

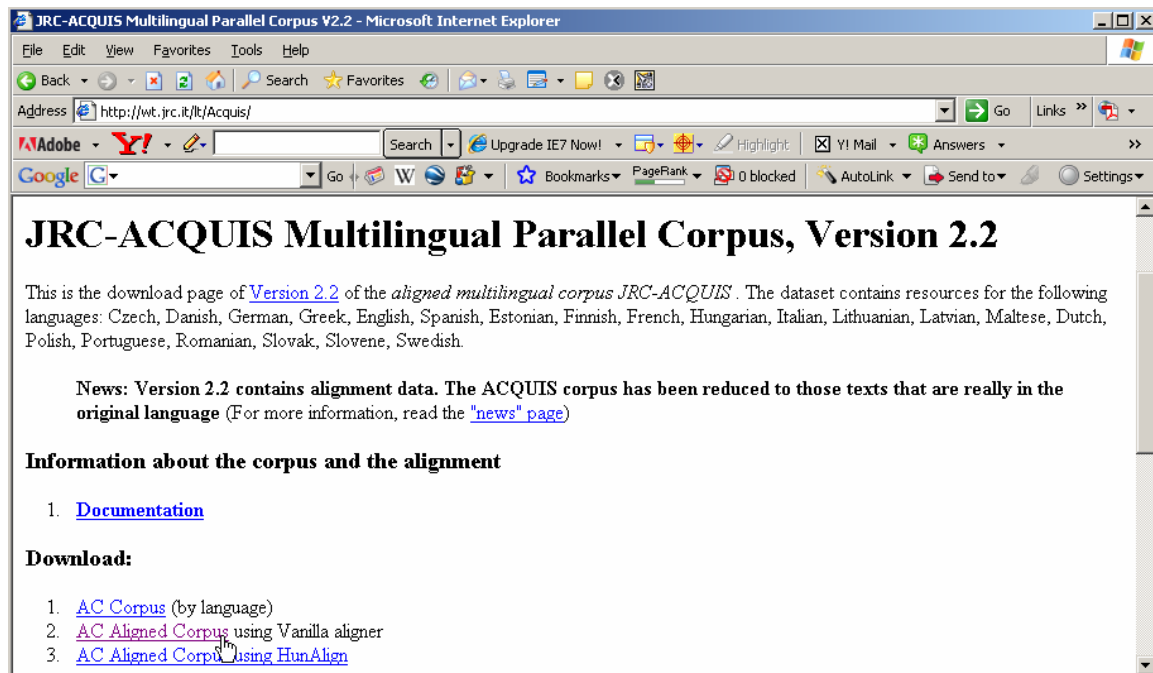
How to compile an aligned corpus

(on Windows system)

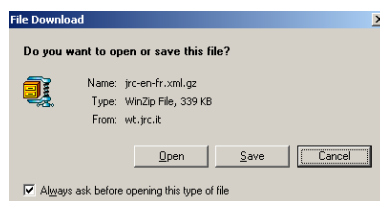
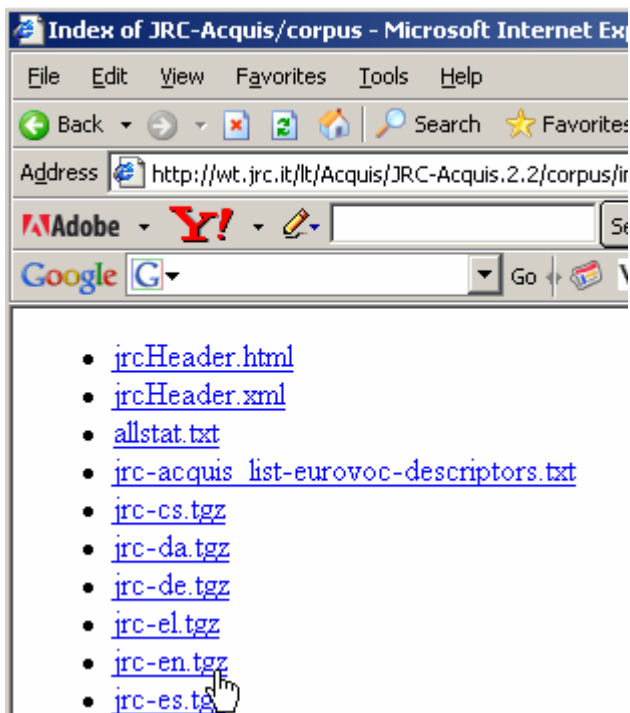
Get the two corpora

