Report: Adding PBC bibles and UDHR unicode texts to the 100LC Corpus

Chris Bentz

06.02.2019

1 Current status of the 100LC

With texts from the PBC and UDHR in unicode included, the current 100LC version features 114 texts of roughly 10M tokens in 62 of the 100 target languages. Numbers of texts, genera, and very rough estimations of number of tokens per language can be seen in `progress_100LC.csv`. Details of how texts were selected and added to the 100LC are given below. Also, there are some open to-do points in terms of adding further bibles, and in terms of basic text processing (adding white spaces). However, the latter might wait until the overall text collection has been finished.

2 Universal Declaration of Human Rights

The UDHR in unicode was downloaded on 30th January 2019 from the web[1]. At this point, it includes 454 files, i.e. translations of the UDHR. Overall, **43 languages** match with the 100LC sample. UDHR texts for these are added to the respective language specific folder in `100LC/Corpus`. Further information on these languages can be found in the file `progress_100LC.csv` The folder for UDHR texts is named `professional` according to the genus specification we have chosen earlier.

[1] https://unicode.org/udhr

2.1 Copyright for the UDHR

The original copyright statement given as a header in every UDHR text file is:

© 1996 – 2009 The Office of the High Commissioner for Human Rights
This plain text version prepared by the "UDHR in Unicode"
project, https://www.unicode.org/udhr.

Additionally, the following line on "Permissions" can be found at the original UDHR website[2]:

If UDHR translations or materials are reproduced, users should make reference to this website as a source by providing a link.

2.2 Multiple UDHR translations

For some language there are multiple UDHR translations available. These include:

- Mandarin Chinese (cmn): udhr_cmn_hans, udhr_cmn_hant (Chinese Mandarin (simplified) and Chinese Mandarin (Traditional))
- German (deu): udhr_deu_1996, udhr_deu_1901 (probably highly redundant)
- Modern Greek (ell): udhr_ell_monotonic, udhr_ell_polytonic (difference in markings of stress accents, i.e. "tone")
- Vietnamese (vie): udhr_vie, udhr_vie_han (written in Latin and Chinese characters)

2.3 UDHR texts: To-do list

- **Add white spaces around punctuation**
  The UDHR texts are not preprocessed to have white spaces around punctuation. This would have to be done to later facilitate accurate tokenization.

[2]https://www.ohchr.org/EN/UDHR/Pages/Introduction.aspx

- **Add white spaces between words**
  For some languages, white spaces (cmn, vie, mya) between words should be added if possible. Check for further issues with writing systems.

3 Parallel Bible Corpus

The PBC [1] was cloned from the github repository of Michael Cysouw end of January 2019. The respective website[3] is currently not available unfortunately. At this point, the corpus folder in the github repository contains 1774 files, i.e. bible translations. The bibles for overall **49 languages** were added to the 100LC/Corpus folder. Further information on these languages can be found in the file `progress_100LC.csv`. The folder for Bible texts is named `non-fiction/written` according to the genus specification we have chosen earlier.

3.1 Copyright in the PBC

Some bibles in the PBC were not included in the 100LC corpus due to copyright issues, namely, if the copyright information given in the file header explicitly prohibits the redistribution of that particular text, or if the organization responsible for the translation is known to explicitely prohibit any usage of the texts (e.g. all the "newworld" bibles come from the Watch Tower Society). For example, the "copyright_long" given for the Kannada (kan) bible text (kan-x-bible.txt) states:

These Scriptures [...] May be distributed without modification in electronic form for non-commercial use. However, they may not be hosted on any kind of server (including a Web or ftp server) without written permission [...]

Another example includes the Bible in Hebrew (heb-x-bible.txt):

The Habrit Hakhadasha/Haderekh is copyrighted and has been made available on the internet for your personal use only. Any other use including, but not limited to, copying or reposting on the internet is prohibited. These Scriptures may not be altered or modified in any form and must remain in their

[3] http://paralleltext.info/data/

original context. These Scriptures may not be sold or otherwise offered for sale.These Scriptures are not shareware and may not be duplicated. These scriptures are not public domain.

