Data in Brief 50 (2023) 109565

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

ChatSubs: A dataset of dialogues in Spanish, Catalan, Basque and Galician extracted from movie subtitles for developing advanced conversational models

Ksenia Kharitonova[a], Zoraida Callejas[a,b,*], David Pérez-Fernández[c], Asier Gutiérrez-Fandiño[d], David Griol[a]

[a] Department of Software Engineering, University of Granada, Granada, Spain
[b] Research Centre for Information and Communication Technologies (CITIC-UGR), University of Granada, Granada, Spain
[c] Universidad Autónoma de Madrid, Madrid, Spain
[d] LHF Labs, Bilbao, Spain

ARTICLE INFO

Article history:
Received 1 August 2023
Revised 4 September 2023
Accepted 5 September 2023
Available online 14 September 2023

Dataset link: ChatSubs: A dataset of movie dialogues in Spanish, Catalan, Basque and Galician (Original data)

Keywords:
Dialogue
Conversation
Chatbots
Conversational AI
Speech
Natural language processing

ABSTRACT

The ChatSubs dataset [5] contains dialogue data in Spanish and three of Spain's co-official languages (Catalan, Basque, and Galician). It has been obtained from OpenSubtitles, from which we have gathered the movie subtitles in our languages of interest and processed them to generate clearly segmented dialogues and their turns. The data processing code is publicly accessible. The result is 206.706 JSON files with more than 20 million dialogues and 96 million turns, which represents one of the biggest dialogue corpus available, as other similar datasets in better resourced languages do not reach 500k dialogues or present less defined conversations. Thus, the ChatSubs dataset is an ideal resource for research teams that are interested in training dialogue models in Spanish, Catalan, Basque, and Galician.

© 2023 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

* Corresponding author.
E-mail address: zoraida@ugr.es (Z. Callejas).
Social media: @zoraidacallejas (Z. Callejas)

https://doi.org/10.1016/j.dib.2023.109565
2352-3409/© 2023 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Specifications Table

Subject	Artificial intelligence, Natural language processing
Specific subject area	Conversational systems
Data format	Textual data, python scripts
Type of data	We contribute a dataset of textual JSON files that contain movie dialogues extracted after filtering and cleaning the raw subtitle files. We also provide the code in Python that was used to implement the processing.
Data collection	The raw subtitles are accompanied by the metadata that includes the information about the subtitle language in form of tags. Using those tags we filtered the raw data to include only subtitles in Spanish, Catalan, Basque and Galician which belong to the scope of the Conversa research project. However, the code we provide allows for building a similar dataset in any language.
Data source location	Primary data taken from www.opensubtitles.org
Data accessibility	Repository name: Zenodo Data identification number: 10.5281/zenodo.8220853 *Direct URL to data:* https://zenodo.org/record/8220853

1. Value of the Data[1]

- Dialogue systems able to generate natural and context-aware conversations require extensive, structured dialogue data that emulates human speech, specifies dialogue turns, and maintains semantic separation. Previous work on large dialogue corpora either lacked clear dialogue segmentation [4], had noisier data leading to a more arbitrary separation heuristic [1] or had insufficient volume [2]. Unlike those corpora, our corpus of movie dialogues includes a cleaner segmentation of both dialogue and turn segmentation, preventing trained models from producing generic responses [1,4] and contains more than 20 millions of dialogues, ample data for training coherent conversational systems.
- Due to its expansive and varied movie source base, the data encompasses a broad spectrum of conversational topics. It closely mirrors real-life speech, featuring elements such as dialects, idioms, and slang.
- Languages different from English, particularly under-resourced ones, lack substantial conversational textual data. While creating a Spanish corpus is easier, it is challenging for local languages like Catalan, Basque and Galician [6]. To our knowledge, we provide one of the most extensive dialogue data corpus in these languages.
- ChatSub can be employed to train AI systems that rely on dialogue, including chatbots, voice assistants, and other interactive agents. Furthermore, its application extends across all fields of artificial intelligence and machine learning that involve text processing, especially given the large volume of data, making it ideal for systems that use text as input and/or output.
- We hope that this dataset will be of use to academic, public and industrial researchers as well as to the general public that is interested in conversational and interactive applications. This data can be used for non-commercial training and testing purposes.