Download the two corpora (here : en for English, fr for French)

Open a browser on “<http://langtech.jrc.it/>” follow the link “**The [JRC-Acquis aligned parallel corpus](#)**” then “[Download the JRC-Acquis corpus](#)”

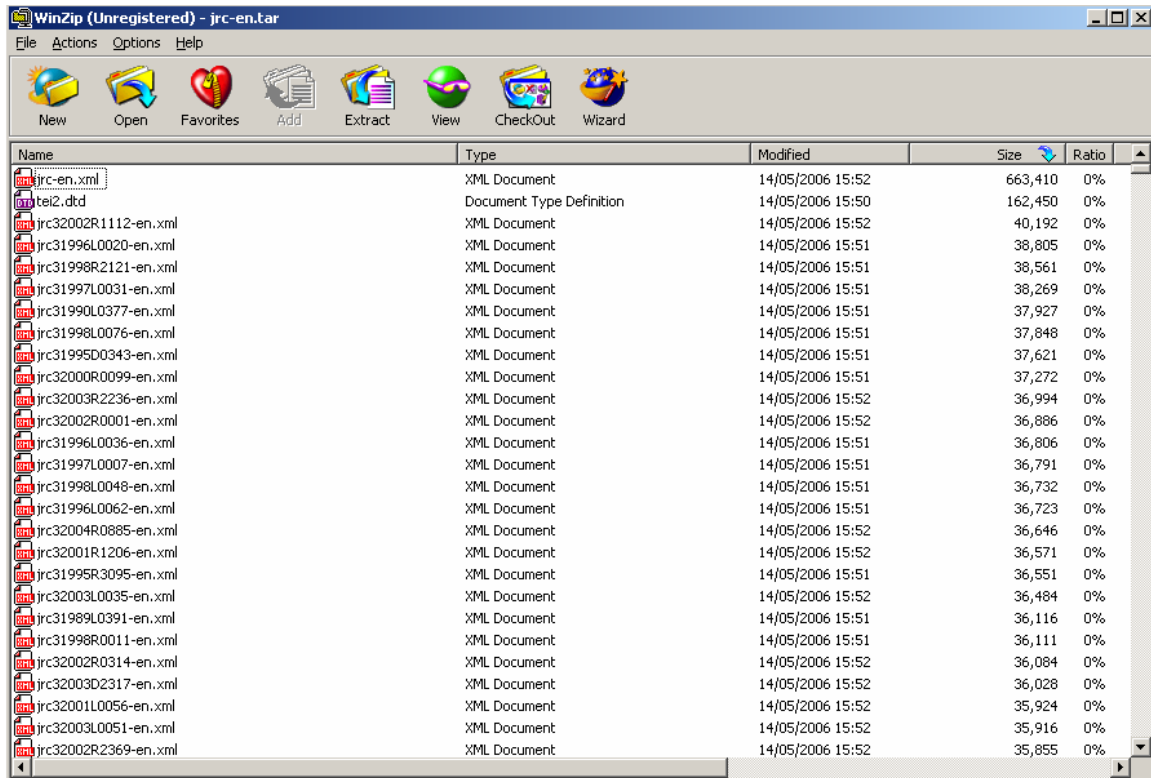


Click on the first link (AC Corpus) then select the language needed “jrc-en.tgz”

