| **Author 1 First Middle Last Name**  (matriculation number)  **Documentation of NAME data set** |
| --- |
| Linguistic Data Science Lab  Ruhr University Bochum  Course/semester: Polysemy of connectors, winter semester 2022-23  Instructor: Dr. Claudia Roch, M.A.  Submission date: 30.01.2023 |
| [firstname.lastname@ruhr-uni-bochum.de](mailto:firstname.lastname@ruhr-uni-bochum.de) |

**Abstract**

This NEW template for documentations (3 pages, bibliography excluded) is supposed to provide the relevant sections to inform about in the data set documentation process which is an important part in data management. Good research practice includes dissemination of research results to ensure that data is handled and archived correctly. The reasons to document data

The reason for documenting the dataset is, because it provides metadata such as their origin, data size, column specifications, their importance and the reason for creating/using the dataset. This kind of documentation is necessary especially in the research field, when we are working as a group. So that, our fellow mates will be able to know the work of one another. Also, it can be used for future reference or to extend the dataset further.

The template follows recommendations for documenting data sets, in particular for NLP by Bender & Friedman (2018) and McMillan-Major et al. (2021).

.

# Dataset Description

## Dataset Summary

Name: Corpus of subordinated connectives in English Movie Dialogues.

[Link to repository](https://github.com/Linguistic-Data-Science-Lab), [Link to paper](https://aclanthology.org/), contact: NAME of a contact person that is responsible

In this documentation we describe the collection, processing and annotation of a corpus of subordinated connectives in English Movie Dialogues. It contains 400 sentences for the connectives ‘as’. ‘when’, ‘since’ and ‘after’ that were labeled with information whether they can express a temporal and/or a causal relation.

[list of links to relevant information about the dataset available elsewhere including the dataset paper, leaderboards, contact information of at least one person in case of further questions, a free text summary]

## Supported Task

[“What is this dataset used for?” the tasks supported, the original purpose for creating the data]

The main purpose of creating this dataset is to analyze whether the subordinate connectives namely ‘as’. ‘when’, ‘since’ and ‘after’ in the dialogue sentences are acting as temporal or causal or both. Temporal and causal relations between connectives have previously been studied by Bethard et al. (2008), O’Gorman et al. (2016), Dunietz et al. (2017) who also created annotated resources for connectives of a broader range. We decided not to use their annotation schemas or labels, because some of their sense distinctions were too fine-grained.

## Languages

The dataset is in English (BCP-47 code: en). There is no available information about which varieties of English are represented in the movies, but at least American English is supposed to be contained.

[A language tag from BCP-47[[1]](#footnote-0) identifying the language variety; prose description of the language variety]

# Dataset Structure

## Data Instances

The data set comprises 400 movie dialogue instances contained in csv files.

[size of the dataset, examples of data points]

## Data Fields

* A: line\_ID
  + It includes an unique value attached to each dialog spoken by individual characters. All the values starts with an identifier 'L' and ends with an unique numerical value.
* B: character\_ID
  + It also includes an unique value for each specific character. All the values starts with an identifier 'u' and ends with an unique numerical value.
* C: movie\_ID
  + It shows the unique values for the movie.
* D: movie\_titles
  + It shows the name/ title for the movie.
* E: genre
  + It shows the category/kind of the movie.
* F: year
  + It shows the year the movie was released.
* G: script\_urls
  + This indicates the urls of the specified movie from which the dialogue has been taken
* H: character\_name
  + It represents the character in the movie who delivered the dialogue.
* I: dialogue
  + This column contains the text of the dialogues spoken by the respective characters in the movie.
* J: connective
  + information about the form of one of the four English connectives that were annotated (‘*after*’, ‘*as*’, ‘*since*’, ‘*when*’)
* K: temporality
  + It distinguishes between the Temporal and Non- Temporal Clause.[ Signifies time in which the dialogue has been delivered ].
* L: causality

It distinguishes between the Causal and Non- Causal Clause.

[description of the features]

## Data Splits/ Data Statistics

[number of training and test instances, distribution, technical information a user might need to train a model]

In general, the train and test split ratio is 80:20 / 70:30. Again, it varies according to the tasks and individuals training the model. For example, in some cases the dataset will be splitted in the ratio of 60:40, in order to avoid overfitting. In special cases, we can also split it as train, validation and test. This time, the ratio can be 60:20:20 / 40:30:30 (depending on the task requirement).

# Dataset Creation

## Curation Rationale

[“Which texts were included and what were the goals in selecting texts, both in the original collection and in any further sub-selection?” (Bender & Friedman 2018:590)]

Dataset from Movie corpora is considered. Since, it is more expressive as it uses dialogue used in day-to-day life. Thus, this corpus would be useful to distinguish the clauses.

We chose the Movie Dialog Corpus as a dataset to investigate the temporal and causal senses of connectives in an informal conversational genre that is relatively similar to human communication – although the conversations are mainly fictional and follow the movie scripts. We expected to annotate sentences with a simpler structure than complex sentences found in newswire text. The text genres that previous approaches have looked at involve news (Dunietz et al. 2017, Cao et al.2022, Tan et al. 2022), congressional hearings (Dunietz et al.2017), fables and Reddit posts (Cao et al. 2022), but not movie dialogues to our knowledge.

We created a sub-selection of the original collection that comprised only sentences containing the connectives *after, as, since*, and *when*. While *after* and *since* code temporal anteriority or a sequence of events in time, *as* and *when* indicate simultaneity, so the connectives we study show different underlying temporal configurations.

