ParCourE: A Parallel Corpus Explorer for a Massively Multilingual Corpus

Ayyoob Imani[1], **Masoud Jalili Sabet**[1], **Philipp Dufter**[1],
Michael Cysouw[2], **Hinrich Schütze**[1]

[1]Center for Information and Language Processing (CIS), LMU Munich, Germany
[2]Research Center Deutscher Sprachatlas, Philipps University Marburg, Germany.
{ayyoob, masoud, philipp}@cis.lmu.de

Abstract

With more than 7000 languages world-wide, multilingual natural language processing (NLP) is essential both from an academic and commercial perspective. Researching typological properties of languages is fundamental for progress in multilingual NLP. Examples include assessing language similarity for effective transfer learning, injecting inductive biases into machine learning models or creating resources such as dictionaries and inflection tables. We provide ParCourE, an online tool that allows to browse a word-aligned parallel corpus, covering 1334 languages. We give evidence that this is useful for typological research. ParCourE can be set up for any parallel corpus and can thus be used for typological research on other corpora as well as for exploring their quality and properties.

1 Introduction

While ≈7000 languages are spoken (Eberhard et al., 2020), the bulk of NLP research addresses English only. However, multilinguality is an essential element of NLP. It not only supports exploiting common structures across languages and eases maintenance for globally operating companies, but also helps save languages from digital extinction and fosters more diversity in NLP techniques.

There are extensive resources that can be used for massively multilingual typological research, such as WALS (Dryer and Haspelmath, 2013), Glottolog (Hammarstrm et al., 2020), BabelNet (Navigli and Ponzetto, 2012) or http://panlex.org. Many of them are manually created or crowdsourced, which guarantees high quality, but limits coverage, both in terms of content and languages.

We work on the Parallel Bible Corpus (PBC) (Mayer and Cysouw, 2014), covering 1334 languages. More specifically, we provide a word-aligned version of PBC, created using state-of-the-art word alignment tools. As word alignments
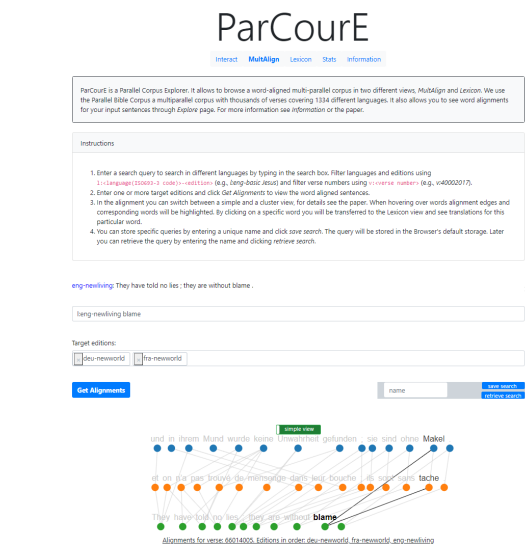

Figure 1: Screenshot of the ParCourE interface. It provides a word-aligned version of the Parallel Bible Corpus (PBC) spanning 1334 languages. Users can search for sentences in any language and see their alignments in other languages from MULTALIGN page. Alternatively they can feed their parallel sentences to INTERACTIVE view and see their word level alignments. They can look up translations of words in other languages, automatically induced from word alignments, from the LEXICON view (This page is interconnected with MULTALIGN). Statistics of the corpus is calculated and shown in the Stats view.

themselves are only of limited use, we provide an interactive online tool[1] that allows effective browsing of the alignments.

The main contributions of this work are: **i)** We provide a word-aligned version of the Parallel Bible Corpus (PBC) spanning 1334 languages and a total of 20M sentences ('verses'). For the alignment we use the state-of-the-art alignment methods SimAlign (Jalili Sabet et al., 2020) and Eflomal (Östling and Tiedemann, 2016a). **ii)** We release ParCourE,

---

[1]http://parcoure.cis.lmu.de/

a user interface for browsing word alignments, see the MULTALIGN view in Figure 1. We demonstrate the usefulness of ParCourE for typological research by presenting use cases in §6. **iii)** In addition to browsing word alignments, we provide an aggregated version in a LEXICON view and compute statistics that support assessing the quality of the word alignments. The two views (MULTALIGN and LEXICON views) are interlinked, resulting in a richer user experience. **iv)** ParCourE has a generic design and can be set up for any parallel corpus. This is useful for analyzing and managing parallel corpora; e.g., errors in an automatically mined parallel corpus can be inspected and flagged for correction.

