Creating a Massively Parallel Bible Corpus

Thomas Mayer, Michael Cysouw
Research Unit Quantitative Language Comparison
Philipps University of Marburg
thomas.mayer@uni-marburg.de, cysouw@uni-marburg.de

Abstract

We present our ongoing effort to create a massively parallel Bible corpus. While an ever-increasing number of Bible translations is available in electronic form on the internet, there is no large-scale parallel Bible corpus that allows language researchers to easily get access to the texts and their parallel structure for a large variety of different languages. We report on the current status of the corpus, with over 900 translations in more than 830 language varieties. All translations are tokenized (e.g., separating punctuation marks) and Unicode normalized. Mainly due to copyright restrictions only portions of the texts are made publicly available. However, we provide co-occurrence information for each translation in a (sparse) matrix format. All word forms in the translation are given together with their frequency and the verses in which they occur.

Keywords: Bible corpus, parallel text, comparable corpus

1. Introduction

In recent years, linguists have become more and more aware of the necessity to collect significant amounts of primary data for as many languages as possible (Abney and Bird, 2010). This involves various steps, among others the elicitation of texts from native speakers and the digitization of existing material. In order to make larger amounts of data available to linguists, a central step is to pre-process the texts and provide them in a well-defined format.

In this paper, we report on the compilation of a corpus of primary data based on translations of the Bible in more than 800 different languages with more to follow soon.[1] In the context of creating such a massively parallel corpus, the main task is to make sure that the parallel structure of the texts is guaranteed. Another important aspect is to enable researchers to easily exploit the parallel structure for language comparison.

While bilingual text corpora have been popular among computational linguists since the advent of statistical machine translation (Brown et al., 1988), there have also been some efforts to compile parallel texts in more than one language. The most widely used multilingual text is the Europarl[2] corpus, a collection of proceedings of the European Parliament, which includes versions in 21 European languages. There also exist parallel texts for literary works (e.g. Harry Potter, Le Petit Prince, Master i Margarita) or translations from the web (e.g., OPUS, `http://opus.lingfil.uu.se`), mostly available for a set of closely related languages. However, only very few of them are freely available or can be regarded as massively parallel texts in the strict sense (Cysouw and Wälchli, 2007).

No other book has been translated into so many languages over such a long period of time as the Bible. Starting with its first translation, the so-called Septuagint, in 300 BC, the Bible is to the present day the object of the most intense translation activity worldwide (Noss, 2007). A growing number of Bible translations are now available in electronic form on the internet. Yet until now there is no large-scale parallel Bible corpus that allows researchers to easily get access to Bible texts and its parallelism in very many different languages.[3] In this paper, we will report on our ongoing effort to compile such a massively parallel Bible corpus for language research.

2. Bible translations

The current status of Bible translations is regularly summarized by the United Bible Societies in their annual Scripture Language Report. The figures of the most recent report from 2012 are given in Table 2.. By way of comparison, the total number of translations (either portions, New Testaments or complete Bibles) increased from 2,167 in 1996 to 2,551 in 2012. The number of complete Bible translations rose from 355 to 484 in the same time. In other words, almost 7% of the 7,105 known living languages for which Ethnologue[4] contains information have a complete Bible translation, 36% have at least portions of the Bible translated.

3. Current status of the Bible corpus

For the first version of the Bible corpus we collected translations from PNGscriptures[5] (188 texts), Bible.is[6] (372 texts), Scripture Earth[7] (197 texts) and Unboundbible[8] (97 texts). In addition, we included 140 Bible translations that were collected by Östen Dahl and were not already available from the resources mentioned above. The total number of unique Bible translations in the collection is currently 994. The translations have been assigned 837 different ISO-639-3 language codes.[9] The geographical distribution

[1] available at `http://paralleltext.info/data/`

[2] `http://www.statmt.org/europarl/`

[3] See (Resnik et al., 1999) and Mark Davies's Polyglot Bible (`http://davies-linguistics.byu.edu/polyglot/`) for earlier efforts.

[4] `http://www.ethnologue.com`, accessed on April 24th, 2013.

[5] `http://pngscriptures.org`

[6] `http://www.bible.is`

[7] `http://www.scriptureearth.org/`

[8] `http://unbound.biola.edu`

[9] The ISO codes for the Bible texts have been checked by means of text comparison (diffs and trigram similarities).

