Chinese Ministry of Foreign Affairs Press Conferences Corpus (CMFA PressCon)

Michal Mochtak and Richard Q. Turcsanyi

15 January 2023 (v3.0)

Contents

Overview

This is a codebook for an original corpus of the Chinese Ministry of Foreign Affairs Press Conferences (CMFA PresCon) mapping two decades of Chinese diplomatic discourse and priorities in China's foreign policy. The dataset is organized around a question/response structure extracted from the official transcripts of press conferences held between 15 October 2002 and 31 December 2022 and in its current version (v3) counts 28 447 data points.

Citation Information

id_orig

Original ID assigned to collected question-response dyad; ordered by date.

id_split

ID assigned to question-response dyad extracted from a multi-topical question. It contains the information on the original question-response ID (X_) and an index (_y) capturing the split.

id

ID assigned to question-answer dyad after the multi-topical split; ordered by date and multi-topical split (in a sequence).

link

Link originally used for accessing the transcript of a press conference (some of them do not work anymore).

spokesperson

Spokesperson holding the press conference.

day

Day of the press conference.

month

Month of the press conference.

year

Year of the press conference.

date

Full date of the press conference.

question

Question(s) asked by a member of the press corps.

answer

Answer(s) provided by the spokesperson.

question_lem

A lemmatized version of "question". Lemmatization was done using Trankit and XLM-Roberta-base model.

answer_lem

A lemmatized version of "answer". Lemmatization was done using Trankit and XLM-Roberta-base model.

who_asked

Author of the question (news outlet). Period covered: 24 March 2020 - 31 December 2022.

q_loc

Named entities for locations extracted from a transcribed question.

q_per

Named entities for persons extracted from a transcribed question.

q_org

Named entities for organizations extracted from a transcribed question.

q_misc

Named entities for miscellaneous references extracted from a transcribed question.

a_loc

Named entities for locations extracted from a transcribed answer.

a_per

Named entities for persons extracted from a transcribed answer.

a_org

Named entities for organizations extracted from a transcribed answer.

a_misc

Named entities for miscellaneous references extracted from a transcribed answer.