## Source Data

The dataset was downloaded from the website "Kaggle".

Link to the source data: <https://www.kaggle.com/datasets/Cornell-University/movie-dialog-corpus>

The dataset was created by Cornell University and last updated in 2017. The corpus is based on the conversations of the movie scripts.

There were 220,579 conversations which were extracted from 10,292 pairs of characters from 617 movies including 9,035 characters. There were 304,713 utterances.

The dataset also consisted 2 types of metadata. The movie metadata were genres, release year, IMDB rating, number of IMDB votes, IMDB rating. Secondly, character metadata contained gender of 3,774 characters, and position on movie credits of 3,321 characters.

## Data collection/extraction

[tools, API, search terms]

The tab-separated source file ‘movie\_lines.tsv’ was loaded as a pandas DataFrame into Colab’s version of Jupyter notebooks. The initial dataframe contained the fields ‘lineID’, ‘characterID’, ‘movieID’, ‘characterName’, and ‘textUtterance’ for 293201 instances. Meta-data for the film title, genre and script urls were added from the files ‘movie\_titles\_metadata.tsv’ and ‘raw\_script\_urls.tsv’ to provide the annotators with background information and context. After that a sub-string search for the connectives with variants of capital and lower initial letter (e.g. ‘As’, ‘as’) was performed on the column containing the utterance and a new column for the corresponding identified connective added which resulted in 13696 instances. The dataframe was then exported as a csv-file (encoding=’utf8’, sep=’;’).

## Data processing

[details of any software or other tools used to process results]

The file was further processed in Microsoft Excel to prepare for the annotation. The curator filtered the spreadsheet manually for instances to be annotated and discard polysemous variants of the connectives.

## Source Language producers

[Age, Gender, (Race/ethnicity), Native language, Socioeconomic status, Number of different speakers represented, Presence of disordered speech]

## Annotations

The annotators (4 master’s students) annotated the data on the basis of “temporality” and “causality” of a sentence. Each annotator annotated 100 sentences. In this case the annotators considered the movie dialogue that has been delivered by each character in the movie that represents the data. The annotation guidelines specify that, the “temporality” indicates the events in which, one incident occurred after the other incident or it is on the other way around or the incidents occurred simultaneously. The “causality” refers to the fact that one incident took place because of the other incident or it can be interpreted as, one event happened as a result of another event. The annotators used 5 different labels namely TEMPORAL for representing temporality, NON-TEMPORAL for indicating non-temporality, CAUSAL for designating causality ,NON-CAUSAL for signifying non-causal events and an UNSURE for specifying the unsure condition of the annotator..

## Annotation process

The annotation was carried out in the time span December 2022 to January 2023. Annotators took breaks during their annotation phases.

Annotators reported difficulties to make decisions in the case of direct or indirect causality, while the annotation of temporal information was felt to be less difficult.

We calculated an inter-annotator agreement for 50 sentences between each annotator and curator. The kappa coefficient were in the range between 0.4 - 0.68. Although this result is between moderate and substantial, the annotators showed the same tendency to judge more instances as causal than the curator. .

## Annotator information

The annotators are students of linguistic data science, with an age range from 25 to years, . Their native languages are Bengali (1), Hindi (1), Malayalam (1), Tamil (1) and … , including women and men who gave their ethnicity as European (), … The annotators spent a semester working on the topic and annotation included a training phase for 20 sentences.

[Age, Gender, (Race/ethnicity), Native language, Socioeconomic status, Training in linguistics/other relevant discipline]

## Personal and Sensitive Information

[status of Personal Identifying (PI) data in the dataset; When depositing data involving human participants, authors must ensure that all datasets have been de-identified in accordance with the]

# Considerations for Using the Data

## Social Impact of the Dataset

## Discussion of Biases

[consider specifically social biases]

## Other Known Limitations

[common surface correlations]

# Additional Information

## Dataset Curators

[“Curators are involved in the selection of which data to include, by selecting individual documents, by creating search terms that generate sets of documents, by selecting speakers to interview and designing interview questions, and so forth.” (Bender & Friedman 2018:588)]

## Licensing Information

## ([FAIR Data Principles](https://www.nature.com/articles/sdata201618))

## Citation Information

## Contributions

# References/Bibliography

Bender, Emily M. & Friedman, Batya (2018). Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics, 6*, pp. 587–604. URL: <https://aclanthology.org/Q18-1041.pdf> [19.01.2023]

McMillan-Major, Angelina, Osei, Salomey, Rodriguez, Juan Diego, Ammanamanchi, Pawan Sasanka, Gehrmann, Sebastian, & Jernite, Yacine (2021). Reusable Templates and Guides For Documenting Datasets and Models for Natural Language Processing and Generation. *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics* (GEM 2021), pp. 121–135. URL: <https://aclanthology.org/2021.gem-1.11.pdf> [19.01.2023]

# Appendix

In the appendix you might include additional information that you want to ship with your work, but not have inside the text.

Please always attach a **signed declaration of independence** of your work:

I, \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_, confirm that the work for the term paper with the title: "\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_" was solely undertaken by myself and that no help was provided from other sources as those allowed. All sections of the paper that use quotes or describe an argument or concept developed by another author have been referenced, including all secondary literature used, to show that this material has been adopted to support my work.

Place / Date Signature

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Cf. <https://www.ietf.org/rfc/bcp/bcp47.txt>; <https://www.w3.org/International/questions/qa-choosing-language-tags> [↑](#footnote-ref-0)