Annotation Scheme

The references in the dataset point either to case law (blue rectangles in the upper part of the diagram below) or to literature (larger green rectangles in the lower part). Moreover, references are either explicit (left side of the diagram) or implicit (right side of the diagram). Every annotated reference consists of basic units and, in case of the implicit references, of a link to the related explicit reference.

First type of basic units is so called constituents. These categorize individual annotated spans of text. Constituents are either related to references to case law (blue rectangles in the middle of the diagram below), literature (green rectangles) or to both types of references (red rectangles). Second type of basic units is so called attributes (rectangles with interrupted red frame). These are related to either a constituent (content) or the annotated reference as a whole and present a specific value that provides additional information about the reference or its content (polarity, depth of discussion and type of the content).

Arrows in the diagram indicate how each annotated reference is put together from different basic units.

Examples:

1) The c:id constituent, which identifies the referred case, is a basic unit representing the corresponding annotated text span containing this information.

2) The constituent content is an annotated text span indicating why the specific reference to case law or literature is being made. This constituent is directly tied to the attribute content type. A single annotated reference may contain multiple content constituents with different content type attributes.

3) Every explicit case law reference c:ref(expl) requires to be assigned several constituents and attributes in order to be constructed and identified properly. As indicated in the diagram, it requires one or multiple c:id constituents, one c:court constituent, one c:date constituent, one c:type constituent, one POI constituent, one or multiple content constituents (with a set value of the attribute content type), and also set values of the attributes polarity and depth of discussion.

4) Similarly, every implicit literature reference l:ref(impl) requires to be assigned a different set of constituents and attributes in order to be constructed and identified properly. As shown in the diagram, it requires one POI constituent, one or multiple content constituents (with a set value of the attribute content type), one implicit constituent and also set values of the attributes polarity and depth of discussion. Most importantly, as it is an implicit reference, it requires a set link to one explicit literature reference l:ref(expl), which allows to understand which previous reference is implied in the implicit reference.
Two annotators have annotated every document in the dataset. These annotators were not familiar with each other’s work. Every document annotated by two annotators has been passed to editor/curator, whose task was to compare the annotated documents, and mark every annotated reference, constituent and attribute as either correct or incorrect. If editor/curator marked work of both annotators as incorrect, he was then required to re-annotate the reference, and therefore to create correct editorial annotations. Raw corpus contains all annotations (correct, incorrect and created by editor). Gold corpus contains only correct annotations (either marked as correct or created by editor).

Constituents and References (EN to CZ)

Diagram on page 2 is in English, however name convention used in corpus is in Czech. To account for this, we provide comparison of English (used throughout this ReadMe) and Czech (used in corpus) designations.

<table>
<thead>
<tr>
<th>Constituents</th>
<th>Czech</th>
</tr>
</thead>
<tbody>
<tr>
<td>c:id</td>
<td>J:Identifikátor</td>
</tr>
<tr>
<td>c:court</td>
<td>J:Soud</td>
</tr>
<tr>
<td>c:date</td>
<td>J:Datum</td>
</tr>
<tr>
<td>c:type</td>
<td>J:Druh</td>
</tr>
<tr>
<td>l:author</td>
<td>L:Autor</td>
</tr>
<tr>
<td>l:title</td>
<td>L:Název</td>
</tr>
<tr>
<td>l:other</td>
<td>L:Další údaje</td>
</tr>
<tr>
<td>POI</td>
<td>Element</td>
</tr>
<tr>
<td>content</td>
<td>Argument</td>
</tr>
<tr>
<td>content type</td>
<td>Druh argumentu</td>
</tr>
<tr>
<td>implicit</td>
<td>Implicitníidentifikace</td>
</tr>
<tr>
<td>polarity</td>
<td>Sentiment</td>
</tr>
<tr>
<td>depth of discussion</td>
<td>Koeficient</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>References</th>
<th>Czech</th>
</tr>
</thead>
<tbody>
<tr>
<td>c:ref(expl)</td>
<td>Reference-Judikatura</td>
</tr>
<tr>
<td>l:ref(expl)</td>
<td>Reference-Literatura</td>
</tr>
<tr>
<td>c:ref(impl)</td>
<td>Reference-Judikatura-implicitní</td>
</tr>
<tr>
<td>l:ref(impl)</td>
<td>Reference-Judikatura-explicitní</td>
</tr>
</tbody>
</table>
Corpus File Structure

```
[
  (
    "txt": "text of a court decision",
    "constituents": [
      {
        "id": "unique id that can be used to tie the constituent to a reference",
        "start": "starting offset in the text",
        "end": "ending offset in the text",
        "type": "type of the constituent",
        "isCorrect": "true if marked correct by an editor, false otherwise",
        "isEditorial": "true if created by an editor, false otherwise",
        "author": "author’s id"
      }
    ]
  ]
```

Examples (in Python)

1. Load Documents.

   ```python
   with open('corpus.json', 'r') as f:
       docs = json.load(f)
   ```

2. Print all c:id constituents from the first document.

   ```python
   doc = docs[0]
   txt = doc['txt']
   c_id_constituents = [c for c in doc['constituents']
       if c['type'] == 'Identifikator']
   for c_id in c_id_constituents:
       start = c_id['start']
       end = c_id['end']
       print(txt[start:end])
   ```

3. Count all c:id constituents across the corpus.

   ```python
   count_c_id = 0
   for doc in docs:
       c_id_constituents = [c for c in doc['constituents']
           if c['type'] == 'Identifikator']
       count_c_id += len(c_id_constituents)
   ```
Suggested Citation


Acknowledgment

Supported by the Czech Science Foundation under grant no. GA17-20645S.