pidgin	0.196
Turkic	0.192
Chibchan	0.191
Chipaya-Uru	0.184
Mapudungu	0.184
Tacanan	0.174
Mixe-Zoquean	0.167
Uralic	0.167
Constructed language	0.161
Witotoan	0.160
Muskogean	0.160
Panoan	0.155
Huavean	0.150
North Bougainville	0.150
Nakh-Daghestanian	0.149
Jivaroan	0.148
Indo-European	0.148
Arauan	0.144
Guaykuruan	0.144
Austro-Asiatic	0.142
Misumalpan	0.141
Karajá	0.140
Tarascan	0.140
South Bougainville	0.138
Matacoan	0.137
Kartvelian	0.137
Guajiboan	0.136
Sino-Tibetan	0.135
Totonacan	0.135
Australian	0.135
Border	0.134
Uto-Aztecan	0.132
Afro-Asiatic	0.131
Cahuapanan	0.126
Japonic	0.125
Yele-West New Britain	0.124
Eastern Trans-Fly	0.123
Paezan	0.122
Mongolic	0.121
Tucanoan	0.112
Zaparoan	0.111
Niger-Congo	0.110
Cariban	0.109
Language isolate	0.105
Barbacoan	0.097
Jicaquean	0.097
Ramu-Lower Sepik	0.096
Trans-New Guinea	0.096
East Bird's Head-Sentani	0.092
West Papuan	0.089
Nilo-Saharan	0.088
Austronesian	0.088
Torricelli	0.085
Puinavean	0.084
Chocoan	0.079
Koreanic	0.079
Tupian	0.078
Tor-Kwerba	0.075
Chukotko-Kamchatkan	0.075
Jean	0.073
Zamucoan	0.072
Mascoyan	0.071
Nambikwara	0.071
Sepik	0.071
Mayan	0.070
Otomanguean	0.070
East New Britain	0.066
Yaguan	0.064
Yanomaman	0.061
Hmong-Mien	0.061
Creole	0.044
Maxakalian	0.017

Table 2: Low type–token ratios are an indicator of a language's morphological richness. We report type–token ratios, averaged over each language family, sorted by this ratio. When there are several translations within a single language, we average these first. (Tai-Kadai would be expected to have a low type–token ratio. It does not because we did not segment the text in languages that do not indicate word boundaries with spaces.))

take this a step further, releasing fine-grained morphosyntactic analysis and generation tools for more than 1,000 languages.

Additional practical resources can be derived from the Bible: translation matrices of named entities (Wu et al., 2018), low-resource transliteration tools (Wu and Yarowsky, 2018), and bitext for multilingual translation (Mueller et al., 2020).

6. Typological Analysis

Beyond overt NLP applications, the large Bible corpus allows us to investigate the typology of the world's languages. We find that the breadth and diversity of our Bible corpus aligns well with the typological diversity of the world's languages.[6]

6.1. Feature Compilation

In order to compile a typological description of this corpus, we utilize the URIEL typological database (Littell et al., 2017) and supplement this with a surface-level parser of Ethnologue (Eberhard et al., 2019) typology descriptions in order to leverage the most recent entries. This parser functions similarly to the original parser described by Littell et al. (2017)—using simple Boolean logic on contents of Ethnologue descriptions. As most Ethnologue entries have similar phrasing or terminology, a language can be classified with a particular typological feature if the description contains one of a handful of phrases used on Ethnologue to describe that feature, but does not contain *any* of the handful of phrases used to describe the *absence* of that feature. These feature–phrase mappings were handcrafted using a sorted list of the most common Ethnologue typology entries.[7]

While the set of typological features for each language may be incomplete—e.g., a typological description of English mentions ordering of words in 5 of its 14 typological entries but does not mention affixes, plurals, nominative/accusative, or clitics—the parser is accurate in its ability to classify 10 of these 14 with the exception of the 4 entries describing phonological features.

