Corpus linguistics

2023-11-24

Corpus Building, XML, ELAN

Corpus building & compilation

- An important part of corpus linguistics is *corpus construction*

- Barth & Schnell (2022) distinguish three types of corpora

    - General corpus
    - Language documentation corpus
    - Research corpus

Barth, Danielle & Stefan Schnell. 2022. *Understanding Corpus Linguistics.* London: Routledge.

Corpus building & compilation

- General corpus
    - ‘over-studied’, large, can be linked to audio-visual content, pre-existing texts
    - for language planning, general reference
- Language documentation corpus
    - ‘under-studied’, small, generally linked to audio-visual content (circa 1990s),
    - for linguistic description, community work (e.g. dictionary production), indigenous pedagogy
- Research corpus
    - small, some type of special **annotation**, narrow or focused research question

Barth, Danielle & Stefan Schnell. 2022. *Understanding Corpus Linguistics.* London: Routledge.

Processing texts

- Processing a text can involve
  - Audio-visual recording
  - transcription
  - time-aligned at a specific level
    - phone
    - utterance
  - annotated at a specific level
    - phone
    - phoneme
    - grapheme
    - utterance ...
  - (translation into metalanguage)

Barth, Danielle & Stefan Schnell. 2022. *Understanding Corpus Linguistics.* London: Routledge.

Transcription and Annotation

- **Transcription** and **Annotation** are done in the first place to make the text searchable.

- If you want to know when people tend to pause or interrupt people in actual speech these have to be transcribed

- See transcription conventions from du Bois 1993, in Barth & Schnell 2022:101

Barth, Danielle & Stefan Schnell. 2022. *Understanding Corpus Linguistics.* London: Routledge.

Software for corpus construction

- ELAN
  - time-aligned with play back option
  - allows hierarchically organized tiers for annotation & multiple levels of transcription
  - typically for utterance level alignment
- FLEx
  - allows semi-automatic interlinear morph-level glossing
  - used for construction of lexicon or dictionary
  - has no linked audio
- Praat
  - for phone level alignment
  - for phonetic research

Barth, Danielle & Stefan Schnell. 2022. *Understanding Corpus Linguistics.* London: Routledge.

Software / Mark-up languages

- The softwares read,import and export files in different **mark-up languages**.

- Praat uses **.TextGrid** files

- ELAN uses **.eaf** files, a type of **XML file**

- Flex uses **.flextext** files, a type of **XML file**

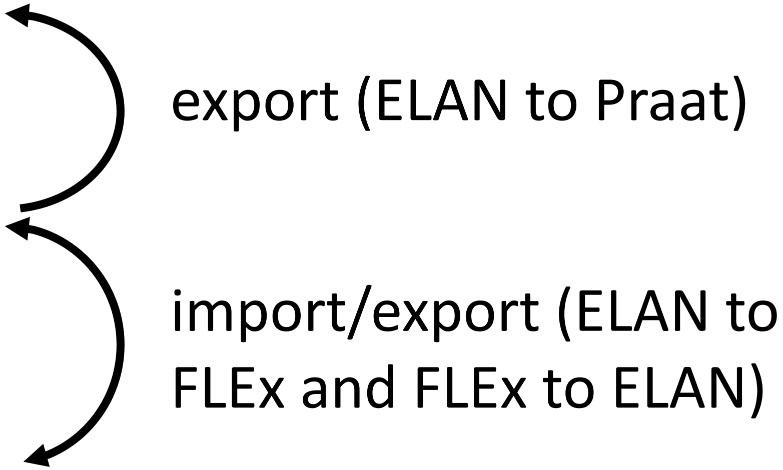
Software / Mark-up languages

- You can import and export files between the programs

- Praat uses **.TextGrid** files

- ELAN uses **.eaf** files, a type of **XML file**

- Flex uses **.flextext** files, a type of **XML file**



export (ELAN to Praat)

import/export (ELAN to FLEx and FLEx to ELAN)

XML grammar

- Tags
- Comments
- Document declaration
- Root element
- Trees and nodes
- Parsing an xml file

XML tags

- Tags are written in between pointy brackets

  <sometag>

- The tag must be succeeded by a **closing tag**

  </sometag>

- Tag names cannot begin with numbers or contain any of the following characters:

  ; @ # $ % ^ ( ) + ? =

  . – should be avoided

Based on Pozrikidis, C. 2013 XML in Scientific Computing. CRC Press

XML tags

- Tags

‘The linguistics class’

```
<sentence>
        <word>The</word>
         <word>linguistics</word>
        <word>class</word>
</sentence>
```

Based on Pozrikidis, C. 2013 XML in Scientific Computing. CRC Press

XML tags

- Tags have attributes, written with ‘=‘ and in quotations “”

‘The linguistics class’

```
<sentence id = “1”>
          <word pos= “D”>The</word>
           <word pos=“N”>linguistics</word>
          <word pos=“N”>class</word>
</sentence>
```

