1 Data Format Task 1

The data is available in json format (see figure below). The unit of analysis is a paragraph. The json file includes a list of “Sentences” and a list of Annotations. For Task 1, Sentences correspond to paragraphs from the Bundestag parliamentary debates and each document includes all contributions of one speaker on one specific agenda item (see example below).


```
{
   "Sentences": [
      {
         "SentenceId": 18,
         "Tokens": [
            "Wir",
            "als",
            "AfD",
            "lehnen",
            "diese",
            "Totalüberwachung",
            "der",
            "Bürger",
            "ab",
            "."
         ]
      },
   ],
   "Annotations": [
      {
         "Message": [
            "18:4",
            "18:5",
            "18:6",
            "18:7"
         ],
         "Source": [
            "18:0",
            "18:1",
            "18:2"
         ],
         "Cue": [
            "18:3"
         ],
         "PTC": [
            "18:8"
         ]
      },
      { ... } ] }
```

- Each item in the list of Sentences is a dictionary with a SentenceId and a list of Tokens.
- Each item in the list of Annotations is a dictionary for a speech event (where *speech* also includes acts of *thought* and *writing*), encoding the cue word(s) and roles for this particular speech event.

For illustration, see the example above where we have a single cue word (Token 3 in Sentence 18, i.e., “ablehnen”). The verb particle of the cue word has been separated from the verb and is encoded as `PTC`. The cue word “ablehnen” is linked to a `Source` (“Wir”) and a `Message` (“diese Totalüberwachung der Bürger”). To increase readability, the examples only show the filled roles.

A cue can consist of one word (e.g., ”sagen“, ”schreiben“, ”denken“, ”Rede“) or can include a multi-word construction (e.g., ”Rede halten“, ”Informationen geben“, ”Gedanken machen“). We use the label `PTC` to encode verb prefixes that have been separated from the cue word (see example above). We use the same label to encode obligatory reflexive pronouns, such as ”sich“ in ”sich$_{\texttt{PTC}}$ Gedanken$_{\texttt{CUE}}$ machen$_{\texttt{CUE}}$“).

The example above includes only one cue (and therefore only one dictionary in the list of Annotations). The cue has the paragraph id "18" and the token id "3". To retrieve the word form for this cue, you can extract the token with id 3 (i.e., the fourth token in the `Tokens` list) from the 18th paragraph in the list of `Sentences`, which is ”lehnen“. This is a particle verb and the verb prefix is encoded as `PTC` (paragraph 18, token id 8 $\rightarrow$ ”ab“). In addition to the cue word and its verb particle, the `Annotations` dictionary also includes the roles for this cue (the complete list of roles is: Source, Message, Addresse, Topic, Medium and Evidence). For more information, please refer to the annotation guidelines (`https://github.com/umanlp/SpkAtt-2023`).

In some speeches, the boundaries have not been identified correctly so that some sentences have been split up, with the first part of the sentence included in one paragraph and the second part in another paragraph. The example below illustrates such a case. Here, the source for the text fragment 10 is included in the previous sentence fragment with id 9, as indicated by the Source id (”9:0“).

The same mechanism is used to encode roles that span over multiple sentences in cases where not all sentences are included in the same paragraph.

```
{ "Sentences": [ {...},
     {
       "SentenceId": 9,
       "Tokens": [
          "Sie",              # Source 9:0
          ",",
          "verehrte",
          "Kollegen",
          "der",
          "FDP",
          ",",
          "fordern",          # Cue 9:7
          "im",
          "Übrigen",
          "nach",
          "wie",
          "vor",
          "eine",             # Message 9:13
          "Abschaffung",      # Message 9:14
          "des",              # Message 9:15
          "NetzDG",           # Message 9:16
          ","
       ]
    },
    {
       "SentenceId": 10,
       "Tokens": [
          "haben",
          "aber",
          "die",
          "letzten",
          "zwei",
          "Jahre",
          "tatenlos",
          "verstreichen",
          "lassen",
          ",",
          "diesen",           # Message 10:10
          "völlig",           # Message 10:11
          "überkommenen",     # Message 10:12
          "Ansatz",           # Message 10:13
          "Ihrer",            # Message 10:14
          "Politik",          # Message 10:15
          "zu",
          "überdenken",       # Cue 10:17
          "."
       ] } ],
```

```
48
49    "Annotations": [
50       { ... },
51       {
52          "Message": [
53             "9:13",
54             "9:14",
55             "9:15",
56             "9:16"
57          ],
58          "Source": [
59             "9:0"
60          ],
61          "Cue": [
62             "9:7"
63          ]
64       },
65       {
66          "Message": [
67             "10:10",
68             "10:11",
69             "10:12",
70             "10:13",
71             "10:14",
72             "10:15"
73          ],
74          "Source": [
75             "9:0"              # Source for Cue 10:17 is included
76          ],                    # in the previous paragraph
77          "Cue": [              # with SentenceId 9
78             "10:17"
79          ]
80       },
81       { ... } ] }
```

The next example illustrated the annotation of multiple cues in the same sentence, encoded as a list of dictionaries in ”Annotations“.

```
    {
       "SentenceId": 6,
       "Tokens": [
          "Im",
          "Übrigen",
          "befinde",              # 1. Cue (6:2)
          "ich",
          "mich",
          "damit",
          "offensichtlich",
          "auch",
          "in",
          "weitgehender",
          "Übereinstimmung",      # 1. Cue (6:10)
          "mit",
          "der",
          "Kanzlerin",
          ",",
          "die",
          "auf",
          "eine",
          "entsprechende",
          "Frage",                # 2. Cue (6:19)
          "vor",
          "wenigen",
          "Minuten",
          "hier",
          "in",
          "diesem",
          "Hause",
          "genau",
          "so",
          "geantwortet",          # 3. Cue (6:29)
          "hat",
          "."
       ]
    },
    {
      ...
    }
  ],
                     # see next page for Annotations
```

```
  "Annotations": [
     {
       ...
     },
     {      {
        "Source": [
           "6:3"
        ],
        "Cue": [          # 1. Cue
           "6:2",
           "6:10"
        ],
        "PTC": [
           "6:4"
        ]
     },
     {
        "Cue": [           # 2. Cue
           "6:19"
        ],
     },
     {
        "Message": [
           "6:27",
           "6:28"
        ],
        "Source": [
           "6:15"
        ],
        "Topic": [
           "6:16",
           "6:17",
           "6:18",
           "6:19"
        ],
        "Cue": [           # 3. Cue
           "6:29"
        ],
     }
```

The first cue, *sich in Übereinstimmung befinden*, includes two cue words, ”befinde“ and ”Übereinstimmung“ and a reflexive pronoun (”mich“), encoded as PTC. This cue also has a Source (”ich“). The second cue, *Frage*, does not have any roles. The third cue, *geantwortet*, has a Source (”die“), a Message (”genau so“) and a Topic (”auf eine entsprechende Frage“).

For more information on the annotations, see the Annotation Guidelines, available from our github repository.