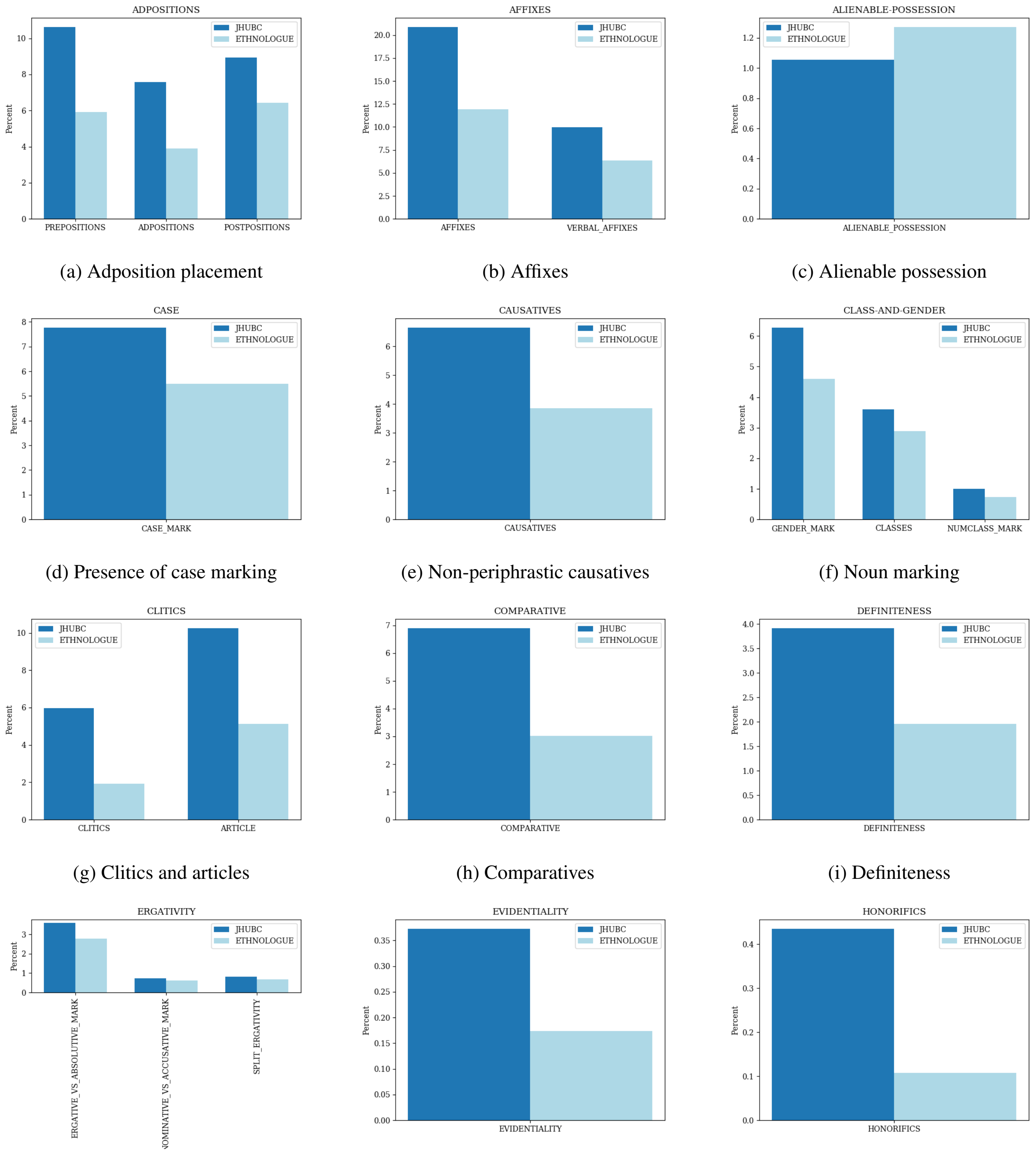
Additionally, some languages have *no* typological entry on Ethnologue and are unsupported by URIEL. To reconstruct a feature set for these languages, we assume that a language family possesses a common set of features. We reconstruct typological features of a language family by taking the intersection of all daughter languages with a typological description. This common set is then applied to all daughter languages *without* a typological description.

Accounting for potential defects of these databases, we present each feature as the percent of JHUBC languages containing this feature in parallel to all Ethnologue languages containing this feature. While some percentages are lower than we might anticipate—approximately 4.5 percent of all Ethnologue languages mark gender—this shows evidence of the previously mentioned incompleteness of the databases. Despite this, we see high agreement between the representations in Ethnologue and those in our corpus, supporting its value as a tool for typologically informed exploration. Twenty-four major features are illustrated in Figure 2 and Figure 3.

As Ethnologue is a human-generated database, it has artifacts of human-biased labels of these typological features—specifically the fact that labels are not exhaustive. Ethno-

[6] While the JHUBC only contains languages with a writing system, Ethnologue also contains languages that exist only orally or in sign. We use all Ethnologue languages when comparing them to ours.