Continent or Region	Portions	Testaments	Bibles	Total
Africa	216	343	189	748
Asia	206	267	146	619
Oceania	135	273	40	448
Europe	110	41	63	214
North America	40	31	8	79
Caribbean, Central & South America	101	302	37	440
Constructed Languages	2	0	1	3
TOTAL	810	1257	484	2,538

Table 1: Statistical summary of languages in which at least one book of the Bible had been registered as of 2012 (Source: `http://www.unitedbiblesocieties.org/sample-page/bible-translation/`).

of the languages in the Bible corpus is shown in Figure 1. The number of languages per family is given in Table 3. We have several hundred further translations in our pipeline to be added in the near future.

In the 994 translations we found 41,964 different verse numbers (including apocrypha books that are not in the biblical canon of the 66 books). This number is much higher than the 31,102 verses that make up the King James Version. Some verses only occur in a very small number of languages. The verse that is most widely available in the texts is Mark 1:7, which has entries in 976 translations. The gospel according to Mark is usually considered to have the largest coverage (Nida, 1972, p. xvi) with respect to the number of translations.[10]

The number of verses per translation varies widely. The average number of verses per translation is 10,707 (with a standard deviation of 7,727 verses). The largest number of verses (36,986) is in the text of the English King James Version, which includes many apocrypha books. The smallest number of verses can be found in the text for the Papua New Guinea language Wedau [wed], which lists only 677 verses. The average number of words per translation is 408,973 (standard deviation: 367,572). The average vocabulary size (number of types) is 21,176 (standard deviation: 15,134).

4. File formats

Each Bible translation is prepared for further processing and stored in different files, which are made available as a .zip data package.[11] The actual text is contained in the Bible .txt file (Section 4.1.), whereas the .wordforms files (Section 4.2.) give an alphabetic listing of all word forms in the texts (with frequency of occurrence). Further, the sparse matrix .mtx files (Section 4.3.) provide a word $\times$ verse matrix of all word forms with the information in which verses each word form occurs.

All file names adhere to the conventions of the BCP 47.[12] The general structure of the file names is ISO-x-bible-TRANSLATION-VERSION where ISO gives the closest possible ISO 639-3 language code and the 'x' is the separator for private codes in BCP 47. The 'bible' tag indicates that it is part of the parallel Bible corpus (as we plan to add further massively parallel corpora in the future), while the TRANSLATION tag shows the name of the specific translations.[13] Finally, the VERSION tag gives the version number within our Bible corpus, which allows us to correct errors while retaining backwards compatibility. All old versions of the texts will remain accessible. The file names also serve as the identifiers for the website. Each verse in a Bible translation is thus given a unique URL. For example, `http://paralleltext.info/data/mri-x-bible-maori-v1/41/001/003/` gives access to Mark 1:3 of the Maori Bible translation (version 1).

We extracted bare base texts of the Bible from the websites without any headings, footnotes or cross-references, retaining capitalization as found in the original. The actual text in our corpus is tokenized and Unicode normalized. For the tokenization step, we separate all characters with white space that do not belong to the Unicode categories 'Ll' (Letter, Lowercase), 'Lu' (Letter, Uppercase), 'Lm' (Letter, Modifier), 'Lo' (Letter, Other), 'Lt' (Letter, Titlecase), 'LC' (Letter, Cased), 'Zs' (Separator, Space) and 'Nd' (Number, Decimal Digit), which serves the purpose to split all punctuation marks and other non-alphabetic symbols from words (by inserting spaces between words and punctuation). We also performed a manual check to correct for those languages where any of the non-alphabetic symbols represents a sound of a language. For instance, in the Bible translation of the Austronesian language Arifama-Miniafia [aai], the right single quotation mark (0x2019) stands for the glottal stop (Wakefield, 1992). Unfortunately, in many such cases the original texts available to us used the quotation mark both for the glottal stop as well as for marking quotations, so the separation of the two uses involved quite some manual work. In general, it turned out to be impossible to automatically separate punctuation from word forms without many errors. Therefore, all texts have been manually corrected.

[10] In general, there seems to be no overall agreement on the order in which books are translated and thus the availability of individual books. Ellingworth notes that the first book to be published in Tongan was Jonah in 1831 (Ellingworth, 2007, p. 136).

