1 Data Format Task 1

As a variation on the data format described in Dataformat_Task1_a.pdf, we also provide our data in an alternative format which merges the Tokens and Annotations into one data structure, thus making it easier to inspect the data. Here, the json dictionary keys are the paragraph ids. For each paragraph, we include the following information, each as a list:

- **Tokens**: the list of tokens for this paragraph
- **Annotations**: a list of dictionaries, where each item in the list is a dictionary with a cue and the respective roles for this cue.

```
{
   "18": {                        # Sentence id
      "Tokens": [                 # Tokens for sentence 18
         "Wir",
         "als",
         "AfD",
         "lehnen",
         "diese",
         "Totalüberwachung",
         "der",
         "Bürger",
         "ab",
         "."
      ],
      "Annotations": [            # Annotations for sentence 18
                                  # (list of dictionaries, with one
                                  # dictionary per cue word(s))
         {
            "Message": [          # Message
               "18:4",            # sentence_id:token_id
               "18:5",
               "18:6",
               "18:7"
            ],
            "Cue": ["18:3"],      # Cue word(s)
            "PTC": [              # Particles (separable verb prefixes
               "18:8"             # and reflexive pronouns)
            ],
            "Source": [           # Source
               "18:0",
               "18:1",
               "18:2"
            ],
         } ] },  ... }
```

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 Gedanken machen“).

In the example above, we only have one cue (and therefore only one dictionary in the Annotations list). 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 list) from paragraph 18, 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 include the roles for this cue (the complete list of roles is: Source, Message, Addresse, Topic, Medium and Evidence. For more information, see the annotation guidelines).

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 the next 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 ids (”9:0“, ”9:1“, ...).

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

```
{
  {
    "9": {
      "Tokens": [
         "Sie",
         ",",
         "verehrte",
         "Kollegen",
         "der",
         "FDP",
         ",",
         "fordern",
         "im",
         "Übrigen",
         "nach",
         "wie",
         "vor",
         "eine",
         "Abschaffung",
         "des",
         "NetzDG",
         ","
      ],
      "Annotations": [
         {
            "Message": [
               "9:13",
               "9:14",
               "9:15",
               "9:16"
            ],
            "Source": [
               "9:0"
            ],
            "Cue": [
               "9:7"
            ]
         }
      ]
   },
```


                        # to be continued on next page

```
  "10": {
     "Tokens": [
        "haben",
        "aber",
        "die",
        "letzten",
        "zwei",
        "Jahre",
        "tatenlos",
        "verstreichen",
        "lassen",
        ",",
        "diesen",
        "völlig",
        "überkommenen",
        "Ansatz",
        "Ihrer",
        "Politik",
        "zu",
        "diskutieren",
        "."
     ],
     "Annotations": [
        {
           "Message": [
              "10:10",
              "10:11",
              "10:12",
              "10:13",
              "10:14",
              "10:15"
           ],
           "Source": [
              "9:0"
           ],
           "Cue": [
              "10:17"
           ]
        }
     ]
  },
  ...
}
```

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

```
{
   "7": {
      "Tokens": [
         "Im",
         "Übrigen",
         "befinde",              # cue 1 (7:2)
         "ich",
         "mich",
         "damit",
         "offensichtlich",
         "auch",
         "in",
         "weitgehender",
         "Übereinstimmung",      # cue 1 (7:10)
         "mit",
         "der",
         "Kanzlerin",
         ",",
         "die",
         "auf",
         "eine",
         "entsprechende",
         "Frage",                # cue 2 (7:19)
         "vor",
         "wenigen",
         "Minuten",
         "hier",
         "in",
         "diesem",
         "Hause",
         "genau",
         "so",
         "geantwortet",          # cue 3 (7:29)
         "hat",
         "."
      ],


                                 # to be continued on next page
```

```
    "Annotations": [
       {
          "PTC": [
             "7:4"
          ],
          "Source": [
             "7:3"
          ],
          "Cue": [
             "7:2",
             "7:10"
          ]
       },
       {
          "Cue": [
             "7:19"
          ]
       },
       {
          "Message": [
             "7:27",
             "7:28"
          ],
          "Source": [
             "7:15"
          ],
          "Topic": [
             "7:16",
             "7:17",
             "7:18",
             "7:19"
          ],
          "Cue": [
             "7:29"
          ]
       }
    ]
}, ... }
```


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.