2. Data Description

ChatSubs contains 206.706 JSON files in Spanish, Catalan, Basque and Galician with 20.254.311 dialogues and 96.925.151 turns (see Table 1 for a detailed breakdown).

The archive that we share contains four datasets, one for each language: open_subtitles_ca (Catalan), open_subtitles_es (Spanish), open_subtitles_eu (Basque), open_subtitles_gl (Galician). Every folder contains the CSV metadata file (see Table 2). The metadata file follows the original tab-delimited format that accompanies the dump of the raw subtitles.

[1] https://www.opensubtitles.org/addons/export_languages.php (Last accessed: July 2023)

Table 1
Main features of the ChatSubs dataset.

Language	Number of output files	Number of dialogues	Number of turns	Average length of dialogues (# of turns)
ca	1.148	91.945	420.178	4,63
gl	659	63.098	284.547	4,35
eu	844	115.796	533.755	4,39
es	204.055	19.983.472	95.686.671	5,05
Total	206.706	20.254.311	96.925.151	-

Table 2
Information in metadata file (export.txt).

Field	Meaning
MovieID	Identifier of the movie
IDSubtitleFile	Identifier of the subtitle file
SubLanguageID	Subtitle language code (ISO 639-1)
IDSubtitle	Identifier for the subtitle
SubActualID	Actual CD or disc number for the subtitle (applicable for multi-disc movies or TV show episodes)
SubSumCD:	Total number of CDs or discs in a subtitle set
SubFormat	Format of the subtitle file (srt, sub, etc)
MovieName	Name of the movie
MovieYear	Year the movie came out
MovieImdbID	IMDB identifier of the movie
UserRank	User ranking
SubDownloadsCnt	The amount of times the subtitle was downloaded
SeriesIMDBParent	For a series episode - a IMDB identifier for a parent series
SeriesSeason	For a series episode - number of a series season
SeriesEpisode	For a series episode - number of a series episode
SubHearingImpaired	Whether a subtitle is adapted for hearing impaired audience
Encoding	Encoding of a subtitle file

Each JSON file has a unique *IDSubtitleFile* as a name, same as the original file it is extracted from.

To ensure replicability, the folder structure of the original dump is fully preserved in ChatSubs. The file structure can be understood using the 1953288724.jsonl file from the open_subtitles_ca dataset as an example. Here, the last four digits of the filename, i.e., 8724, are reversed, resulting in 4278. Starting from the root open_subtitles_ca, this reversed sequence forms a series of subfolders leading to the JSON file. The full path becomes open_subtitles_ca/4/2/7/8/1953288724.jsonl.

As explained in the next section, the delimitation of dialogues was not evident. Also the original excerpts corresponding to the time period between the beginning and the end of a subtitle appearance on the screen do not necessarily represent a dialogue turn. In order to clearly identify dialogues and be sure that the dialogue turn is correctly preserved we implemented a thorough cleaning, merging and processing of the original subtitle parts. The code corresponding to this procedure to process the raw dumps and obtain ChatSubs JSON files is shared on GitHub and in Zenodo[2].

As shown in Fig. 1, every JSON file contains a dictionary with the key 'dialogues' that has as value an array of strings where every string corresponds to one separate dialogue in the film. The dialogues in the array are presented in chronological order by their appearance in the movie. Every dialogue string contains an array of dialogue turns separated by the newline separator.