Based on Pozrikidis, C. 2013 XML in Scientific Computing. CRC Press

XML tags

- Note you always have the option of writing an attribute as another nested tag

```
<sentence id = “1”>
        <word><dem>The</dem></word>
         <word><noun>linguistics</noun></word>
        <word><noun>class</noun></word>
</sentence>
```

Based on Pozrikidis, C. 2013 XML in Scientific Computing. CRC Press

XML tags

- XML tags cannot ‘branch cross’

This:

<word><dem>The</dem></word>

Rather than this:

<dem><word><The</dem></word>

Based on Pozrikidis, C. 2013 XML in Scientific Computing. CRC Press

XML comments

- A lot of programming languages allow you to write in **comments**
- Basically the purpose of this in XML is so you can write comments that are ignored by **XML parsers**
- **You use** <!--...-->

```
<sentence id = “1”> <!--  This is the first utterance of our corpus, id gives us information about that -->
        <word><dem>The</dem></word>
         <word><noun>linguistics</noun></word>
        <word><noun>class</noun></word>
</sentence>
```

Based on Pozrikidis, C. 2013 XML in Scientific Computing. CRC Press

XML comments

- In programming **commenting out** lines of code is really important for debugging
- Comments are also important for understanding the author’s purpose in writing the code
- (Let’s see what happens if we ignore this line)
- We’ll see what this looks like in R

Based on Pozrikidis, C. 2013 XML in Scientific Computing. CRC Press

XML Document declaration

- The first line of an xml file **declares** that the file contains an *xml* document

```
<?xml version="1.0“ encoding =“utf-8”?>
<sentence id = “1”> <!--  This is the first utterance of
our corpus, id gives us information about that -->
        <word><dem>The</dem></word>
         <word><noun>linguistics</noun></word>
        <word><noun>class</noun></word>
</sentence>
```

Based on Pozrikidis, C. 2013 XML in Scientific Computing. CRC Press

Root elements

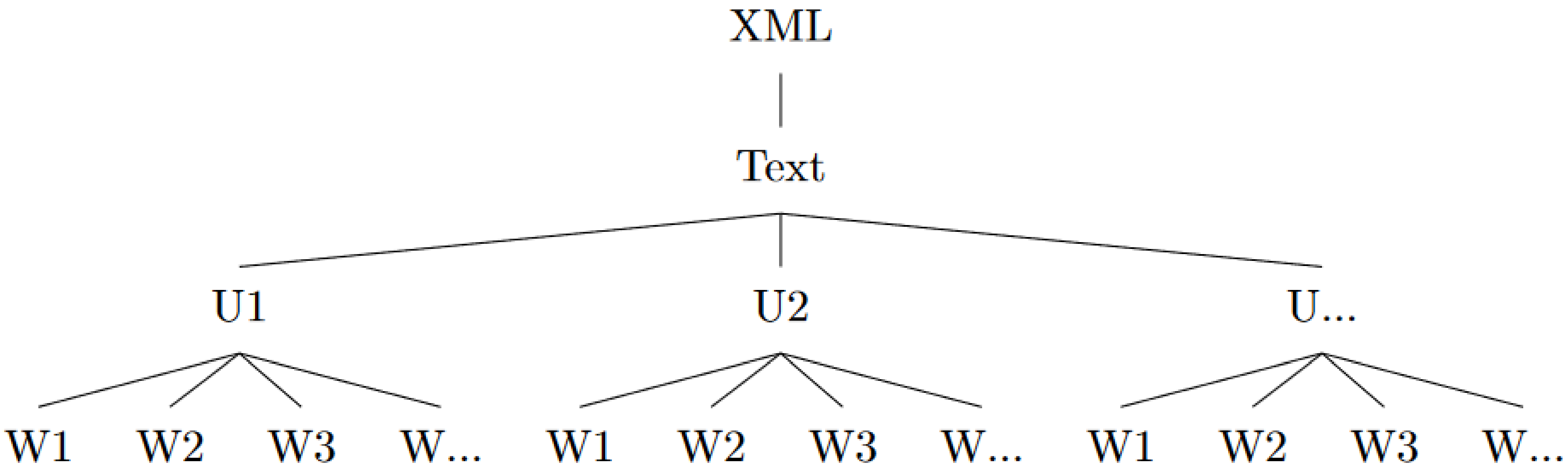
- The first tag after the declaration is a a **root element**

```
<?xml version="1.0“ encoding =“utf-8”?>
<story>
<sentence id = “1”> <!--  This is the first utterance of our
corpus, id gives us information about that -->
        <word><dem>The</dem></word>
         <word><noun>linguistics</noun></word>
        <word><noun>class</noun></word>
        <word><verb>is</verb></word>
        <word><adjective>exciting</adjective></word>
</sentence>
</story>
```