[11] `http://dataprotocols.org/data-packages/`

[12] `http://tools.ietf.org/html/bcp47`, accessed on October 9th, 2013.

[13] The translation tag can either represent a dialectal difference for a language or a specific translation. For instance, "wosera" and "maprik" are the translations of two dialects of Ambulas [abt], whereas "elberfelder" stands for a particular translation of the Bible into Standard German [deu].

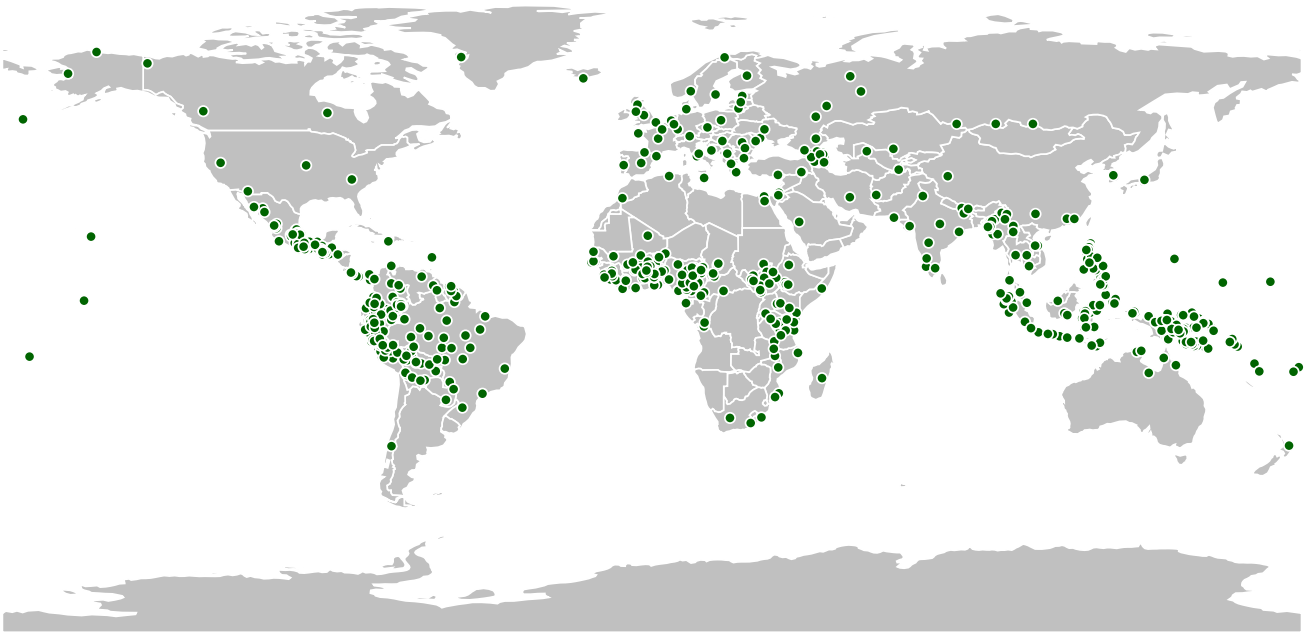

Figure 1: Geographical distribution of all languages in the Bible corpus

All texts are normalized using Unicode's Normalization Form C (NFC), which performs a canonical decomposition followed by a canonical composition.[14] This is to make sure that all symbols with diacritics are represented in the same form within each text. In addition, we provide metadata on translations together with their copyright information as far as it is indicated on the websites. The texts include no analyses. These will be added as stand-off annotations that will be prepared by ourselves or provided by others in the future.

4.1. Text files

The format for the Bible texts has the structure as shown in Figure 2. Each line contains two elements which are separated by a TAB. The first element is the verse ID and the second element contains the actual text. The verse ID contains information about the book, chapter and verse number and is structured as follows (e.g. line 3 in Figure 2 being 40001003):

- the first two digits represent the number of the book (e.g. 40 refers to the first book in the New Testament, the Gospel according to Matthew). The correspondences between book names and numbers are given in Table 2.
- the next three digits indicate the chapter (e.g. 001 refers to the first chapter in the book)
- the last three digits show the verse number (e.g. 003 refers to the third verse in the chapter)

One of the advantages of numerical verse IDs is that portions of the Bible text can be easily selected by giving the range of IDs that correspond to the portion. For example, selecting all books of the New Testament can be achieved by extracting all IDs that are larger than 40,000,000, i.e., all IDs that start with a book number of 40 or larger. Another benefit of numerical verse IDs is that they can directly serve as indices in a (sparse) matrix, which is useful when analyzing parallel texts with matrix algebra (Mayer and Cysouw, 2012).