2 Related Work

Word Alignment is an important tool for typological analysis (Lewis and Xia, 2008) and annotation projection (Yarowsky et al., 2001; Östling, 2015; Asgari and Schütze, 2017). Statistical models such as IBM models (Brown et al., 1993), Giza++ (Och and Ney, 2003), fast-align (Dyer et al., 2013) and Eflomal (Östling and Tiedemann, 2016b) are widely used. Recently, neural models were proposed, such as SimAlign (Jalili Sabet et al., 2020), Awesome-align (Dou and Neubig, 2021), and methods that are based on neural machine translation (Garg et al., 2019; Zenkel et al., 2020). We use Eflomal and SimAlign for generating alignments.

Resources. There are many online resources that enable typological research. WALS (Dryer and Haspelmath, 2013) provides manually created features for more than 2000 languages. We prepare a multiparallel corpus for investigating these features on real data. http://panlex.org is an online dictionary project with 2500 dictionaries covering 5700 languages and BabelNet (Navigli and Ponzetto, 2012) is a large semantic network covering 500 languages, but their information is generally on the type level, without access to example contexts. In contrast, ParCourE supports the exploration of word translations across 1334 languages in context.

Another line of work uses the **Parallel Bible Corpus (PBC)** for analysis. Asgari and Schütze (2017) investigate tense typology across PBC languages. Xia and Yarowsky (2017) created a multiway alignment based on fast-align (Dyer et al., 2013) and extracted resources such as paraphrases for 27 Bible editions. Wu et al. (2018) used align-

ments to extract names from the PBC.

One of the first attempts to index the Bible and align words in multiple languages were Strong's numbers (Strong, 2009[1890]); they tag words with similar meanings with the same ID. Mayer and Cysouw (2014) created an inverted index of word forms. Östling (2014) align massively parallel corpora simultaneously. We use the Eflomal word aligner by the same authorsostling2016efficient.

Finally, we review work on **Word Alignment Browsers.** Gilmanov et al. (2014)'s tool supports visualization and editing of word alignments. Akbik and Vollgraf (2017) use co-occurrence weights for word alignment and provide a tool for the inspection of annotation projection. Aulamo et al. (2020)'s filtering tool increases the quality of (mined) parallel corpora. Graën et al. (2017) rely on linguistic preprocessing, target corpus and word alignment exploration, do not show the graph of alignment edges and do not provide a dictionary view. While there is commonality with this prior work, ParCourE is distinguished by both its functionality and its motivating use cases: an important use case for us are typological searches; linguistic preprocessing is not available for many PBC languages; ParCourE can be used as an interactive explorer (but is not a fully-automated pipeline for a specific use case); our goal is not annotation; we use state-of-the-art word alignment methods. However, much of the complementary functionality in prior work would be useful additions to ParCourE. Another source of useful additional functionality would be work on embedding learning (Dufter et al., 2018; Kurfal and Östling, 2018) and machine translation (Tiedemann, 2018; Santy et al., 2019; Mueller et al., 2020) for PBC.

3 Features

ParCourE's user facing functionality can be divided into three main parts: MULTALIGN and LEXICON views and interconnections between the two.

3.1 Multiparallel Alignment Browser: MULTALIGN

ParCourE allows the user to search through the parallel corpus and check word alignments in a multiparallel corpus. An overview of MULTALIGN is shown in Figure 2.

