Language-classified Open Subtitles (LACLOS)
Download, extraction, and quality assessment

Technical Report

Adrien Barbaresi
Berlin-Brandenburgische
Akademie der Wissenschaften
barbaresi@bbaw.de

Contents

1 Introduction

1.1 Interest

Psycholinguistics In the context of psychological research, text corpora such as subtitles are used to derive frequency lists. The frequencies are used in correlations tests with reaction and latency times gained in experiments based on lexical decision and/or naming. In that sense, the quality of the corpus resource is related to its explanatory power and prediction potential.

More specifically, Brysbaert and New (2009) showed word frequencies gained from movie subtitles were superior to frequencies from classical sources in explaining variance in the analysis of reaction times from lexical decision experiments. In fact, the explanatory power of subtitles in psychological experiments has been found to be high, not only for American English, used in the first experiment (Brysbaert & New, 2009), but for many languages, ranging from Dutch (Keuleers, Brysbaert, & New, 2010) to Chinese (Cai & Brysbaert, 2010).

The reason for this superiority is still somewhat unclear (Brysbaert et al., 2011). It may stem from the fact that subtitles resemble spoken language, while traditional corpora are mainly compiled from written language (Heister & Kliegl, 2012). In that sense, it seems feasible to draw an analogy between subtitles and spoken language.

Linguistics Besides, subtitle corpora may also be relevant to linguistic studies, not only in the form from frequency lists. They may offer a more down-to-earth language sample which is closer to everyday life and spoken corpora. This is at least what the psychological studies suggest, saying that the subtitles are better predictors because they are less abstract than traditional written corpora.

Potential advantages in the case of lexicography include the discovery of new words and senses, or example sentences for words/senses which are known to exist but cannot be found in standard written corpora.

A general linguistic interest resides in the investigation of language use beyond traditional written corpora, as well as in the exploration of language patterns close to or derived from spoken variants, for example on a syntactic level.

Computational linguistics Potential interests in computational linguistics include language modeling, since there are tasks for which subtitles corpora may perform better than other types of corpora, and tools hardening, concerning morphology or word sense disambiguation for instance.

Interest at the BBAW At the BBAW, the construction of a subtitles corpus originates from the DLexDB project (Heister et al., 2011). Its purpose is to complement the use of the DWDS corpus (Geyken, 2007) to derive frequency lists. Moreover, the subtitles have found their way into comparative studies concerning the corpora of the DWDS project (Barbaresi & Würzner, 2014) or other specific web corpora (Barbaresi, 2013).

1.2 Overview

Corpus building included the following main phases, which are detailed in the sections below:

1. Search for subtitles files
2. Retrieval of metadata
3. Data download
4. Data processing

The processing chain takes text files as input, more precisely several subtitles formats (MicroDVD, SubViewer, SAMI, SSA, TXT). The output is in form of text files (TXT format) or XML format following the TEI guidelines.[1]

2 Retrieval

Source The subtitles are retrieved from the OpenSubtitles project[2], a community-based web platform for the distribution of movie and video game subtitles.

Search for subtitles files The subtitle files are searched for using two different sources: first by sifting through the dumps provided by OpenSubtitles and carrying out crosschecks to discover other resources; and second by querying the XMLRPC API systematically, i.e. for each known subtitle ID, in order to find those who are in German according to metadata.

Retrieval of metadata The full metadata are also retrieved using the XMLRPC interface for the texts classified as being in German. Each video document is identified by an IMDB number which could theoretically make metadata completion using other sources possible (for example `imdb.com` itself).

Drawbacks The drawbacks experienced during retrieval are twofold: on one hand there are growing restrictions on download frequency, and on the other hand the quality of the website in terms of information architecture could be improved (database access is sometimes inconsistent).

3 Processing

3.1 Processing steps

Data processing encompasses the following major steps:

1. Normalization
    - Unicode conversion and repairing
      The default working format is UTF-8.
    - Identification of subtitle format
      There are five main known formats: MicroDVD, SubViewer, SAMI, SSA, and TXT.

