Linguistic Annotation in the TEI

LRBS 2025

Digital Editing

OXFORD TEXT ARCHIVE

CLARIN-UK

What is linguistic annotation?

- According to TEI guidelines 16.4: Annotation determined by an analysis of the linguistic features of the text.
- Syntactic analysis – analysis of sentence and clause structure.
- Lexical analysis – word-level analysis of syntax and semantics; lemmatization and normalization are applied at this level.
  - Annotation at word-level is often referred to as ‘tagging’.
- Morphological analysis – identification of morphemes within a word.

Why would you use linguistic annotation?

- Used largely for linguistic corpora.
  - Corpus linguistics – the study of language in a written context.
  - Implementation dependent on corpus design.
- Useful for enhancing search functionality in digital editions and increasing the usability of your text.
  - Search by lemma.
  - Search by POS.
  - Search by semantic field.
- Downsides of linguistic annotation
  - Reduced readability.
  - Not a ‘plain text’.

What does it look like?

Basic example

```
<head>THE SCOTS APOSTACY.</head>
<l>IS't come to this? what? ſhall the Cheekes of Fame</l>
<l> Stretch't with the breath of learned <hi>Lowdens</hi> name </l>
<l>Be flagg'd againe, and that great peice of Sence</l>
<l>As rich in Loyaltie, as Eloquence,</l>
<l>Brought to the Teſt, be found a tricke of State?</l>
```

Syntactic analysis

See TEI Guidelines 18.1

```
<head>THE SCOTS APOSTACY.</head>
<l n="1"/>
<s><cl>IS't come to this? </cl></s>
<s><cl>what? </cl></s>
<s><cl>ſhall the Cheekes of Fame <l n="2"/>
        <cl>Stretch't with the breath of learned <hi>Lowdens</hi> name</cl>
        <l n="3"/>Be flagg'd againe,</cl>
    <cl><cl>and that great peice of Sence <l n="4"/><cl>As rich in Loyaltie, as
                Eloquence,</cl>
            <l n="5"/>Brought to the Teſt</cl>, be found a tricke of State?</cl></s>
```

What does it look like?

Lexical analysis

See TEI Guidelines 16.4

```
<head>
    <w pos="AT" sem="Z5">THE</w>
    <w pos="JJ" sem="Z2">SCOTS</w>
    <w norm="Apostasy" pos="NN1" sem="A1.7-/S9">APOSTACY</w>
    <w pos="." sem="PUNC">.</w>
</head>
<l>
    <w norm="Is it" pos="VBZ/PPH1" sem="A3+/Z8">IS't</w>
    <w pos="VV0" sem="A4.1[i1.2.1">come</w>
    <w pos="II" sem="A4.1[i1.2.2">to</w>
    <w pos="DD1" sem="Z8">this</w>
    <w pos="?" sem="PUNC">?</w>
    <w pos="DDQ" sem="Z8">what</w>
    <w pos="?" sem="PUNC">?</w>
    <w norm="Shall" pos="VM" sem="T1.1.3">ſhall</w>
    <w pos="AT" sem="Z5">the</w>
    <w norm="Cheeks" pos="NN2" sem="B1">Cheekes</w>
    <w pos="IO" sem="Z5">of</w>
    <w pos="NN1" sem="X2.2+">Fame</w>
</l>
```

Basic example

```
<head>THE SCOTS APOSTACY.</head>
<l>IS't come to this? what? ſhall the Cheekes of Fame</l>
<l> Stretch't with the breath of learned <hi>Lowdens</hi> name </l>
<l>Be flagg'd againe, and that great peice of Sence</l>
<l>As rich in Loyaltie, as Eloquence,</l>
<l>Brought to the Teſt, be found a tricke of State?</l>
```

What does it look like?