In the **search field** (a(1)), the user can enter a text query and select (a(2)) multiple sentences for alignment. For narrowing the search scope, the
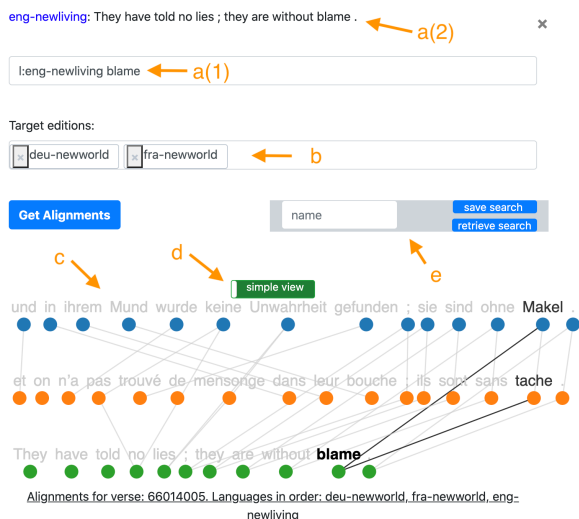
Figure 2: An overview of the MULTALIGN view. a) Search field for selecting sentences [a(1)] and the list of selected sentences [a(2)]. Any language can be used for the source sentence – in this case, it is English. b) Search bar for selecting the target languages. c) The alignment graph for the selected sentences in the source and the target languages. d) Switch button for simple view / cluster view. e) Save and retrieve search results

language and edition of the text segment can be specified in the beginning, e.g., by typing *l:eng-newworld2013*. Similarly, *v:40002017* specifies a verse ID.

PBC has 1334, so showing alignments for all translations of a sentence is difficult. We provide a drop-down (b) to select a subset of target languages for display.

For each sentence, a graph of alignment edges between selected languages is shown (c). By hovering over a word, the alignments of that word will be highlighted. Above each alignment graph, there is a button to switch between **Simple view** and **Cluster view** (d). In the simple view, when hovering over a word, only the alignment edges connected to that word are highlighted; in the cluster view, all words in a cluster (neighbors of neighbors) that are aligned together will be highlighted. We do not actually run any clustering algorithm on the alignment graph. Instead we simply highlight words that are up to two hops away from the hovered word. This helps spot a group of words across languages that have the same meaning.

Creating queries for typology research can take time. Thus, MULTALIGN allows the user to **save and retrieve** (e) queries.
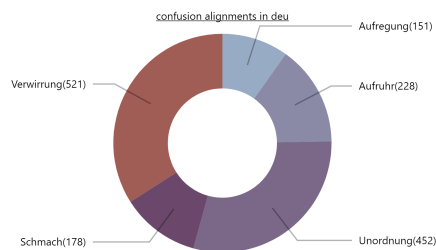

Figure 3: LEXICON view example: for the English word "confusion", there are five frequent translations in German. "Unordnung" literally means "disorder" and "Verwirrung" means "bewilderment".

3.2 Lexicon View: LEXICON

The MULTALIGN view allows the user to focus on word alignments on the sentence level and study the typological structure of languages in context. The LEXICON view focuses on word translations. The user can specify a source language by selecting the language code. This is to distinguish words with the same spelling in different languages. The user can search for one or multiple word(s) and specify target language(s). A pie chart for each target language depicting translations of the word is generated. Figure 3 shows German translations of "confusion" and the number of alignment edges for each. Word alignments are not perfect, so pie charts may also contain errors.

3.3 Interconnections

Both MULTALIGN and LEXICON views provide important features to the user for exploring the parallel corpus. For many use cases (cf. §6), the user may need to go back and forth between the views. For example, if she notices an error in the word alignment, she may want to check the LEXICON statistics to see if one of the typical translations of an incorrectly aligned word occurs in the sentence.

Thus, the two views are interconnected. In the MULTALIGN view, the user will be transferred to the LEXICON statistics of a word by clicking on it. This will open the LEXICON view, showing the search results for the selected word. Conversely, if the user clicks on one of the target translations in the LEXICON view, the MULTALIGN view will show sentences where this correspondence is part of the word alignment between source and target translation.

# editions	1758	# verses	20,470,892
# languages	1334	# verses / # editions	11,520
		# tokens / # verses	28.6

Table 1: PBC corpus statistics

3.4 Alignment Generation View: INTERACTIVE

The views mentioned so far provide the ability to search over the indexed corpus. This is useful when the main corpus of interest is fixed and the user has generated its alignments.

