Expanding the content model of annotationBlock

Alexandre Bartz, Juliette Janes, Laurent Romary, Philippe Gambette, Rachel Bawden, Pedro Ortiz Suarez, Benoît Sagot, Simon Gabay

Expanding the content model of `annotationBlock`

Alexandre Bartz[1], Juliette Janes[2], Laurent Romary[3], Philippe Gambette[4], Rachel Bawden[3], Pedro Ortiz Suarez[1,3], Benoît Sagot[3], and Simon Gabay[5]

[1]Sorbonne Université (France)
[2]PSL - Ecole nationale des chartes (France)
[3]Inria, Paris (France)
[4]LIGM, Univ. Eiffel, CNRS, Marne-la-Vallée (France)
[5]Université de Genève (Switzerland)

Linguistic annotation benefits from ISO specifications such as the Morphosyntactic Annotation Framework (MAF), whose recommendations have been added to the TEI P5[1] (ISO-24611 2012; Stührenberg 2012). Relying on feature structures (cf. ex. 1), these recommendations have however not been fully integrated into the TEI stand-off annotation model (Bański et al. 2016) and, for instance, it is currently impossible to encode feature structures within the `listAnnotation` and `annotationBlock` elements.

```
<seg>
    <w xml:id="s1w1">lõg</w>
    <w xml:id="s1w2">tems</w>
</seg>
<spanGrp type="wordForm">
   <span target="#s1w1" ana="#s1fs1"/>
   <span target="#s1w2" ana="#s1fs2"/>
</spanGrp>
<fs xml:id="s1fs1">
   <f name="lemma">
      <string>long</string>
   </f>
</fs>
<fs xml:id="s1fs2">
   <f name="lemma">
      <string>temps</string>
   </f>
</fs>
```

Example 1: TEI encoding following the MAF

[1]https://tei-c.org/release//doc/tei-p5-doc/en/html/FS.html.

With the multiplication of annotation tools, the case is becoming more complex and is no longer limited to low-level linguistic annotation (lemma, POS, etc.) using the `fs` element. For instance, it is becoming more and more common for medievalists (Stutzmann 2011) and modernists (Gabay and Barrault 2020) working on normalisation tasks to encode various levels of transcription (*lõg tems* → *long tems*) or offer a version which is fully aligned with contemporary spellings (*lõg tems* → *longtemps*). New elements, such as `reg`, would therefore be extremely useful (cf. Ex. 2).

```
<seg corresp="#s1">
    <w xml:id="s1w1">lõg</w>
    <w xml:id="s1w2">tems</w>
</seg>
[…]
<spanGrp type="wordForm">
   <span target="#s1w1" ana="#s1reg1"/>
   <span target="#s1w2" ana="#s1reg2"/>
</spanGrp>
<reg type="reg" xml:id="s1reg1">long</reg>
<reg type="reg" xml:id="s1reg2">tems</reg>
[…]
<spanGrp type="formNorm">
   <span target="#s1w1 #s1w2" ana="#s1norm1"/>
</spanGrp>
<reg type="norm" xml:id="s1norm1">longtemps</reg>
```

Example 2: Annotation of linguistic normalisation

Sadly, in the case of multiple levels of embedded stand-off data, the current version of the guidelines promotes a multiplication of `standOff` elements with a semantically inappropriate `seg` to store the necessary annotation (cf. Ex. 3).

```
<standOff type="linguistic">
   <seg corresp="#s1">
      <spanGrp>[…]</spanGrp>
      <fs>[…]</fs>
   </seg>
</standOff>
<standOff type="norm">
   <seg corresp="#s1">
      <spanGrp>[…]</spanGrp>
      <reg>[…]</reg>
   </seg>
</standOff>
```

Example 3: Annotation of linguistic normalisation

Using the data of the *E-ditiones* project, we will make a case for a more appropriate use of `standOff` with one `listAnnotation` per annotation type and an extended version of the `annotationBlock` content model (cf. Ex. 4). This model, also an ISO recommendation (ISO-24624 2016), has originally been created to identify the reference features needed to transcribe spoken resources that are anchored on a single reference timeline, but also for integrating mechanisms to encompass most usual transcription conventions. It therefore needs to include not only the `fs` element for reference annotation, but also additional information about editorial transcription such as normalisation (`reg`) or any other philological intervention on the text (cf. the `model.pPart.transcriptional` class: `add`, `corr`, `damage`, `del`, `handShift`, `mod`, `orig`, `redo`, `reg`, `restore`, `retrace`, `secl`, `sic`, `supplied`, `surplus`, `unclear`, `undo`).