Morphological Analysis

See TEI Guidelines 18.1.2

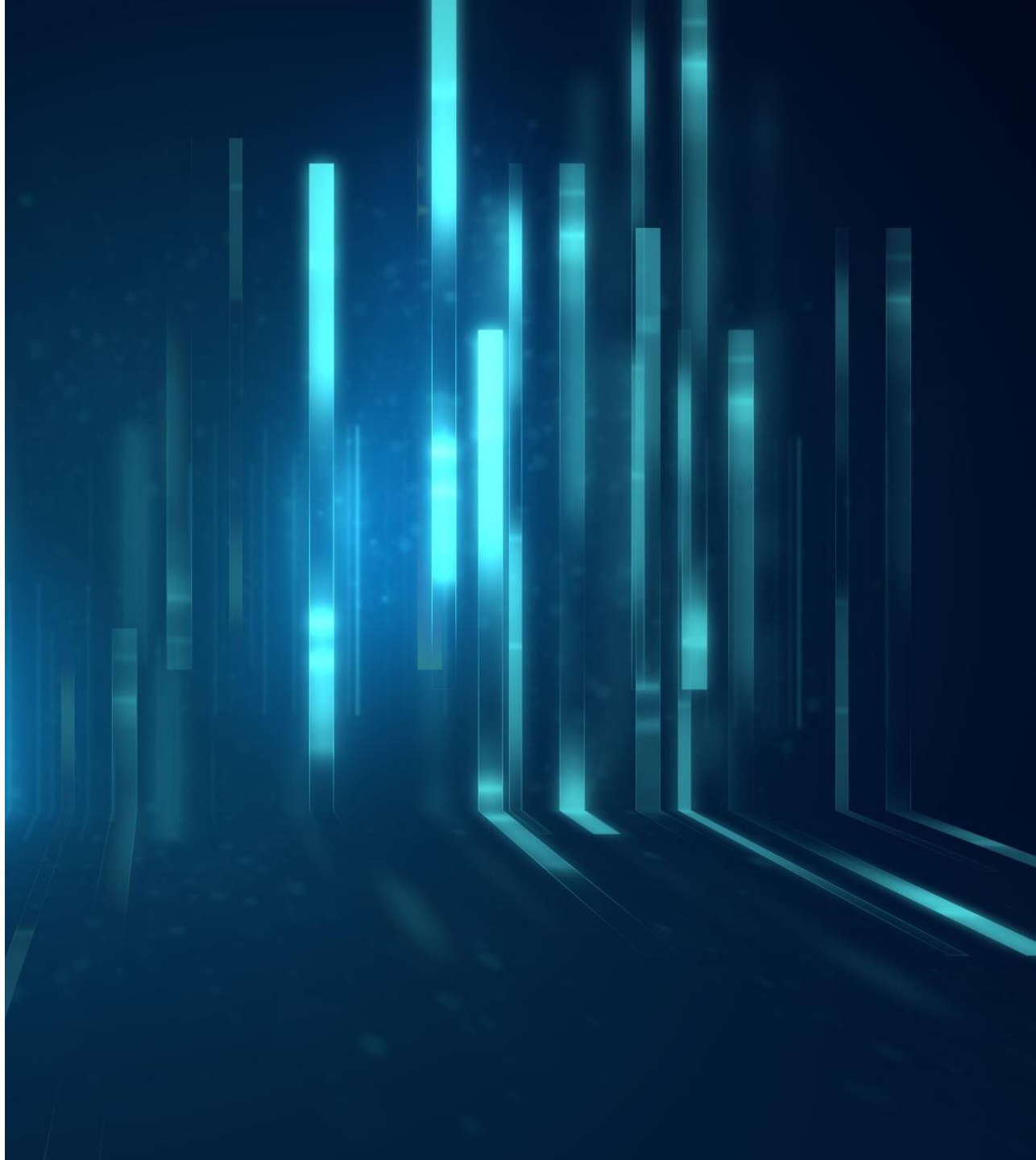
```
<head>
    <w pos="AT" sem="Z5"><m type="base">THE</m></w>
    <w pos="JJ" sem="Z2"><m type="base">SCOTS</m></w>
    <w norm="Apostasy" pos="NN1" sem="A1.7-/S9"><m type="base"><m type="prefix"
                >APO</m><m type="root">STACY</m></m></w>
    <w pos="." sem="PUNC">.</w>
</head>
<l>
    <w norm="Is it" pos="VBZ/PPH1" sem="A3+/Z8"><m type="base">IS</m><m
            type="contraction" baseForm="it">'t</m></w>
    <w pos="VV0" sem="A4.1[il.2.1"><m type="base">come</m></w>
    <w pos="II" sem="A4.1[il.2.2"><m type="base">to</m></w>
    <w pos="DD1" sem="Z8"><m type="base">this</m></w>
    <w pos="?" sem="PUNC">?</w>
    <w pos="DDQ" sem="Z8"><m type="base">what</m></w>
    <w pos="?" sem="PUNC">?</w>
    <w norm="Shall" pos="VM" sem="T1.1.3"><m type="base">fhall</m></w>
    <w pos="AT" sem="Z5"><m type="base">the</m></w>
    <w norm="Cheeks" pos="NN2" sem="B1"><m type="base">Cheek</m><m type="suffix"
            baseForm="s">es</m></w>
    <w pos="IO" sem="Z5"><m type="base">of</m></w>
    <w pos="NN1" sem="X2.2+"><m type="base">Fame</m></w>
</l>
```

Basic example

```
<head>THE SCOTS APOSTACY.</head>
<l>IS't come to this? what? fhall the Cheekes of Fame</l>
<l> Stretch't with the breath of learned <hi>Lowdens</hi> name </l>
<l>Be flagg'd againe, and that great peice of Sence</l>
<l>As rich in Loyaltie, as Eloquence,</l>
<l>Brought to the Teſt, be found a tricke of State?</l>
```

Tools for implementing linguistic annotation

- USAS tagger (for part-of-speech and semantic tagging) – can be implemented through Wmatrix and LancsBox.
- VARD (for normalization).
- NLTK (Natural Language Toolkit) (for lemmatization).



Demo of the USAS tagger