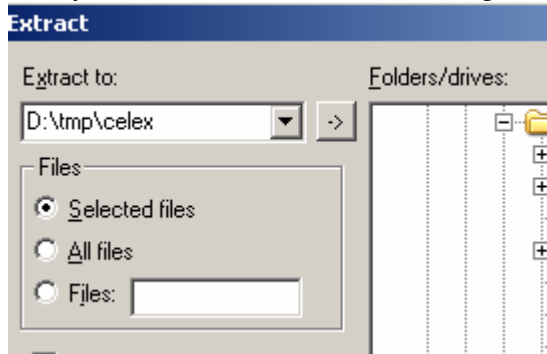


Your computer may display a warning:

Depending on your computer settings, you may save first the “unzipped” version of the file (in which case you should rename it as “jrc-en.tar”, before being able to “extract” all the files from it). If you use “winzip” you should be able to see something like:



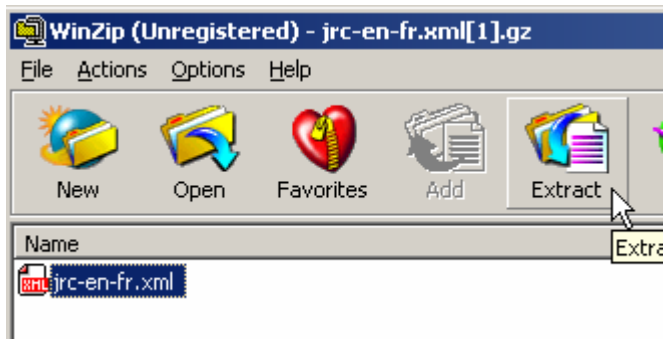
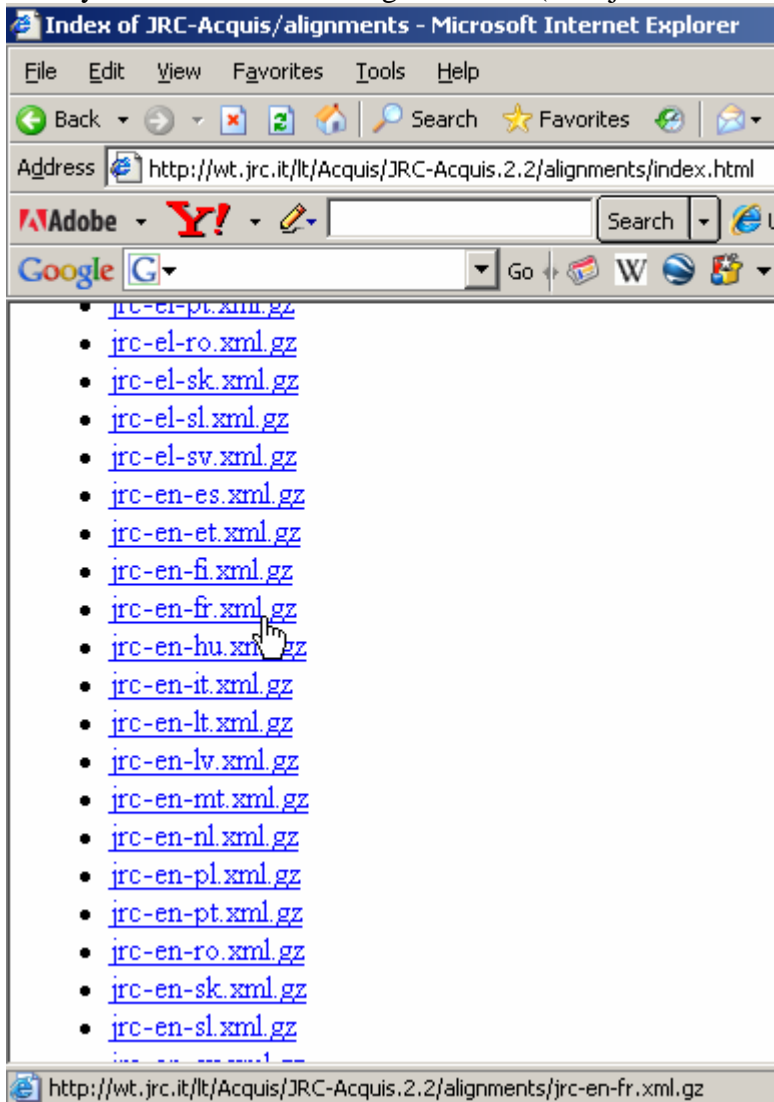
then you can extract all the files in a given directory (here d:\tmp\Celex)



Do the same with the other language (here French).