[7] Ethnologue conveniently delimits its entries using ";". A description can be split on ";", which results in a list of typological entries for each language.



(a) Adposition placement

(b) Affixes

(c) Alienable possession

(d) Presence of case marking

(e) Non-periphrastic causatives

(f) Noun marking

(g) Clitics and articles

(h) Comparatives

(i) Definiteness

(j) Ergativity

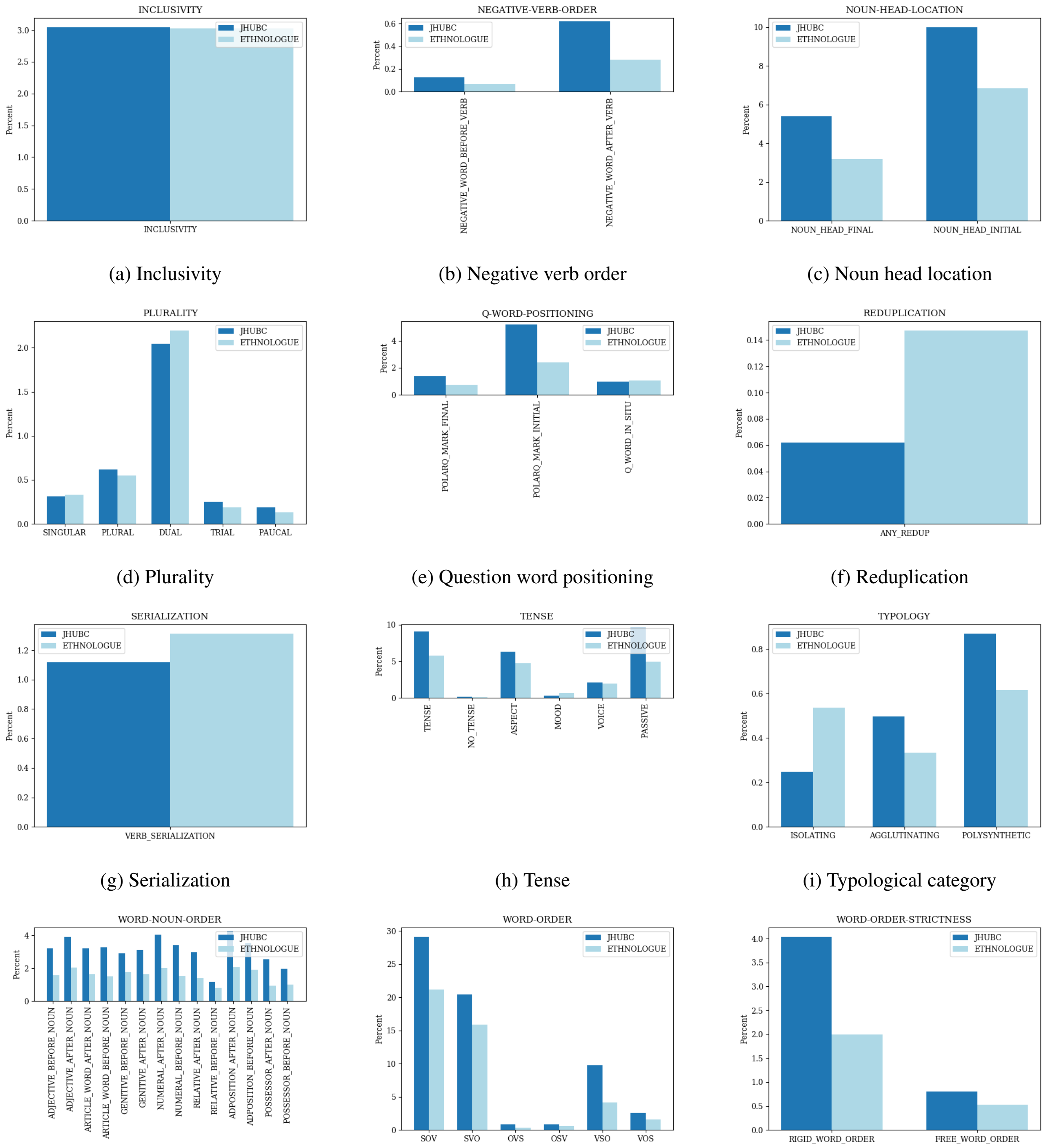
(k) Evidentiality

(l) Honorifics

Figure 2: Typological features of our languages, compared against Ethnologue.

logue only lists features when they are remarkable in the particular language. For instance, while the English language contains both indefinite and definite articles, Ethnologue does not mention this fact, favoring more unique features–such as word order or free stress.

