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*

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

- Open *s3aw.xml* in *<oxYgen/> Editor*
- Go to the search window and write: //token[@pos='n']/wf

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='n']/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='n'`)

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].

```
/toponimos/lista_toponimos/toponimo [Codigo='1.105']/Oficial

//toponimo[Codigo='1.105']/Oficial
```

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

```
/toponimos/lista_toponimos/toponimo[Codigo='1.601'][Altitud>800]/Nombre

/toponimos/lista_toponimos/toponimo[Codigo='1.601' and Altitud>800]/Nombre
```

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`).

```
count(/toponimos/lista_toponimos/toponimo/Nombre[lower-case(.)=lower-case(following-sibling::Oficial)])

count(//Nombre[lower-case(.)=lower-case(following-sibling::Oficial)])

count(//toponimo[lower-case(Nombre)=lower-case(Oficial)])
```

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].

```
count(//def[contains(.,'como')])
```

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

```
count(//q[contains(.,'persona')]|//def[contains(.,'persona')])
```

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.

```
//entry[gramGrp/pos='adj.']
```

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

```
count(//entry[gramGrp/pos='vintr.']|//entry[sense/gramGrp/pos='vintr.'])

count(//entry[.//gramGrp/pos='vintr.'])

count(//entry[.//pos='vintr.'])
```

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

```
//entry[@id='g_d0e7458']/gramGrp/pos

//pos[ancestor::entry[@id='g_d0e7458']]
```

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.

```
//entry/form[1]/orth[count(../../sense)>=3]

//entry/form[1]/orth[count(ancestor::entry/sense)>=3]
```

```
//entry[count(sense)>=3]/form[1]/orth
```

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

```
//entry[.//xr[lbl='Sin.']/ref='corredor']/form[1]/orth

//entry/form[1]/orth[../..//ref[.='corredor' and ../lbl='Sin.']]

//entry/form[1]/orth[ancestor::entry//ref[.='corredor' and ../lbl='Sin.']]
```

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

```
//entry[.//xr[lbl='Sin.']]/form[1]/orth

//entry/form[1]/orth[../..//xr[lbl='Sin.']]

//entry/form[1]/orth[ancestor::entry//xr[lbl='Sin.']]
```

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

```
//entry[.//q[contains(., 'enfrentó')] or
.//def[contains(.,'enfrentó')]]/form[1]/orth

//entry/form[1]/orth[ancestor::entry//q[contains(. ,'enfrentó')]]|
//entry/form[1]/orth[ancestor::entry//def[contains(. ,'enfrentó')]]

//entry/form[1]/orth[ancestor::entry//q[contains(.,'enfrentó')] or
ancestor::entry//def[contains(.,'enfrentó')]]
```

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)

```
//wf
```

3.2 Get the textual form of `wf` elements

```
//wf/text()
```

Counting

3.3 How many words, sentences

```
count(//token)

count(//s)
```

3.4 How many words with sense attached [805]

```
count(//token[lexsn])
```

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

```
count(//token[lexsn[.!='Unknown']]) [Resultado: 785]
```

3.5 How many ambigous words? [11]

```
count(//token[count(lexsn)>1])
```

3.6 [optational] Get id and lemmas of ambiguous words

```
//token[count(lexsn)>1]/@id | //token[count(lexsn)>1]/@lemma

//token[count(lexsn)>1]/(@id |@lemma)
```

Note: Using `for` in XPath 2.0:

```
for $t in //token[count(lexsn)>1] return concat($t/@id, " ", $t/@lemma)
```

Atributess

3.7 Different POS tags [existen 16 categorías diferentes]

```
distinct-values(//@pos)

distinct-values(/context/s/token/@pos)
```

3.8 How many nouns? [389]

```
count(//token[@pos='n'])
```

3.9 Lemmas of nouns

```
//token[@pos='n']/@lemma
```

3.10 Lemmas of ambiguous words

```
//token[count(lexsn)>1]/@lemma
```

Axis, etc

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

```
distinct-values(/context/s/token[@interp]/@pos)
```

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

```
distinct-values(//token[lexsn[starts-with(., "be%")]]/wf)
```

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

```
//token[@interp>1]/(wf | @lemma | following-sibling::token[@lemma][1]/@lemma)
```

Nota: Using `for` in XPath 2.0:

```
for $t in //token[@interp>1] return concat ($t/wf, ", ", $t/@lemma, ", ",
$t/following-sibling::token[@lemma][1]/@lemma)
```