The INTERACTIVE view allows the user to study the alignments between arbitrary input sentences that are not necessarily in the corpus. Since the input sentences are not part of a corpus, INTERACTIVE uses SimAlign to generate alignments for all possible pairs of sentences. Similar to MULTALIGN, the INTERACTIVE view shows the alignment between the input sentences.

4 Experimental Setup

Corpus. We set up ParCourE on the PBC corpus provided by Mayer and Cysouw (2014). The version we use consists of 1758 editions (i.e., translations) of the Bible in 1334 languages (distinct ISO 639-3 codes). Table 1 shows corpus statistics. We use the PBC tokenization, which contains errors for a few languages (e.g., Thai). We extract word alignments for all possible language pairs. Since not all Bible verses are available in all languages, for each language pair we only consider mutually available verses.

PBC aligns Bible editions on the verse level by using verse-IDs that indicate book, chapter and verse (see below). Although one verse may contain multiple sentences, we do not split verses into individual sentences and consider each verse as one sentence.

Retrieval. Elasticsearch[2] is a fast and scalable open source search engine that provides distributed fulltext search. The setup is straightforward using an easy-to-use JSON web interface. We use it as the back-end for ParCourE's search requirement. We find that a single instance is capable of handling the whole PBC corpus efficiently, so we do not need a distributed setup. For bigger corpora, a distributed setup may be required. We created two types of inverted indices for our data: an edge-ngram in-

dex to support search-as-you-type capability and a standard index for normal queries.

Alignment Generation. SimAlign (Jalili Sabet et al., 2020) is a recent word alignment method that uses representations from pretrained language models to align sentences. It has achieved better results than statistical word aligners. For the languages that multilingual BERT (Devlin et al., 2019) supports, we use SimAlign to generate word alignments. For the remaining languages, we use Eflomal (Östling and Tiedemann, 2016a), an efficient word aligner using a Bayesian model with Markov Chain Monte Carlo (MCMC) inference. The alignments generated by SimAlign are symmetric. We use atools[3] and the grow-diag-final-and heuristic to symmetrize Eflomal alignments.

Lexicon Induction. We exploit the generated word alignments to induce lexicons for all 889,111 language pairs. To this end, we consider aligned words as translations of each other. For a given word from the source language, we count the number of times a word from the target language is aligned with it. The higher the number of alignments between two words, the higher the probability that the two have the same meaning. We filter out translations with frequency less than 5%.

5 Backend Design

An overview of our architecture can be found in Figure 4. The code is available online.[4]

Parallel Data Format. We use the PBC corpus format (Mayer and Cysouw, 2014): each verse has a unique ID across languages / editions, the *verse-ID*. The verse-ID is an 8-digit number, consisting of two digits for the book (e.g., 41 for the Gospel of Mark), three digits for the Chapter, and two digits for the verse itself. There are separate files for each edition. In each edition file, a line consists of the ID and the verse, separated by a tab.

Indexing. We identify a PBC verse using the following format: {verse-ID}@{language-code}-{edition-name}. We use this identifier to save and retrieve sentences with Elasticsearch. In addition, we store all metadata identifiers within Elasticsearch. Thus, we can search for a sentence by keyword, sentence number (= verse-ID), language code, or edition name.

ParCourE also supports the Corpus Alignment

---

[2]https://www.elastic.co/

[3]https://github.com/clab/fast_align
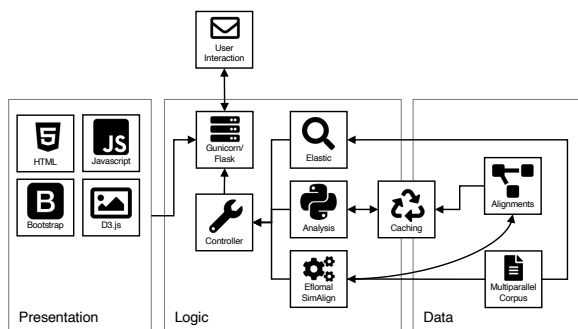[4]https://github.com/cisnlp/parcoure

Figure 4: Overview of the system architecture. We use a standard front-end stack with d3.js for visualization. The backend is written in Python, which we use for computing alignments and performing analyses such as lexicon induction. We use Elasticsearch for search. The input is a multiparallel corpus for which all alignments are precomputed. For speeding up the system we use smart caching algorithms for our analyses. Icons taken without changes from `https://fontawesome.com/license`.

