Dataset	Train	Dev.	Test	Total
GSW-BE-Novel	2,667	218	183	3,251
GSW-BE-Wikipedia	–	180	67	247
GSW-VS-Radio	463	100	50	613
GSW-ZH-Wikipedia	–	45	50	95
GSW-BE-Bible	–	–	126	126
GSW-Archimob	40,159	2,710	2,710	45,579
GSW-ZH-Lexicon1	1,527	–	–	1,527
GSW-BE-Lexicon2	1,224	–	–	1,224

Table 1: GSW/DE parallel datasets partitioned for MT training, tuning and testing, with sizes in numbers of parallel sentences. The lexicons (last two lines) were not used for testing, and 183 additional lines from GSW_BE_Novel are kept apart for future testing.

written originally in GSW and then translated into DE. Among the growing body of literature published in Swiss German, we found only one volume translated into High German and available in electronic form: *Der Goalie bin ig* (in English: *I am the Keeper*), written in Bernese by Pedro Lenz in 2010. The DE translation stays close to the original GSW-BE text, therefore sentence-level alignment was straightforward, resulting in 3,251 pairs of sentences with 37,240 words in GSW-BE and 37,725 words in DE.

GSW-BE-Wikipedia and **GSW-ZH-Wikipedia**. The Alemannic version of Wikipedia[2] appeared initially as a promising source of data. However, its articles are written not only in Swiss German, but also in other Alemannic dialects such as Alsatian, Badisch and Swabian. As its contributors are encouraged to write in their own dialects, only a few articles are homogeneous and have an explicit indication of their dialect, using an Infobox with one of the six labels indicated above. Among them, even fewer have an explicit statement indicating that they have been translated from High German (which would make the useful as parallel texts). We identified two such pages and sentence-aligned them to serve as test data: "*Hans Martin Sutermeister*" translated from DE into GSW-BE and "*Wädenswil*" from DE into GSW-ZH.[3]

GSW-VS-Radio. A small corpus of Valaisan Swiss German (also called *Wallisertiitsch*) has been collected at the Idiap Research Institute (Garner et al., 2014).[4] The corpus consists of transcriptions of a local radio broadcast[5] translated into High German.

GSW-BE-Bible. The Bible has been translated in several GSW dialects, but the only electronic version available to us were online excerpts in Bernese.[6] However, this is not translated from High German but from a Greek text, hence the alignment with any of the German Bibles is problematic.[7] We selected the contemporary *Gute Nachricht Bibel* (1997) for its modern vocabulary, and generated parallel data from four excerpts of the Old and New Testament, while acknowledging their particular style and vocabulary. The following excerpts were aligned: *Üse Vatter*, *D Wienachtsgschicht*, *Der barmhärzig Samaritaner* and *D Wält wird erschaffe*.

GSW-Archimob. Archimob is a corpus of standardized Swiss German (Samardžić et al., 2016), consisting of transcriptions of interviewees speaking Swiss German, with a word-align normalized version in High German.[8] The interviews record memories of WW II, and all areas of Switzerland are represented. In most cases, the normalization provides the corresponding High German word or group of words, but in other cases it is Swiss German with a standardized orthography devised by the annotators. Using a vocabulary of High German, we filtered out all sentences whose normalizations included words outside this vocabulary. In other words, we kept only truly High German sentences, along with their original Swiss German counterparts, resulting in about 45,000 GSW/DE word-aligned sentence pairs.

GSW-ZH-Lexicon and **GSW-BE-Lexicon**. The last two parallel resources are vocabularies, i.e. lists of GSW words with their DE translation. As such, they are useful for training our research systems, but not for testing them. The first one is based on *Hoi Zäme*, a manual of Zürich Swiss German intended for High German speakers. The data was obtained by scanning the printed version, performing OCR[9] and manually aligning the result. Although the book contains also parallel sentences, only the bilingual dictionary was used in our study, resulting in 1,527 words with their translations. A similar dictionary for Bernese (GSW-BE vs. DE) was found online[10] with 1,224 words for which we checked and corrected the alignments.

2.3. Monolingual Resources

The Phonolex dictionary, a phonetic dictionary of High German,[11] was used for training our grapheme-to-phoneme converter (see Section). It contains High German words with their phonetic transcriptions.

[2] http://als.wikipedia.org

[3] These pages are respectively available at https://de.wikipedia.org/wiki/Hans_Martin_Sutermeister (High German), https://als.wikipedia.org/wiki/Hans_Martin_Sutermeister (Bernese), https://de.wikipedia.org/wiki/W%E4denswil (High German), and https://als.wikipedia.org/wiki/W%E4denswil (Zurich Swiss German).

[4] www.idiap.ch/dataset/walliserdeutsch

[5] Radio Rottu, http://www.rro.ch.

[6] www.edimuster.ch/baernduetsch/bibel.htm

[7] www.die-bibel.de/bibeln/online-bibeln/

[8] http://www.spur.uzh.ch/en/departments/korpuslab/ArchiMob.html

[9] Tesseract: https://github.com/tesseract-ocr/

[10] www.edimuster.ch/baernduetsch/woerterbuechli.htm

[11] www.bas.uni-muenchen.de/forschung/Bas/BasPHONOLEXeng.html. We also use it to find OOV words.