The organization of the Bible into books, chapters and verses allows for a relatively fine-grained parallel structure of the corpus. However, the verse structure is not always identical for all Bible texts. Sometimes one verse in the first language contains the same information as two or more verses in the other. In this case, we list the relevant verse ID in the file without giving any text on the same line. An empty verse with only its verse ID shows that the content was probably merged with its previous verse(s). In contrast, the absence of a verse ID indicates that the content of that verse is not translated. For reasons of copyright protection, only the book of Mark is currently freely available for download. Please contact us directly for complete access.

4.2. Wordform files

Apart from the text files, we also preprocess each translation into two separate word forms and matrix files, which contain information about the complete available bible text. The .wordforms file contains all the word forms in the text (according to the tokenization procedure) in alphabetical order together with their frequency in the text. The line number in the word file serves as an ID for the word form in the sparse matrix file (see below). The word form "Aaron" in line 18 of Figure 3 occurs 352 times in this Bible translation and has the ID 18.

4.3. Sparse matrix files

In addition to the text and word files, we provide a sparse matrix .mtx file for each Bible translation in the corpus. The

[14] The normalization is performed with the `normalize()` method in Python's `unicodedata` package.

```
40001001 TAB The book of the generations of Jesus Chris... LF
40001002 TAB The son of Abraham was Isaac ; and the so ... LF
40001003 TAB And the sons of Judah were Perez and Zerah... LF
40001004 TAB And the son of Ram was Amminadab ; and the... LF
40001005 TAB And the son of Salmon by Rahab was Boaz ; ... LF
40001006 TAB And the son of Jesse was David the king ; ... LF
...
```

Figure 2: The file format for Bible texts

matrix file contains a list of all word forms together with the information in which verses they occur. So the matrix contains the information which words occur in which sentence, but not in which order. In effect, this is a randomized sentence structure, which is not copyright-protected but can still be used to extract co-occurrence statistics.

The information is given in the matrix market format (.mtx), which is used for exchanging and storing (sparse) matrices. The first (non-commented) line of the matrix file gives three integers, with the first two standing for the number of rows and columns in the matrix. The last integer indicates the number of entries in the (sparse) matrix. For instance, the example in Figure 4 shows a 13,487 $\times$ 41,964 sparse matrix with 718,568 entries. The subsequent lines in the example show entries for the word form ID 18 (the word 'Aaron' in Figure 3) together with its verse IDs (columns in the matrix). The 79,776th line in Figure 4 thus indicates that the word form 'Aaron' occurs in the 29,787th verse, which stands for the verse ID 44007040 (Acts 7:40). A correspondence table of all 41,964 verses to their verse IDs is also provided in the package.

```
...
Aaron                352
Abaddon                4
Abagtha                1
Abanah                 1
Abarim                 5
Abba                   3
Abda                   2
...
```

Figure 3: The .wordforms format for Bible texts

5. Acknowledgements

The authors would like to thank Östen Dahl for sharing his collection of Bible texts with us and giving much appreciated information on additional resources. We are also grateful to Bernhard Wälchli for his advice as well as Matthew Dryer and Harald Hammarström for useful hints to further resources. The present work has been funded by the Deutsche Forschungsgemeinschaft (DFG) within the project "Algorithmic corpus-based approaches to typological comparison".

6. References

Abney, S. and Bird, S. (2010). The human language project: Building a universal corpus of the world's languages. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden*, pages 88–97. Association for Computational Linguistics.

Brown, P. F., Cocke, J., Della-Pietra, S. A., Della-Pietra, V. J., Jelinek, F., Mercer, R. L., and Roossin, P. S. (1988). A statistical approach to language translation. In *Proceedings of the 12th International Conference on Computational Linguistics (COLING-88)*, pages 71–76.

Cysouw, M. and Wälchli, B. (2007). Parallel texts: using translational equivalents in linguistic typology. *Sprachtypologie und Universalienforschung STUF*, 60(2):95–99.