[2] https://github.com/conversa-ai/ChatSubs and ⟨13:italic ⟩https://zenodo.org/record/8220853⟨/13:italic⟩ (Last accessed: August 2023)

```
{
   “dialogues”:
      [
          "Sentinela Diária.\nUm momento, por favor.",
          "O prefeito não quer que o artigo do preço do gás seja impresso.\nAté que tudo seja resolvido.\nDiga ao prefeito que estou insultado.\nEu nunca arriscaria a integridade deste jornal por uma jogada política dele.\nMuito bem, senhor.",
          "Sei que sente falta da sua mãe.\nEu também.",
          …
     ]
}

English translation:
{
   “dialogues”:
      [
          "Daily watch.\nA moment, please.",
          “The mayor doesn't want the gas price article in print.\nUntil everything is resolved.\nTell the mayor I'm insulted.\nI would never risk the integrity of this newspaper for his political move.\nVery well, sir.”,
          "I know you miss your mother.\nMe too.",
          …
     ]

}
```


Fig. 1. Extract from a JSON file in Galician from the ChatSubs dataset (1952874000.jsonl).

For example a dialogue *"Sentinela Diária.\nUm momento, por favor."* from Fig. 1 corresponds to two turns:

1. Sentinela Diária. (Daily watch)
2. Um momento, por favor. (A moment, please)

3. Experimental Design, Materials and Methods

3.1. Experimental design, materials and methods

A dialogue can be defined as a spoken or written communication exchange in which multiple participants take turns contributing to the discussion on a particular subject matter or within a specific context.

To generate ChatSubs, we have addressed three main challenges, related to the delimitation of the dialogues and turns within each movie subtitles and the cleaning and processing of the dataset.

On the one hand, conversational analysis underlines a turn-taking principle that "there is one and only one person speaking while speaker change recurs with minimal gap and minimal overlap" [3]. Considering this premise, to identify the multiple dialogues within movie subtitles, we examined the time elapsed by setting a predefined time threshold, and establishing the start of a new dialogue when the time gap between consecutive dialogue turns is longer than this threshold. To find an appropriate threshold, a human annotator analyzed several threshold values[3] with two Spanish and one Basque movie. The threshold of 1 second provided the most coherent and semantically consistent separation of dialogues across all the movies.

This protocol is based on the conceptualization of subtitle sequences as a continuous flow of speaker turns, in which the time between turns of the same dialogue should be relatively short. In speech analysis the concept of inter-pausal units[4] is based on a similar premise, and previous research has shown that humans are typically very good at keeping the gaps between turns short, often with just a 200ms gap [7].

The use of a time threshold is even more feasible when processing subtitles, as in movies it is typical to take some time to properly introduce a new scene after transitioning from the previous one. This introduction is vital for helping the audience understand the context, location, and characters related to the new scene.

On the other hand, in order to separate or merge the subtitles into dialogue turns we exploited the layout of a standard subtitle file. The subtitles are parts of text that correspond to what a movie character says within the time interval while the subtitle is shown on screen. However, it is essential to note that a subtitle may not always match a dialogue turn since a dialogue turn marks the change of the speaker taking the lead in the conversation. For instance, several consecutive subtitles may capture a monologue that fits within a single dialogue turn.

Usually these cases are explicitly indicated in the subtitle format. When two consecutive subtitles form a single utterance from the same speaker, their corresponding end and beginning are marked with "...". Fig. 2 shows an example of subtitles that contain several turns. In this case, there are two turns:

- "Al recostarme hacia atrás y susurrar establecí una posición física dominante.". (Translation to English: After I leaned back and whispered I established a dominant physical position)
- "Qué bueno.". (Translation to English: "Excellent.").

Another situation is when a single subtitle contains two turns as parts of a dialogue. They are grammatically marked with "-" at the beginning. In Fig. 3 we see two such cases with 4 turns:

[3] [0.75, 1.0, 1.25] sec

[4] Stretches of audio from one speaker without any silence exceeding a certain amount (such as 200ms)

```
96
00:05:42,575 --> 00:05:45,660
Al recostarme hacia atrás y susurrar...

97
00:05:45,695 --> 00:05:48,536
...establecí una posición
física dominante.

98
00:05:49,716 --> 00:05:50,740
Qué bueno.
```

Fig. 2. Extract from the original subtitle of the movie with ID 1952430000 (see English translation of the dialogue above).

```
8
00:00:21,187 --> 00:00:22,950
- Es muy cierto.
- Nos vemos, Jim.

9
00:00:23,023 --> 00:00:24,547
- Kev, buen fin de semana.
- Gracias.
```

Fig. 3. Extract from the original Spanish subtitle of the movie with ID 1952430000 (see English translation of the dialogue above).