[1] http://www.tei-c.org/
[2] http://opensubtitles.org

2. Text cleaning

   - Removal of markup and text cleaning, based on file format detection
     Mostly time specifications, advertisements, typography, and so-called ASCII art.

3. Formatting of the output

   - Optional fusion of frames into sentences
     Performed by a basic sentence boundary detection.
   - Optional conversion from text to XML TEI format.

3.2 Example

Raw data:

```
909
01:28:16,334 --> 01:28:19,202
<i>Ich genieße einfach</i>
<i>den Rest des Sommers.</i>

910
01:36:09,932 --> 01:36:13,141
Copyright EUROTAPE 2013
Untertitel: Cosima Ertl u. a.
```

Result:

```
Ich genieße einfach
den Rest des Sommers.
```

3.3 Known issues and design decisions

Format and encoding Format-related issues include the existence of several formats as well as encoding and markup irregularities in both encoding and markup, so that robustness is paramount.

There are obviously cases where UNIX-tools such as *file* and *iconv* fail to detect the proper encoding or translate it properly, probably because of previous unduly assessed encodings. A typical case are fir instance files which most probably were natively encoded in Windows-/CP-1252 but which were processed and destructively re-encoded as being latin-1/ISO-8859-1. LACLOS does not fix this problem, it merely contains the damage by applying a series of oneliners.

Content Content-related issues the existence of several subtitles for the same film which require heuristics to choose the potentially better one, flaws of OCR methods used on subtitles which require error correction, multilingual documents, and spam or advertising.

- Partially addressed issues

  1. There may be several versions (i.e. files) for the same film, although this problem rarely occurs concerning German subtitles since they are more than ten times less numerous than the English ones for example. In the case where there are cases where several files are available, heuristics can be used to choose from the different versions. The default is to select the subtitle file which has been downloaded the most.

2. Multi-lingual documents (see quality assessment below)
3. Spam or advertising, most frequently for a subtitle "brand" or a movie release team, including exotic markup, which is easy to detect, but also full sentences such as "*Normalerweise hat Qualität ihren Preis ... / doch bei uns kriegt ihr sie umsonst !*" [3]

- Problems left untouched

1. There are files which are the result of an optical character recognition which failed partially, leaving vowels out or turning all "i" in "l". It has not been attempted to remedy this phenomenon.
2. Cases have been reported where the DVD-menu and not the actual content of the subtitles have been framed. The files which are concerned are easy to filter out as to their size. There is no automatic procedure to see if a better subtitle file is available and/or to replace it.

Normalization of tokens No normalization of any kind has been attempted on token level, which means that possible divergent orthographic forms, be it because of linguistic variants or because of typos or digitalization mistakes, are left as such.

Paratext within subtitles The paratext within subtitles consists of scene descriptions and indications for hearing impaired. It is not a clear case of markup, since it is linked to film and subtitle content and usually written in plain English.

Nonetheless, as it does not correspond to actual utterances in the video, this paratext was excluded from the content for psycholinguistic purposes and marked as such in the XML export version of the corpus. This is done where the paratext is clearly identifiable, for instance because of particular punctuation styles, but not in the other cases, which include for example story introduction at the beginning of a film or epilogues at the end.

4 Result

4.1 Intrinsic quality assessment

In general, user-generated content on the Web comes with an inherent unevenness to smooth out. Subtitles are no exception, they can be of different origin and nature, but also mixed quality, so that design decisions are not necessarily clear-cut.

Most of the indicators for internal quality assessment are token-based, they can be roughly split into the following categories: N-gram analysis (from tokens/unigrams to 5-grams of tokens), language identification (spell checker and probabilistic models), annotation toolchain, and analysis of results (elementary text statistics).

Frequent n-grams are extracted and must be manually scrutinized in order to find potential caveats.

The spell check library used, `enchant`, allows the use of a variety of spell-checking backends, like aspell, hunspell or ispell, with one or several locales.[4] A significant proportion of unknown words as well as the presence of words in a concurrent language, in that case English, are indicators which can be trigger automatically the exclusion of texts above a certain threshold.

[3] "Normally, quality does have a price tag... / but not with us!"

