XPath

How to run the exercises

This document contains xpath exercises. To execute them, you have two main options:

Using <oXygen/>:

Just open the xml document in the *<oXygen/> Editor* and put the required xpath expression in the window located just below the toolbar, and press *Return*

Using xmllint

From the terminal, use the *xmllint* tool to run the xpath queries. The format is the following:

```
xmllint --xpath "xpath_expression" file.xml
```

For instance, the following command will display the word forms of noun tokens in *s3aw.xml*

```
xmllint --xpath "//token[@pos='d']/wf" s3aw.xml
```

Notice how we use double quotes (") to enclose the whole expression, and single quotes when comparing the value of the attribute (`@pos='d'`)

1. Exercise

We are going to work with the *toponimos.xml* document (whose structure is defined by the DTD in *toponimos.dtd*), which comprises 4.000 toponym of the Basque Country. You should write the XPath expressions that answer the following questions:

1.1 List of the official name (`Oficial`) of the municipalities (whose `codigo` attribute value is `1.105`) [There are 688 municipalities].

1.2 List of the official name (`Nombre`) of hills (`Codigo` with value `1.601`) with more than 800m in height [420 hills].

1.3 In how many toponyms the official name (`Nombre`) is the same as the Basque name (`Oficial`)? [2051].

Note: You will need the XPATH 2.0 function `lower-case` (which is not available in `xmllint`).

2 Exercise

We will now work with the *dbe40_g.xml* document, which corresponds to the letter G of a Spanish dictionary. The structure is defined in *teip4_dict_DBE_simplificado.dtd.* Write and execute the Xpath expressions that answer the following questions:

2.1 How many definitions (`def`) contain the word *como*? [37 definitions].

2.2 How many definitions (`def`) or examples (`q` element inside `eg`) contain the word *persona*? [The correct answer is 63].

2.3 Entries (`entry`) whose POS is `adj.` [62 entries].

Note: Here we refer to the entry level category, not to the category of the senses. That is, we `pos` inside `gramGrp` which are direct child of `entry` element.

2.4 How many entries (`entry`) have POS `vintr` in any of its senses? [12 entries].

2.5 Get the POS value of entry with id `g_d0e7458`. [Result: `sf.`].

2.6 Get the headword (`form[1]/orth`) of entries (`entry`) that have three or more senses (`entry/sense`) [60 entries].

Note: We want senses that are direct children of `entry` elements.

2.7 Get corredor synonyms: headword (`form[1]/orth`) of entries whose synonym (`xr/ref`, where `lbl='Sin.'`) is the word *corredor*. [Result: *galería*].

2.8 [advanced] Headword (`form[1]/orth`) of entries that have some synonym (`xr/ref`, where `lbl='Sin.'`) [109 headwords].

2.9 [advanced] Headword of the entry that contains the word *enfrentó* in its definition or example? [Result: *gallardía*].

Exercise 3

Let's work with the document *s3aw.xml,* whose structure is defined in *s3aw.dtd.* The document contains text that is linguistically annotated at several levels (sentence segmentation, tokenization and sense annotations). Write and execute the Xpath expressions that answer the following questions:

Basic

3.1 Get word forms (`wf` elements)

3.2 Get the textual form of `wf` elements

Counting

3.3 How many words, sentences

3.4 How many words with sense attached [805]

Note that there are words whose sense is `Unknown` (20); what to do with these?Try this expression:

3.5 How many ambiguous words? [11]

3.6 [advanced] Get id and lemmas of ambiguous words

Attributes

3.7 Different POS tags [16 different values]

3.8 How many nouns? [389]

3.9 Lemmas of nouns

3.10 Lemmas of ambiguous words

Axis, etc

3.11 Which is the POS tag of content words (i.e. words manually tagged with senses) [`v`, `n`, `a` and `r`]

3.12 Obtain word forms whose sense tag starts with `"be%"` [`'s`, `Was`, `were`, `is`, `was`, `been` and `be`]

3.13 [advanced] If the word is ambiguous, obtain its word form, its lemma, and the lemma of the word immediately following it [11 triplets]