Encoding (CES)[5] format. One can download parallel corpora in CES format and use our tools to adapt them to ParCourE's input format.

Alignment Computation. Since Eflomal's performance depends on the amount of data it uses for training, we concatenate all editions to create a bigger training corpus for languages that have more than one edition. If language l_1 has two, and language l_2 three different editions, then the final training corpus for this language pair will contain six aligned edition pairs.

System Architecture. ParCourE is built on top of modern open source technologies, see Figure 4. The back-end uses the Flask web framework,[6] Gunicorn web server,[7] and Elasticsearch.[8] The front-end utilizes the Bootstrap CSS framework,[9] and the d3 visualization library.[10] Since all these tools are free and open-source, there is no restriction on setting up and releasing a new ParCourE instance. To extract word alignments, one can use any tool, such as Eflomal, fast_align or SimAlign.

Performance Improvements. For good runtime performance, we precompute the word alignments. Regarding LEXICON, given a query word and a target language, ParCourE first looks for a precomputed lexicon file; if it does not exist, Par-

---

CourE obtains the translations for the query word online. To accelerate the translation process, ParCourE employs Python's multiprocessing library. The number of CPU cores is decided online based on the number of editions available for source and target languages.

For a corpus with 1334 languages, we will end up with 890,445 alignment files and the same number of lexicon files. We cache alignment / lexicon files to speed up access. We use the Last Recently Used (LRU) cache replacement algorithm.

6 ParCourE Use Cases

Languages differ in how they encode meanings/functions. There are various aspects that make such differences an interesting problem when dealing with a dataset that has good coverage of the entire variation of the world's languages. (i) Many such differences between languages are not widely acknowledged in linguistic theory, so to document the extent of variation becomes a discovery of sorts. For example, the fact that interrogative words might distinguish between singular and plural (Figure 6) turns out to be a typologically salient differentiation (Mayer and Cysouw, 2012). (ii) The variation of linguistic marking is even stronger in the domain of grammatical function, like the differentiation between the interrogative and relative pronoun in Figure 6. (iii) In lexical semantics, ParCourE supports the investigation of how languages carve up the meaning space differently (cf. Figure 5), especially when it comes to the ≈1000 low-resource languages covered in PBC. Massively parallel texts are an ideal resource to investigate such variation (Haspelmath, 2003).

Grammatical differences between languages, like differences in word order, have a long history in research on worldwide linguistic variation (Greenberg, 1966; Dryer, 1992). However, being able to look at the usage of word order in specific contexts (and being able to directly compare exactly the same context across languages) is only possible by using parallel texts. For example, specific orders of more than two elements can be directly extracted from the parallel texts, like the order of demonstrative, numeral and noun "these two commandments" in Figure 7 (Cysouw, 2010).

For lack of space, we describe four more use cases only briefly: grammatical markers vs. morphology as devices to express grammatical features (Figure 8); differences in how languages use gram-

Alignments for verse: 42017032. Languages in order: deu-neue, fra-courant1997, eng-amplified

Alignments for verse: 40022027. Languages in order: deu-neue, fra-courant1997, eng-amplified
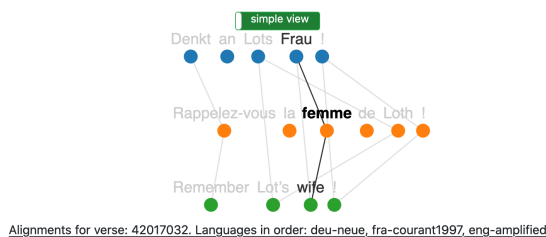
Figure 5: Use case 1, *lexical differentiation*. French "femme" has two different translations in English ("wife" and "woman") whereas German also conflates the two different meanings.
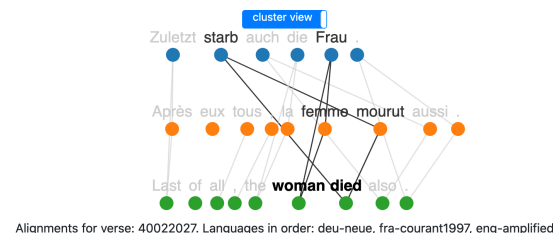