Data

Scripts and data are available at https://github.com/e-ditiones/Annotator.

```
<TEI>
   <text>
      <body>
         <p>
            <seg xml:id="s1">
               <w xml:id="s1w1">lõg</w>
               <w xml:id="s1w2">tems</w>
               <w xml:id="s1w3">a</w>
               <w xml:id="s1w4">geneve</w>
            </seg>
         </p>
      </body>
   </text>
   <standOff>
      <listAnnotation type="linguistic">
         <annotationBlock corresp="#s1">
            <spanGrp type="wordForm">
               <span target="#s1w1" ana="#s1ling1"/>
               <span target="#s1w2" ana="#s1ling2"/>
               <span target="#s1w3" ana="#s1ling3"/>
               <span target="#s1w4" ana="#s1ling4"/>
            </spanGrp>
            <fs xml:id="s1ling1">
               <f name="lemma">
                  <string>long</string>
               </f>
               <f name="pos">
                  <symbol value="ADJqua"/>
               </f>
               <f name="nomb">
                  <symbol value="s"/>
               </f>
               <f name="genre">
                  <symbol value="m"/>
               </f>
               <f name="norm1">
                  <string>long</string>
               </f>
            </fs>
            <fs xml:id="s1ling2">
               <f name="lemma">
                  <string>temps</string>
               </f>
               <f name="pos">
                  <symbol value="NOMcom"/>
               </f>
               <f name="nomb">
                  <symbol value="s"/>
               </f>
               <f name="genre">
                  <symbol value="m"/>
               </f>
```

```
            <f name="norm1">
               <string>temps</string>
            </f>
         </fs>
         <fs xml:id="s1ling3">
            <f name="lemma">
               <string>à</string>
            </f>
            <f name="pos">
               <symbol value="PRE"/>
            </f>
            <f name="norm1">
               <string>à</string>
            </f>
         </fs>
         <fs xml:id="s1ling4">
            <f name="lemma">
               <string>Genève</string>
            </f>
            <f name="pos">
               <symbol value="NOMpro"/>
            </f>
            <f name="norm1">
               <string>Genève</string>
            </f>
         </fs>
      </annotationBlock>
   </listAnnotation>
   <listAnnotation type="normalisation">
      <annotationBlock corresp="#s1">
         <spanGrp type="formNorm">
            <span target="#s1w1 #s1w2" corresp="#s1norm1"/>
            <span target="#s1w3" corresp="#s1norm2"/>
            <span target="#s1w4" corresp="#s1norm3"/>
         </spanGrp>
         <reg xml:id="s1norm1">longtemps</reg>
         <reg xml:id="s1norm2">à</reg>
         <reg xml:id="s1norm3">Genève</reg>
      </annotationBlock>
   </listAnnotation>
   <listAnnotation type="NERD">
      <annotationBlock corresp="#s1">
         <spanGrp type="formNorm">
            <span target="#s1w4" ana="#s1ner1"/>
         </spanGrp>
         <fs xml:id="s1ner1">
            <f name="NER">
               <symbol value="B-loc.adm.town"/>
            </f>
            <f name="wikidata">
               <symbol value="Q71"/>
            </f>
         </fs>
      </annotationBlock>
   </listAnnotation>
</standOff>
</TEI>
```

Example 4: TEI encoding following the MAF

References

Bański, P. et al. (2016). "Wake up, standOff!" In: *TEI Abstracts 2016*. TEI Conference 2016. Vienna, Austria, pp. 35–37. URL: https://tei-c.org/Vault/MembersMeetings/2016/sites/default/files/TEIconf2016_BookOfAbstracts.pdf (visited on 07/22/2021).

Gabay, S. and L. Barrault (June 2020). "Traduction automatique pour la normalisation du français du XVII e siècle". In: *27ème Conférence sur le Traitement Automatique des Langues Naturelles*. TALN2020. Nancy, France: ATALA. URL: https://hal.archives-ouvertes.fr/hal-02596669 (visited on 12/10/2020).

ISO-24611 (2012). *Language Resource management — Transcription of Spoken Language — ISO 24611*. ISO.

ISO-24624 (2016). *Language resource management — Morpho-syntactic annotation framework (MAF) — ISO 24624*. ISO.

Stührenberg, M. (Nov. 5, 2012). "The TEI and Current Standards for Structuring Linguistic Data". In: *Journal of the Text Encoding Initiative* (Issue 3). ISSN: 2162-5603. DOI: 10.4000/jtei.523. URL: https://journals.openedition.org/jtei/523 (visited on 07/22/2021).

Stutzmann, D. (2011). "Paléographie statistique pour décrire, identifier, dater... Normaliser pour coopérer et aller plus loin ?" In: *Kodikologie und Paläographie im digitalen Zeitalter = Codicology and Palaeography in the Digital Age*. Schriften des Instituts für Dokumentologie und Editorik 2. Ed. by **F. Fischer, C. Fritze, and G. Vogeler**, pp. 247–277. URL: https://halshs.archives-ouvertes.fr/halshs-00596970 (visited on 01/09/2020).