Thus, we acknowledge that these percentages do not match the ground truth percentages of the world's languages. For this reason, we compare the Ethnologue languages percentages to this corpus. This means that while Ethnologue, according to our parser, claims only 4.5% of the world's languages have a gender mark, we claim that our corpus is representative as our corpus is composed of languages—of which roughly 6.5% mark for gender (according to the Ethnologue parser). For the same reason, we do not have labels for all languages given a feature. In the case of a feature such as tense—which as a "no tense" label for some languages—we would consider the unlabeled languages as "unknown tense classification" which are omitted from the figures for clarity and brevity.



(a) Inclusivity

(b) Negative verb order

(c) Noun head location

(d) Plurality

(e) Question word positioning

(f) Reduplication

(g) Serialization

(h) Tense

(i) Typological category

(j) Word order regarding nouns

(k) Word order regarding verbs

(l) Word order strictness

Figure 3: Additional typological features of our languages, compared against Ethnologue.

6.2. Language Families

Among the 1611 languages, 95 of 155 language families are represented.[8] 35 language families have at least five members with a Bible in the corpus; 32 families have only one. In Table 3 we present these language families' representation in our corpus.

Additionally, these languages represent a large number and wide range of speakers. Most of these languages (1339) are spoken by one million or fewer speakers with 84 of these languages being spoken by fewer than one thousand speakers. Many of the languages are local vernacular languages (Graddol, 1997). We estimate that these languages are an L1 (first language) of at least 76% of the world population. Notable exceptions include the world's sign languages.

7. Case Study: Projection of Morphosyntactic Information onto English Pronouns

Pronouns form a closed class of words, but in many languages, they inflect for features such as person, gender, case,

[8] The 155 language families are defined by the 155 largest non-overlapping language families described by Ethnologue.

Family	Count	%
Niger–Congo	782	15.77
Austronesian	744	15.00
Indo-European	706	14.23
Trans-New Guinea	380	7.66
Otomanguean	251	5.06
Mayan	197	3.97
Sino-Tibetan	196	3.95
Afro-Asiatic	190	3.83
Nilo-Saharan	122	2.46
Quechuan	100	2.02
Uto-Aztecan	86	1.73
Creole	77	1.55
Maipurean	71	1.43
Turkic	69	1.39
Tucanoan	58	1.17
Tupian	53	1.07
Unclassified	53	1.07
Language isolate	48	0.97
Austro-Asiatic	46	0.93
Sepik	37	0.75
Uralic	37	0.75
Chibchan	34	0.68
Torricelli	34	0.69
Mixe-Zoquean	32	0.65
Totonacan	29	0.58
Australian	28	0.56
Dravidian	28	0.56
Panoan	28	0.56
Cariban	26	0.52
Algic	22	0.44
Tai-Kadai	19	0.33
Jivaroan	18	0.33
Witotoan	15	0.32
Chocoan	14	0.28
Hmong-Mien	14	0.28
Jean	13	0.26
Tacanan	13	0.26
West Papuan	13	0.26
Eskimo-Aleut	12	0.24
Eyak-Athabaskan	12	0.24
Guajiboan	12	0.24
Nakh-Daghestanian	11	0.22
Guaykuruan	10	0.20
Matacoan	10	0.20
Barbacoan	9	0.18
Mongolic	9	0.18
Puinavean	9	0.18
Arauan	8	0.16
Koreanic	8	0.16
Paezan	8	0.16
Ramu-Lower Sepik	8	0.16
Aymaran	7	0.14
Border	7	0.14
Yele-West New Britain	7	0.14
Chipaya-Uru	6	0.12
East Bird's Head-Sentani	6	0.12
Eastern Trans-Fly	6	0.12
Misumalpan	6	0.12
Mixed language	6	0.12
Japonic	5	0.10
South-Central Papuan	5	0.10
Harákmbut	4	0.08
Huavean	4	0.08
Iroquoian	4	0.08
Jicaquean	4	0.08
Mapudungu	4	0.08
South Bougainville	4	0.08
Zaparoan	4	0.08
Arai (Left May)	3	0.06
Cahuapanan	3	0.06
Karajá	3	0.06
Khoe-Kwadi	3	0.06
Mascoyan	3	0.06
Maxakalian	3	0.06
Nambikwara	3	0.06
North Bougainville	3	0.06
Pauwasi	3	0.06
Senagi	3	0.06
Tarascan	3	0.06
Tequistlatecan	3	0.06
Tor-Kwerba	3	0.06
Yaguan	3	0.06
Yanomaman	3	0.06
Constructed language	2	0.04
East Geelvink Bay	2	0.04
East New Britain	2	0.04
Pidgin	2	0.04
Zamucoan	2	0.04
Chukotko-Kamchatkan	1	0.02
Kartvelian	1	0.02
Kiowa-Tanoan	1	0.02
Muskogean	1	0.02
Piawi	1	0.02
Tungusic	1	0.02
Yuat	1	0.02