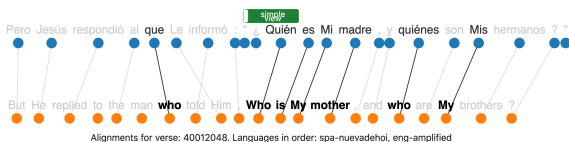Alignments for verse: 40012048. Languages in order: spa-nuevadehoi, eng-amplified

Figure 6: Use case 2, *grammatical differentiation*. English "who" has three different translations in this Spanish example: relative pronoun ("que"), and singular ("quién") and plural ("quiénes") interrogative pronoun.

Alignments for verse: 40022040. Languages in order: swh-union1997, eng-amplified

Figure 7: Use case 3, *word order variation*. The English order is demonstrative, numeral, noun whereas Swahili has noun, demonstrative, numeral.

Alignments for verse: 41015037. Languages in order: eng-amplified, crs

Figure 8: Use case 4, *grammatical markers*. In contrast to English, Seychelles Creole does not inflect verbs for tense and uses the past tense marker "ti" instead.

matical case (Figure 9, ablative/dative in Latin can correspond to five different cases in Croatian); and exploration of paraphrases (Figure 10). See the captions of the figures for more details.

7 Extension to Other Corpora

Our code is available on GitHub and can be generically applied: you can create a ParCourE instance for your own parallel corpus. Parallel corpora are essential for machine translation (MT); ParCourE's functionality is useful for analyzing the quality of a parallel corpus and the difficulty of the translation problem it poses. We give three examples **i)** Incorrect sentence alignments can be identified, e.g., cases in which a target sentence is matched with the merger of two sentences in the source: cf. Figure 11 where a short sentence in English is aligned with German and French sentences that also contain a second sentence that is missing in English. This functionality is particularly helpful for mined parallel corpora that tend to contain er-
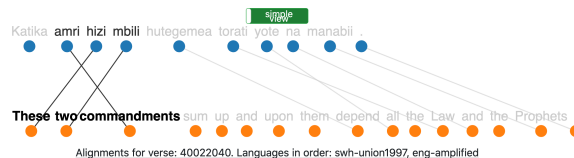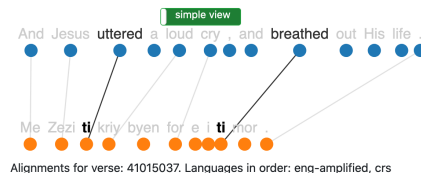
roneous sentence pairs. **ii)** Suppose an MT system trained on the parallel corpus makes a lexical error in a particular context c by mistranslating source word w_s with target word w_t. The LEXICON view can be consulted for w_s and the user can then click on the erroneous target word w_t to get back to a MULTALIGN view of aligned sentence pairs containing w_s and w_t. She can then analyze why the MT system mismatched c with these contexts. Examples of the desired translation are easy to find and inspect to support the formation of hypotheses as to the source of the error. **iii)** For multi-source approaches to MT (Zoph and Knight, 2016; Firat et al., 2016; Libovický and Helcl, 2017; Crego et al., 2010), ParCourE supports the inspection of all input sentences together. The MT system output can also be loaded into ParCourE for a view that contains all input sentences and the output sentence. Since any of the input sentences can be responsible for an error in multi-source MT, this facilitates analysis and hypothesis formation as to what caused a specific error.

7.1 Computing Infrastructure and Runtime

We did all computations on a machine with 48 cores of Intel(R) Xeon(R) CPU E7-8857 v2 with 1TB memory. In this experiment only one core was used.

