

1 Data Format Task 1

The data is available in json format (see figure below). The unit of analysis is a sentence (or whatever spaCy's sentence splitter considers as one). The json file includes a list of Sentences and a list of Annotations. For Task 1, all Sentences come from parliamentary debates from the German Bundestag and each document includes all contributions of one speaker on one specific agenda item (see example below).

```
1 {
2   "Sentences": [
3     {
4       "SentenceId": 18,
5       "Tokens": [
6         "Wir",
7         "als",
8         "AfD",
9         "lehnen",
10        "diese",
11        "Totalüberwachung",
12        "der",
13        "Bürger",
14        "ab",
15        "."
16      ]
17    },
18  ],
19  "Annotations": [
20    {
21      "Message": [
22        "18:4",
23        "18:5",
24        "18:6",
25        "18:7"
26      ],
27      "Source": [
28        "18:0",
29        "18:1",
30        "18:2"
31      ],
32      "Cue": [
33        "18:3"
34      ],
35      "PTC": [
36        "18:8"
37      ]
38    },
```

- 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) (or triggers) 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” (en.: *reject*)). 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 (or trigger) can consist of one word (e.g., “sagen” *say*, “schreiben” *write*, “denken” *think*, “Rede” *speech*) or can include a multi-word construction (e.g., “Rede halten” *give a speech*, “Informationen geben” *give information*, “Gedanken machen” *wonder*). 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_{PTC} Gedanken_{CUE} machen_{CUE}”).

The example above includes only one cue (and therefore only one dictionary in the list of Annotations). The cue has the sentence 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 sentence in the list of Sentences, which is “lehnen”. This is a particle verb and the verb prefix is encoded as PTC (sentence 18, token id 8 → “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, Adresse, Topic, Medium and Evidence). For more information, please refer to the annotation guidelines (<https://github.com/umanlp/SpkAtt-2023>).

In some speeches, the sentence boundaries have not been identified correctly so that some sentences have been split up into fragments. 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.

```

1  { "Sentences": [ {...},
2    {
3      "SentenceId": 9,
4      "Tokens": [
5        "Sie",           # Source 9:0
6        ",",
7        "verehrte",
8        "Kollegen",
9        "der",
10       "FDP",
11       ",",
12       "fordern",       # Cue 9:7
13       "im",
14       "Übrigen",
15       "nach",
16       "wie",
17       "vor",
18       "eine",           # Message 9:13
19       "Abschaffung",   # Message 9:14
20       "des",            # Message 9:15
21       "NetzDG",        # Message 9:16
22       ",",
23     ]
24   },
25   {
26     "SentenceId": 10,
27     "Tokens": [
28       "haben",
29       "aber",
30       "die",
31       "letzten",
32       "zwei",
33       "Jahre",
34       "tatenlos",
35       "verstreichen",
36       "lassen",
37       ",",
38       "diesen",         # Message 10:10
39       "völlig",         # Message 10:11
40       "überkommenen",  # Message 10:12
41       "Ansatz",         # Message 10:13
42       "Ihrer",          # Message 10:14
43       "Politik",        # Message 10:15
44       "zu",
45       "überdenken",     # Cue 10:17
46       ".",
47     ] } ],

```

```

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"
76     ],
77     "Cue": [
78       "10:17"
79     ]
80   },
81   { ... } ] }

```

Source for Cue 10:17 is included
 # in the previous sentence
 # with SentenceId 9

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

```

1
2      {
3        "SentenceId": 6,
4        "Tokens": [
5          "Im",
6          "Übrigen",
7          "befinde",          # 1. Cue (6:2)
8          "ich",
9          "mich",
10         "damit",
11         "offensichtlich",
12         "auch",
13         "in",
14         "weitgehender",
15         "Übereinstimmung",  # 1. Cue (6:10)
16         "mit",
17         "der",
18         "Kanzlerin",
19         ",",
20         "die",
21         "auf",
22         "eine",
23         "entsprechende",
24         "Frage",            # 2. Cue (6:19)
25         "vor",
26         "wenigen",
27         "Minuten",
28         "hier",
29         "in",
30         "diesem",
31         "Hause",
32         "genau",
33         "so",
34         "geantwortet",      # 3. Cue (6:29)
35         "hat",
36         "."
37       ]
38     },
39     {
40       ...
41     }
42 ],
43
44     # see next page for Annotations
45

```

```

46 "Annotations": [
47   {
48     ...
49   },
50   {
51     "Source": [
52       "6:3"
53     ],
54     "Cue": [           # 1. Cue
55       "6:2",
56       "6:10"
57     ],
58     "PTC": [
59       "6:4"
60     ]
61   },
62   {
63     "Cue": [           # 2. Cue
64       "6:19"
65     ],
66   },
67   {
68     "Message": [
69       "6:27",
70       "6:28"
71     ],
72     "Source": [
73       "6:15"
74     ],
75     "Topic": [
76       "6:16",
77       "6:17",
78       "6:18",
79       "6:19"
80     ],
81     "Cue": [           # 3. Cue
82       "6:29"
83     ],
84   }

```

The first cue, "sich in Übereinstimmung befinden" (*being in accordance*), includes two cue words, "befinde" (*being in*) and "Übereinstimmung" (*accordance*) and a reflexive pronoun ("mich"), encoded as PTC. This cue also has a Source ("ich", *I*). The second cue, "Frage" (*question*), does not have any roles. The third cue, "geantwortet" (*answered*), has a Source ("die", *who*), a Message ("genau so", *exactly like that*) and a Topic ("auf eine entsprechende Frage", *to a corresponding question*).

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