Based on Pozrikidis, C. 2013 XML in Scientific Computing. CRC Press

Trees and nodes

- Element nesting results in an **XML tree** originating from the **root**

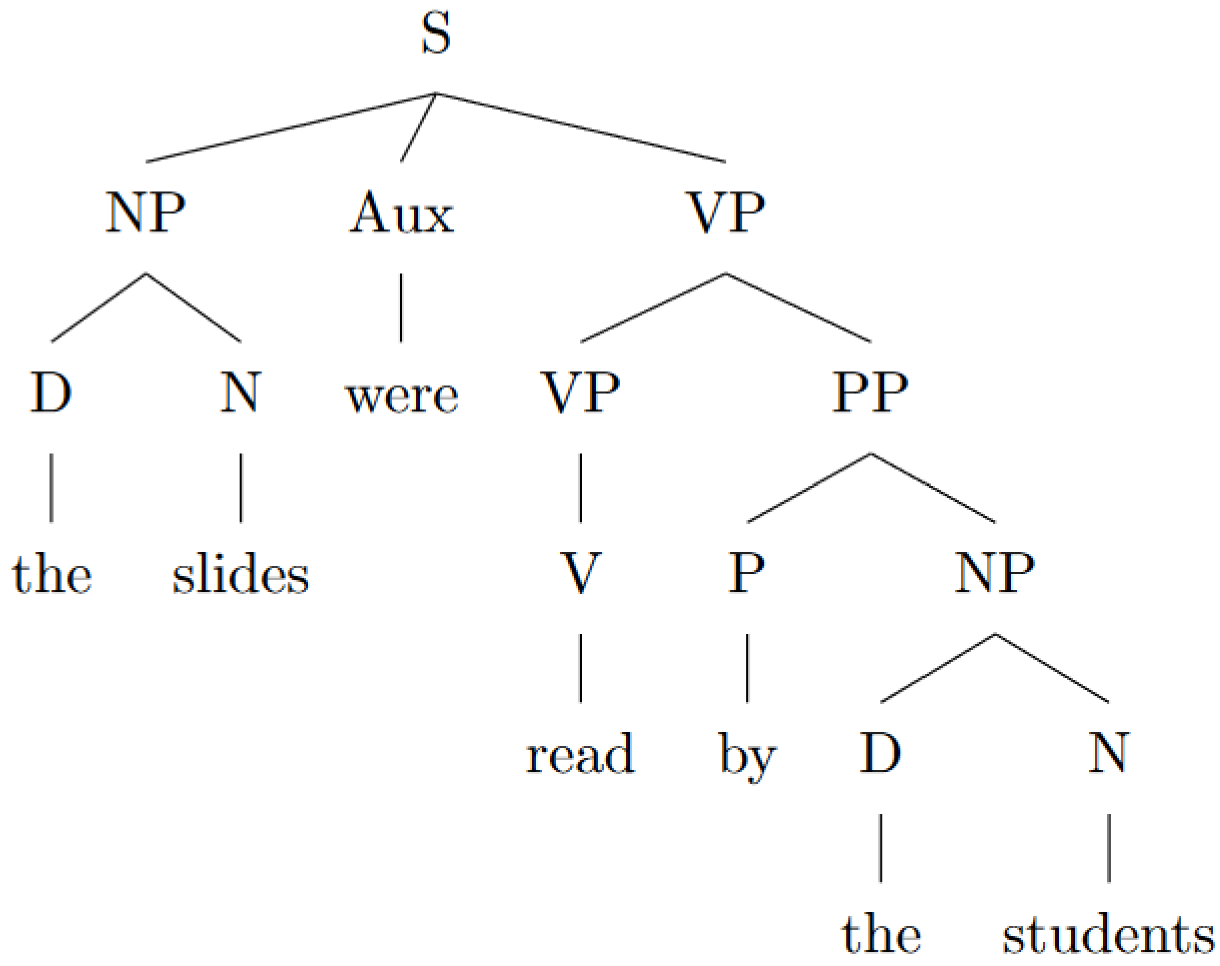


Inspired by materials on
https://alvinntnu.github.io/NTNU_ENC2036_LECTURES/xml.html

Trees and nodes

- Element nesting results in an **XML tree** originating from the **root**

```
<S>
        <NP><D>the</D>
        <N>slides</N></NP>
        <Aux>were</Aux>
        <VP><VP><V>read</V></VP>
        <PP><P>by</P>
        <NP><D>the</D>
        <N>students</N></NP></PP></VP>
</S>
```



Based on Pozrikidis, C. 2013 XML in Scientific Computing. CRC Press

Parsing XML

- Different softwares can **parse** XML code
- XML code has to be **parsed** to be used for corpus analysis
- In this course we will practice using R to parse

Gries, Stefan Th. 2017. *Quantitative Corpus Linguistics with R: A Practical Introduction.* Routledge.
Desaguiler, Guillaume. 2017. *Corpus Linguistics and Statistics with R.* Springer.
Barr, Dale J. 2015. Read ELAN XML files to tidy output. https://github.com/dalejbarr/elan/blob/master/README.md

ELAN

- People do not write up **XML** tagging by hand, but use some type of software that creates the **XML** files.
- A popular software, and that which is most used for language documentation, is **ELAN**.
- Please download if you have not already

https://archive.mpi.nl/tla/elan/download

ELAN and eaf files

- ELAN files are saved as **EAF** files, which are a type of **XML** file.
- EAF files are **time-aligned** to specific recordings – usually at the level of **utterances** (sentences or speech between pauses)
- Download the **WAV** file