We created a corpus of 5 translations in 4 languages, with around 31k parallel sentences (overally 155k sentences) and applied the ParCourE pipeline to it. Runtimes for different parts of the
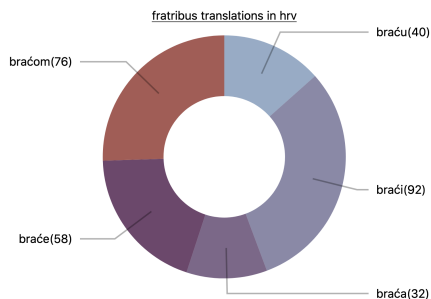
Figure 9: Use case 5, *morphology*. The Latin ending "ibus" in "fratribus" (dative/ablativ plural) corresponds to five different cases in Croatian: accusative, locative/dative, nominative, genitive, instrumental (clockwise starting from "braću").
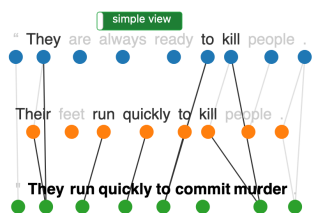

Figure 10: Use case 6, *paraphrases*. PBC is a rich source of paraphrases since high-resource languages have several translations (32 for English). ParCourE can be used to explore these paraphrases. Here, the paraphrases "kill" and "murder" are correctly aligned, "always ready" and "run quickly" are not.

pipeline are reported in Table 2. The installation of the package is straightforward and as shown in the table, it takes around 12 minutes to initiate ParCourE on a small corpus with 4 languages.

Method	Runtime
Conversion from CES to ParCourE format	153
Indexing with Elasticsearch	14
Alignment generation with Eflomal	537
Stats calculation	22
Overall	726

Table 2: Runtime in seconds for each part of the pipeline to initiate a ParCourE instance on a corpus with 4 languages and 31K parallel sentences.

8 Conclusion

Progress in multilingual NLP is an important goal of NLP and requires researching typological properties of languages. Examples include assessing language similarity for effective transfer learning, injecting inductive biases into machine learning models and creating resources such as dictionaries and inflection tables. To serve such use cases, we
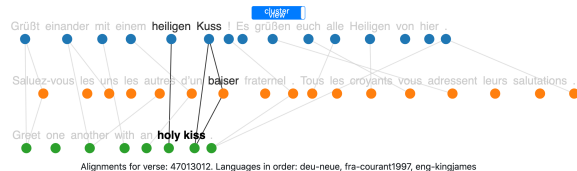

Figure 11: Use case 7, *quality analysis*. ParCourE makes it easy to analyze the quality of the parallel corpus. For this sentence, part of a Bible verse present in German and French is missing in English. Note that the alignment of *holy, heiligen* to French *fraternel* is not discovered.

have created ParCourE, an online tool for browsing a word-aligned parallel corpus of 1334 languages, and given evidence that it is useful for typological research. ParCourE can be set up for any other parallel corpus, e.g., for quality control and improvement of automatically mined parallel corpora.

Acknowledgments

9 Ethical Considerations

Word alignments and lexicon induction as tasks themselves may not have ethical implications. However, working on a biblical corpus requires special consideration of the following issues.

i) The Bible is the central religious text of Christianity and the Hebrew Bible that of Judaism. It contains strong opinions and world views (e.g., on divorce and homosexuality) that are not generally shared. We would like to emphasize that we treat the PBC simply as a multiparallel corpus, and the corpus does not necessarily reflect the opinions of the authors nor of the institutions funding the authors. **ii)** In a similar vein, while the PBC has great language coverage and allows for typological analysis, we need to be aware that languages might not be accurately and completely reflected in the PBC. The language used in the PBC might be outdated and is restricted to a relatively small subset of topics and thus cannot be considered a balanced and complete view of the language. **iii)** We also need to

be aware of selection bias. The PBC only covers a subset of the world's languages. The selection criteria are unknown and may be based on historical and cultural biases that we are not able to assess.

References

Alan Akbik and Roland Vollgraf. 2017. The projector: An interactive annotation projection visualization tool. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 43–48, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Mikko Aulamo, Sami Virpioja, and Jörg Tiedemann. 2020. OpusFilter: A configurable parallel corpus filtering toolbox. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 150–156, Online. Association for Computational Linguistics.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2).