3.2 Multiple bibles for the same language

For several languages (ISO codes) there are different bible translations available in the PBC. In this case, a minimum of one translation and a maximum of three translations was chosen to be included in the 100LC. The criterion to include/exclude translations was a) copyright, b) whether texts come from vastly different time periods (e.g. in German, there is one translation form Luther 1545, one version from 1871, and one from 1997), c) different writing systems (e.g. Hindi devanagari and latin translations), d) whether translations seemed different from a quick inspection of the texts.

It is worth considering whether we should generally only choose one translation per language in order to avoid redundancy. The names of original PBC text files chosen for languages with multiple translations are given below:

- Chamorro (cha): cha-x-bible-2003, cha-x-bible-1908
- German (deu): deu-x-bible-neue, deu-x-bible-elberfelder1871, deu-x-bible-luther1545
- Modern Greek (ell): ell-x-bible-hellenic1, ell-x-bible-modern2009
- English (eng): eng-x-bible-newsimplified, eng-x-bible-kingjames, eng-x-bible-darby
- Basque (eus): eus-x-bible-batua, eus-x-bible-Hautin1571
- Fijian (fij): fij-x-bible-1974
- Finnish (fin): fin-x-bible-1766, fin-x-bible-1992
- French (fra): fra-x-bible-perret, fra-x-bible-louissegond
- Guarani (gug): gug-x-bible
- Hindi (hin): hin-x-bible-bsi, hin-x-bible-latin
- Indonesian (ind): ind-x-bible-suciinjil, ind-x-bible-terjemahanbaru

- Georgian (kat): kat-x-bible
- Korean (kor): kor-x-bible-revised
- Burmese (mya): mya-x-bible-common, mya-x-bible-1835
- Persian (pes): pes-x-bible-1995, pes-x-bible-2007
- Plateau Malagasy (plt): plt-x-bible-1865, plt-x-bible-romancatholic
- Sango (sag): sag-x-bible
- Russian (rus): rus-x-bible-synodal, rus-x-bible-kulakov
- Spanish (spa): spa-x-bible-reinavaleracontemporanea, spa-x-bible-lapalabra
- Swahili (swh): swh-x-bible-union1997
- Tagalog (tgl): tgl-x-bible-1905, tgl-x-bible-1996
- Thai (tha): tha-x-bible-standard2011, tha-x-bible-kjv
- Turkish (tur): tur-x-bible-2009
- Vietnamese (vie): vie-x-bible-2002, vie-x-bible-1926nocompounds
- Yoruba (yor): yor-x-bible-2010

3.3 Bible texts: To-do list

- **Unify verse identifiers to format of the 100LC**
  The texts that come from the PBC are already processed to have white spaces before and after punctuation. This helps considerably with tokenization. They also have verse identifiers as exemplified with the first line of the King James Version in English below:

  01001001 In the beginning God created the heaven and the earth .

  The first two digits of the identifier refer to the book, the next three digits to the chapter, and the last three digits to the verse. See [1] for a list of book identifiers. These identifiers need to be translated into an XML compatible format.

- **Add missing bibles**
  For some more languages there are bibles available independent of the PBC that can be added to the 100LC:

  - Zulu (zul): either scrape from www.wordproject.org, or use XML version at http://christos-c.com/bible/ and transform into raw text file
  - Modern Hebrew (heb): also available as XML at http://christos-c.com/bible/
  - Kannada (kan): also available as XML at http://christos-c.com/bible/
  - Japanese (jpn): also available as XML at http://christos-c.com/bible/
  - Mandarin (cmn): also available as XML at http://christos-c.com/bible/
  - Canela-Kraho (ram): available as pdf at https://scriptureearth.org/data/ram

- **Add white spaces**
  For some languages white spaces (cmn, vie, mya) between words should be added if possible. Check for further issues with writing systems.

References

[1] Thomas Mayer and Michael Cysouw. Creating a massively parallel bible corpus. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 3158–3163, 2014.