Table 3: The count of individual Bibles in a language that is a member of each family. The percent denotes the percent of the JHUBC Bibles that come from that language family.

and clusivity. We leverage the parallelism of the Bible to annotate pronouns in the English Bibles for these features which are otherwise unmarked in English.

We first collect a pronoun list for a small set of languages that mark pronouns for a number of different phenomena such as number, gender, plurality, and clusivity, and annotated these lists with UniMorph-style inflectional information (Sylak-Glassman et al., 2015; Kirov et al., 2018; McCarthy et al., 2020). Next, we word-align the English Bibles with the Bibles in those languages. Projecting from the source onto English, we obtain source–English pronoun hypotheses for each English pronoun. For each feature, we vote among the languages to arrive at a final annotation for English. To mitigate the ubiquity of certain feature values over others (i.e., the nominative case is much more prevalent in our languages than the essive case), we normalize each feature by the number of languages in which it is present. We can then use these annotations to identify pronouns in other languages via alignment.

Clusivity Unlike features such as case and number, clusivity is often a subjective decision made by the translator—the information is not present in the original. Across five languages that mark clusivity (Cebuano, Ilocano, Tagalog, Samoan, and Maori), we obtain a Fleiss's κ of 0.3066, illustrating the subjectivity of the task.

8. Conclusion

We present findings from the creation of a massively parallel corpus in over 1600 languages, the Johns Hopkins University Bible Corpus (JHUBC). The corpus consists of over 4000 unique translations of the Christian Bible and counting. Our data is derived from scraping several online resources and merging them with existing corpora, combining them under a common scheme that is verse-parallel across all translations. We detail our effort to scrape, clean, align, and utilize this ripe multilingual dataset. The corpus captures the great typological variety of the world's languages. We catalog this by showing highly similar proportions of representation of Ethnologue's typological features in our corpus. We also give an example application: projecting pronoun features like clusivity across alignments to richly annotate languages which do not mark the distinction. Copyright restrictions limit our ability to publicly disseminate the data. The dataset is available by contacting the authors.

9. Bibliographical References

Agić, Ž. and Vulić, I. (2019). JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy, July. Association for Computational Linguistics.

Agić, Ž., Hovy, D., and Søgaard, A. (2015). If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 268–272, Beijing, China, July. Association for Computational Linguistics.

Asgari, E. and Schütze, H. (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, September. Association for Computational Linguistics.

Black, A. W. (2019). CMU wilderness multilingual speech dataset. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5971–5975, May.

Chang, P.-C., Galley, M., and Manning, C. D. (2008). Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 224–232, Columbus, Ohio, June. Association for Computational Linguistics.

Christodouloupoulos, C. and Steedman, M. (2015). A massively parallel corpus: the Bible in 100 languages. *Language resources and evaluation*, 49(2):375–395.

Duh, K. (2018). The multitarget TED Talks task. `http://www.cs.jhu.edu/~kevinduh/a/multitarget-tedtalks/`.

Eberhard, D. M., Simons, G. F., and Fennig, C. D. (2019). *Ethnologue: Languages of the World. Twenty-second edition.* Dallas, Texas: SIL International.

Francis, W. N. and Kucera, H. (1964). Brown corpus. *Department of Linguistics, Brown University, Providence, Rhode Island*, 1.

Graddol, D. (1997). *The future of English?: A guide to forecasting the popularity of the English language in the 21st century*. British Council.

Guzmán, F., Chen, P.-J., Ott, M., Pino, J., Lample, G., Koehn, P., Chaudhary, V., and Ranzato, M. (2019). The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6100–6113, Hong Kong, China, November. Association for Computational Linguistics.