Josep Maria Crego, Aurélien Max, and François Yvon. 2010. Local lexical adaptation in machine translation through triangulation: SMT helping SMT. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 232–240, Beijing, China. Coling 2010 Organizing Committee.

Michael Cysouw. 2010. Dealing with diversity: towards an explanation of NP word order frequencies. *Linguistic Typology*, 14(2):253–287.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. *CoRR*, abs/2101.08231.

Matthew S. Dryer. 1992. The Greenbergian word order correlations. *Language*, 68(1):80–138.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

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.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

David M. Eberhard, F. Simons Gary, and D. Fennig (eds.) Charles. 2020. *Ethnologue: Languages of the World*, 23rd edition. SIL International.

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas. Association for Computational Linguistics.

Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Jointly learning to align and translate with transformer models. 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 4453–4462, Hong Kong, China. Association for Computational Linguistics.

Timur Gilmanov, Olga Scrivner, and Sandra Kübler. 2014. SWIFT aligner, a multifunctional tool for parallel corpora: Visualization, word alignment, and (morpho)-syntactic cross-language transfer. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2913–2919, Reykjavik, Iceland. European Language Resources Association (ELRA).

Johannes Graën, Dominique Sandoz, and Martin Volk. 2017. Multilingwis2 extendash explore your parallel corpus. In *Proceedings of the 21st Nordic Conference on Computational Linguistics, NODAL-IDA 2017, Gothenburg, Sweden, May 22-24, 2017*, volume 131 of *Linköping Electronic Conference Proceedings*, pages 247–250. Linköping University Electronic Press / Association for Computational Linguistics.

Joseph H. Greenberg. 1966. *Language Universals: with special reference to feature hierarchies*. Janua Linguarum, Series Minor. Mouton, The Hague.

Harald Hammarstrm, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2020. Glottolog 4.3. Max Planck Institute for the Science of Human History.

Martin Haspelmath. 2003. The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In Michael Tomasello, editor, *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure (Volume 2)*, pages 211–242. Erlbaum, Mahwah, NJ.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Murathan Kurfal and Robert Östling. 2018. Word embeddings for 1250 languages through multi-source projection. In *Seventh Swedish Language Technology Conference*.

William D. Lewis and Fei Xia. 2008. Automatically identifying computationally relevant typological features. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

Jindrich Libovický and Jindrich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 196–202. Association for Computational Linguistics.

Thomas Mayer and Michael Cysouw. 2012. Language comparison through sparse multilingual word alignment. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 54–62, Avignon, France. 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).

Aaron Mueller, Garrett Nicolai, Arya D. McCarthy, Dylan Lewis, Winston Wu, and David Yarowsky. 2020. An analysis of massively multilingual neural machine translation for low-resource languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3710–3718, Marseille, France. European Language Resources Association.

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

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1).

Robert Östling. 2014. Bayesian word alignment for massively parallel texts. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pages 123–127. The Association for Computer Linguistics.

Robert Östling. 2015. Word order typology through multilingual word alignment. 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 205–211, Beijing, China. Association for Computational Linguistics.

Robert Östling and Jörg Tiedemann. 2016a. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Robert Östling and Jörg Tiedemann. 2016b. Efficient word alignment with Markov Chain Monte Carlo. *The Prague Bulletin of Mathematical Linguistics*, 106(1).

Sebastin Santy, Sandipan Dandapat, Monojit Choudhury, and Kalika Bali. 2019. INMT: Interactive neural machine translation prediction. 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): System Demonstrations*, pages 103–108, Hong Kong, China. Association for Computational Linguistics.

James Strong. 2009[1890]. *Strong's exhaustive concordance of the Bible*. Hendrickson Publishers.

Jörg Tiedemann. 2018. Emerging language spaces learned from massively multilingual corpora. In *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference, DHN 2018, Helsinki, Finland, March 7-9, 2018*, volume 2084 of *CEUR Workshop Proceedings*, pages 188–197. CEUR-WS.org.

Winston Wu, Nidhi Vyas, and David Yarowsky. 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. European Language Resources Association (ELRA).

Patrick Xia and David Yarowsky. 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. Asian Federation of Natural Language Processing.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*.

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.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34, San Diego, California. Association for Computational Linguistics.