[4] http://www.abisource.com/projects/enchant/

Additionally, the language identification system `langid.py` (Lui & Baldwin, 2012) is open-source[5] and it incorporates a pre-trained model for 97 languages. The output of the software, i.e. a language code and a confidence interval, is used to find outliers in the subtitles collection.

About 10% of the original files are not used because of encoding errors, improper OCR-use but mostly because they were detected as not being in German.

4.2 Status (2013 version)

11,956 files have been downloaded. According to http://www.opensubtitles.org/de/statistics there were 17,116 available subtitles for German at that point. It is unclear why files are apparently missing, it may be explained by the counting of several subtitle versions for a given film.

10,795 documents remain at the end of the processing chain, meaning that a total of 1,161 files (9.7%) have been blacklisted because of encoding errors, improper OCR-use but mostly because they were detected as not being in German.

There are still irregularities to be detected by about 1.5% of the texts concerning markup or unicode-related issues, but it only affects a minority of tokens (around 0.001% of the collection), so that no further texts have been removed.

The corpus size is 56,276,568 tokens, which makes it an interesting resource, since there are probably enough different texts and enough tokens to cover various enunciation situations as well as to provide somewhat reliable word frequencies.[6]

4.3 Linguistic processing

The corpus has been automatically split into tokens and sentences with the help of WASTE, Word and Sentence Tokenization Estimator (Jurish & Würzner, 2013), a statistical tokenizing approach based on a Hidden Markov Model (HMM), using the standard DTiger model.

Subsequently, the resulting tokens have been assigned with possible PoS tags and corresponding lemmata by the morphological analysis system TAGH (Geyken & Hanneforth, 2006). The HMM tagger *moot* (Jurish, 2003) has then selected the most probable PoS tag for each token given its sentential context. In cases of multiple lemmas per best tag the one with the lowest edit distance to the original token's surface is chosen.

5 Software

The software used to download and preprocess the subtitles, LACLOS[7], is available under an open source license: https://github.com/adbar/laclos

References

Barbaresi, A. (2013). *Two comparable corpora of German newspaper text gathered on the web: Bild & Die Zeit* (Tech. Rep.). ICAR / ENS Lyon.

[5] https://github.com/saffsd/langid.py

[6] Brysbaert and New (2009) expect such a limit to be around 10-15 Mtokens.

[7] The acronym stands for LAnguage-CLassified OpenSubtitles.

Barbaresi, A., & Würzner, K.-M. (2014). For a fistful of blogs: Discovery and comparative benchmarking of republishable German content. In *KONVENS 2014, NLP4CMC workshop proceedings* (pp. 2–10). Hildesheim University Press.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, *58*(5), 412–424.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990.

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies ased on film subtitles. *PLoS One*, *5*(6), e10729.

Geyken, A. (2007). The DWDS corpus: A reference corpus for the German language of the 20th century. In C. Fellbaum (Ed.), *Collocations and Idioms: Linguistic, lexicographic, and computational aspects* (pp. 23–41). Continuum Press.

Geyken, A., & Hanneforth, T. (2006). TAGH: A Complete Morphology for German based on Weighted Finite State Automata. In *Finite State Methods and Natural Language Processing* (Vol. 4002, pp. 55–66). Springer.

Heister, J., & Kliegl, R. (2012). Comparing word frequencies from different German text corpora. In K.-M. Würzner & E. Pohl (Eds.), *Lexical Resources in Psycholinguistic Research* (pp. 27–44). Potsdam Cognitive Science Series. (vol.3)

Heister, J., Würzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., et al. (2011). dlexDB – eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychologische Rundschau*(62), 10–20.

Jurish, B. (2003). *A Hybrid Approach to Part-of-Speech Tagging* (Final Report). Kollokationen im Wörterbuch, Berlin-Brandenburgische Akademie der Wissenschaften.

Jurish, B., & Würzner, K.-M. (2013). Word and Sentence Tokenization with Hidden Markov Models. *JLCL*, *28*(2), 61-83.

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new frequency measure for Dutch words based on film subtitles. *Behavior Research Methods*, *42*(3), 643–650.

Lui, M., & Baldwin, T. (2012). langid.py: An Off-the-shelf Language Identification Tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012).*