Kanungo, T., Resnik, P., Mao, S., Kim, D.-W., and Zheng, Q. (2005). The Bible and multilingual optical character recognition. *Commun. ACM*, 48(6):124–130, June.

Kirov, C., Cotterell, R., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., Faruqui, M., Mielke, S. J., McCarthy, A., Kübler, S., Yarowsky, D., Eisner, J., and Hulden, M. (2018). UniMorph 2.0: Universal morphology. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan, May. European Language Resource Association.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Levy, O., Søgaard, A., and Goldberg, Y. (2017). A strong baseline for learning cross-lingual word embeddings from sentence alignments. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 765–774, Valencia, Spain, April. Association for Computational Linguistics.

Littell, P., Mortensen, D. R., Lin, K., Kairis, K., Turner, C., and Levin, L. (2017). Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14. Association for Computational Linguistics.

Mayer, T. and Cysouw, M. (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, May. European Language Resources Association (ELRA).

McCarthy, A. D., Li, X., Gu, J., and Dong, N. (2019). Improved Variational Neural Machine Translation by Promoting Mutual Information. *arXiv e-prints*, page arXiv:1909.09237v1, Sep.

McCarthy, A. D., Kirov, C., Grella, M., Nidhi, A., Xia, P., Gorman, K., Vylomova, E., Mielke, S. J., Nicolai, G., Silfverberg, M., Arkhangelskij, T., Krizhanovsky, N., Krizhanovsky, A., Kylachko, E., Sorokin, A., Mansfield, J., Ernstreits, V., Pinter, Y., Jacobs, C., Cotterell, R., Hulden, M., and Yarowsky, D. (2020). UniMorph 3.0: Universal morphology. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, Marseilles, France, May. European Language Resources Association (ELRA).

Mueller, A., Nicolai, G., McCarthy, A. D., Lewis, D., Wu, W., and Yarowsky, D. (2020). An analysis of massively multilingual neural machine translation for low-resource languages. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, Marseilles, France, May. European Language Resources Association (ELRA).

Nicolai, G. and Yarowsky, D. (2019). Learning morphosyntactic analyzers from the Bible via iterative annotation projection across 26 languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1765–1774, Florence, Italy, July. Association for Computational Linguistics.

Nicolai, G., Lewis, D., McCarthy, A. D., Mueller, A., Wu, W., and Yarowsky, D. (2020). Fine-grained morphosyntactic analysis and generation tools for more than one thousand languages. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, Marseilles, France, May. European Language Resources Association (ELRA).

Qi, Y., Sachan, D., Felix, M., Padmanabhan, S., and Neubig, G. (2018). When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana, June. Association for Computational Linguistics.

Resnik, P., Olsen, M. B., and Diab, M. (1999). The Bible as a parallel corpus: Annotating the 'book of 2000 tongues'. *Computers and the Humanities*, 33(1):129–153, Apr.

Schlichtkrull, M. and Søgaard, A. (2017). Cross-lingual dependency parsing with late decoding for truly low-resource languages. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 220–229, Valencia, Spain, April. Association for Computational Linguistics.

Summers, D. and Gadsby, A. (1995). *Longman dictionary of contemporary English*. Longman Harlow.

Sylak-Glassman, J., Kirov, C., Yarowsky, D., and Que, R. (2015). A language-independent feature schema for inflectional morphology. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 674–680, Beijing, China, July. Association for Computational Linguistics.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-*

2012), pages 2214–2218, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Ure, J. (1971). Lexical density and register differentiation. *Applications of linguistics*, pages 443–452.

Wagner, R. A. and Fischer, M. J. (1974). The string-to-string correction problem. *J. ACM*, 21(1):168–173, January.

Wu, W. and Yarowsky, D. (2018). A comparative study of extremely low-resource transliteration of the world's languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Wu, W., Vyas, N., and Yarowsky, D. (2018). Creating a translation matrix of the Bible's names across 591 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Wu, W., Nicolai, G., and Yarowsky, D. (2020). Multilingual dictionary based construction of core vocabulary. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, Marseilles, France, May. European Language Resources Association (ELRA).

Xia, P. and Yarowsky, D. (2017). Deriving consensus for multi-parallel corpora: an English Bible study. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 448–453, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

Yarowsky, D., Ngai, G., and Wicentowski, R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*.