- "Es muy cierto". (Translation to English: It is very certain.)
- "Nos vemos, Jim". (Translation to English: See you, Jim.)
- "Kev, buen fin de semana". (Translation to English: Kev, have a good weekend.)
- "Gracias". (Translation to English: Thank you.)

Regarding the processing and cleaning of the raw data, we clean the subtitle sequence by removing the opening and closing credits. Sometimes, subtitles have HTML tags that we also eliminate. Once we have the complete utterances, we create a sequence of dialogue turns, with each turn marked by a specific timestamp at its beginning and end.

4. Limitations

Linguistic and cultural differences can influence the average time gap between dialogue turns. Additionally, different movie-making styles and approaches can also impact the time separation between dialogues. All these factors play a role in determining the appropriate separation threshold.

Also, our approach to segmenting turns within dialogues relies on the proper formatting of the subtitle file, which might not always be consistent.

Finally, since most of the subtitles are derived from translations of the original versions, it is possible that the dialects of the languages being targeted may not be accurately represented.

Ethics Statement

This dataset does not contain any copyrighted or illegal material, and the providers of the source data are in correspondence with the Digital Millennium Copyright Act[5] and general international copyright laws[6]. Commercial use is prohibited.

Data Availability

ChatSubs: A dataset of movie dialogues in Spanish, Catalan, Basque and Galician (Original data) (Zenodo)

CRediT Author Statement

Ksenia Kharitonova: Conceptualization, Methodology, Software, Writing – original draft, Visualization; **Zoraida Callejas:** Validation, Data curation, Writing – review & editing, Supervision, Project administration; **David Pérez-Fernández:** Conceptualization, Methodology, Software; **Asier Gutiérrez-Fandiño:** Data curation; **David Griol:** Validation, Data curation, Writing – review & editing, Supervision, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This dataset and publication is a result of the project CONVERSA (TED2021-132470B-I00) funded by MCIN/AEI/10.13039/501100011033 and by "European Union NextGenerationEU/PRTR".

References

[1] R. Csaky, G. Recski, The Gutenberg dialogue dataset, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, 2021, pp. 138–159. https://aclanthology.org/2021.eacl-main.11/.

[2] C. Danescu-Niculescu-Mizil, L. Lee, Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs, in: Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, 2011.

[3] P. Ten Have, Doing conversation analysis: A practical guide, 2nd ed., Sage Publications, 2007, doi:10.4135/9781849208895.

[4] M. Henderson, P. Budzianowski, I. Casanueva, S. Coope, D. Gerz, G. Kumar, N. Mrkšić, G. Spithourakis, P.-H. Su, I. Vulić, T.-H. Wen, A repository of conversational datasets, in: Proceedings of the First Workshop on NLP for Conversational AI, Association for Computational Linguistics, 2019, pp. 1–10. https://aclanthology.org/W19-4101/.

[5] Ksenia Kharitonova, Zoraida Callejas, David Pérez-Fernández, Asier Gutiérrez-Fandiño, David Griol, ChatSubs: a dataset of movie dialogues in Spanish, Catalan, Basque and Galician (1.1) [Data set], Zenodo (2023), doi:10.5281/zenodo.8220853.

[6] G. Rehm, A. Way (Eds.), European Language Equality: A Strategic Agenda for Digital Language Equality. Cognitive Technologies, Springer, 2023 https://link.springer.com/book/10.1007/978-3-031-28819-7.

[7] G. Skantze, Turn-taking in conversational systems and human-robot interaction: a review, Comput. Speech Lang. (2021) 67 https://www.sciencedirect.com/science/article/pii/S088523082030111X.

[5] https://www.copyright.gov/legislation/dmca.pdf (Last accessed: July 2023)

[6] https://www.opensubtitles.org/en/dmca (Last accessed: July 2023)