Ellingworth, P. (2007). From martin luther to the english revised version. In Noss, P. A., editor, *A History of Bible Translation*, pages 105–139. American Bible Society.

Mayer, T. and Cysouw, M. (2012). Language comparison through sparse multilingual word alignment. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 54–62, Avignon, France (April 23-24).

Nida, E. A. (1972). *The Book of a Thousand Tongues (Second Edition)*. United Bible Socities.

Noss, P. A. (2007). A history of bible translation: Introduction and overview. In Noss, P. A., editor, *A History of Bible Translation*, pages 1–25. American Bible Society.

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*, 33:129–153.

Wakefield, D. C. (1992). Miniafia organised phonology data. Draft.

```
    1  13487 41964 718568
    2  1 1004015
    3  1 1014020
    4  1 1016005
    5  ...
79776  18 29787
79777  18 33014
79778  18 33057
79779  18 33092
79780  19 14141
79781  19 18758
79782  19 19130
79783  ...
```

Figure 4: The .mtx matrix format for Bible texts

Old Testament				**New Testament**			
01	Genesis	20	Proverbs	40	Matthew	59	James
02	Exodus	21	Ecclesiastes	41	Mark	60	1 Peter
03	Leviticus	22	Song of Solomon	42	Luke	61	2 Peter
04	Numbers	23	Isaiah	43	John	62	1 John
05	Deuteronomy	24	Jeremiah	44	Acts	63	2 John
06	Joshua	25	Lamentations	45	Romans	64	3 John
07	Judges	26	Ezekiel	46	1 Corinthians	65	Jude
08	Ruth	27	Daniel	47	2 Corinthians	66	Revelation
09	1 Samuel	28	Hosea	48	Galatians		
10	2 Samuel	29	Joel	49	Ephesians		
11	1 Kings	30	Amos	50	Philippians		
12	2 Kings	31	Obadiah	51	Colossians		
13	1 Chronicles	32	Jonah	52	1 Thessalonians		
14	2 Chronicles	33	Micah	53	2 Thessalonians		
15	Ezra	34	Nahum	54	1 Timothy		
16	Nehemiah	35	Habakkuk	55	2 Timothy		
17	Esther	36	Zephaniah	56	Titus		
18	Job	37	Haggai	57	Philemon		
19	Psalms	38	Zechariah	58	Hebrews		
		39	Malachi				

Table 2: Books of the Bible together with their two-digit code

Family	No. languages	Family	No. languages
Austronesian	136	Barbacoan	2
Niger-Congo	128	Algic	2
Trans-New Guinea	106	Eyak-Athabaskan	2
Otomanguean	66	Yele-West New Britain	2
Indo-European	49	Eastern Trans-Fly	2
Afro-Asiatic	31	Paezan	2
Mayan	23	Border	2
Uto-Aztecan	20	Mapudungu	1
Sino-Tibetan	20	Puinavean	1
Quechuan	19	Ramu-Lower Sepik	1
Nilo-Saharan	16	Arai (Left May)	1
Tucanoan	14	Maxakalian	1
Maipurean	14	East Geelvink Bay	1
Altaic	12	South-Central Papuan	1
Language isolate	12	Huavean	1
Tupian	12	Constructed language	1
Creole	11	East New Britain	1
Sepik	10	Kartvelian	1
Chibchan	8	Tai-Kadai	1
Totonacan	8	Cahuapanan	1
Mixe-Zoquean	7	Arauan	1
Uralic	7	Zaparoan	1
Torricelli	7	Japonic	1
Cariban	6	Pauwasi	1
Australian	4	Chipaya-Uru	1
Austro-Asiatic	4	South Bougainville	1
Panoan	4	Karaj	1
Eskimo-Aleut	4	Pidgin	1
Jivaroan	4	Nambiquaran	1
Witotoan	4	Jicaquean	1
Jean	4	Aymaran	1
North Caucasian	4	Tequistlatecan	1
Chocoan	3	Harkmbut	1
East Birds Head-Sentani	3	North Bougainville	1
Guajiboan	3	Mosetenan	1
Dravidian	3	Senagi	1
West Papuan	3	Iroquoian	1
Tacanan	3	Yaguan	1

Table 3: Language families in the Bible corpus