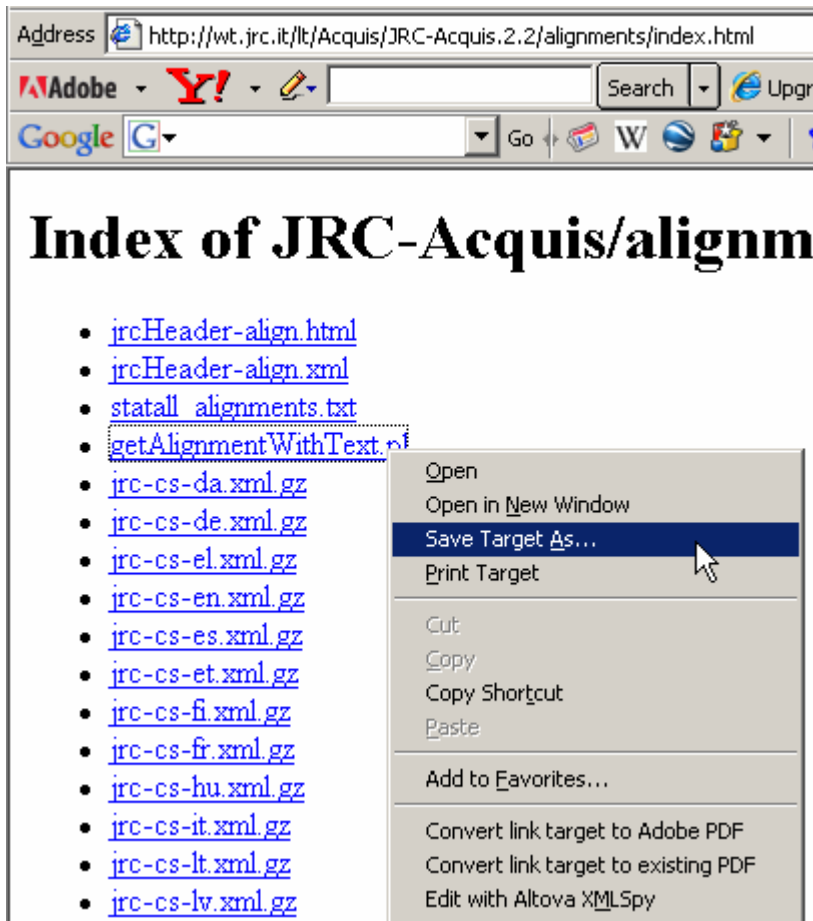
Get the alignment file

Then you have to save the alignment file (here jrc-en-fr.xml)



Get the Perl script

Then download the program itself (getAlignmentWithText.pl)



The screenshot shows a web browser window with the address bar displaying `http://wt.jrc.it/lt/Acquis/JRC-Acquis.2.2/alignments/index.html`. The browser's toolbar includes logos for Adobe, Y!, and Google, along with a search box and navigation buttons. The main content area features the heading "Index of JRC-Acquis/alignm" and a list of links. A context menu is open over the link `getAlignmentWithText.pl`, with the "Save Target As..." option highlighted by the mouse cursor. The menu options include "Open", "Open in New Window", "Save Target As...", "Print Target", "Cut", "Copy", "Copy Shortcut", "Paste", "Add to Favorites...", "Convert link target to Adobe PDF", "Convert link target to existing PDF", and "Edit with Altova XMLSpy".

Address `http://wt.jrc.it/lt/Acquis/JRC-Acquis.2.2/alignments/index.html`

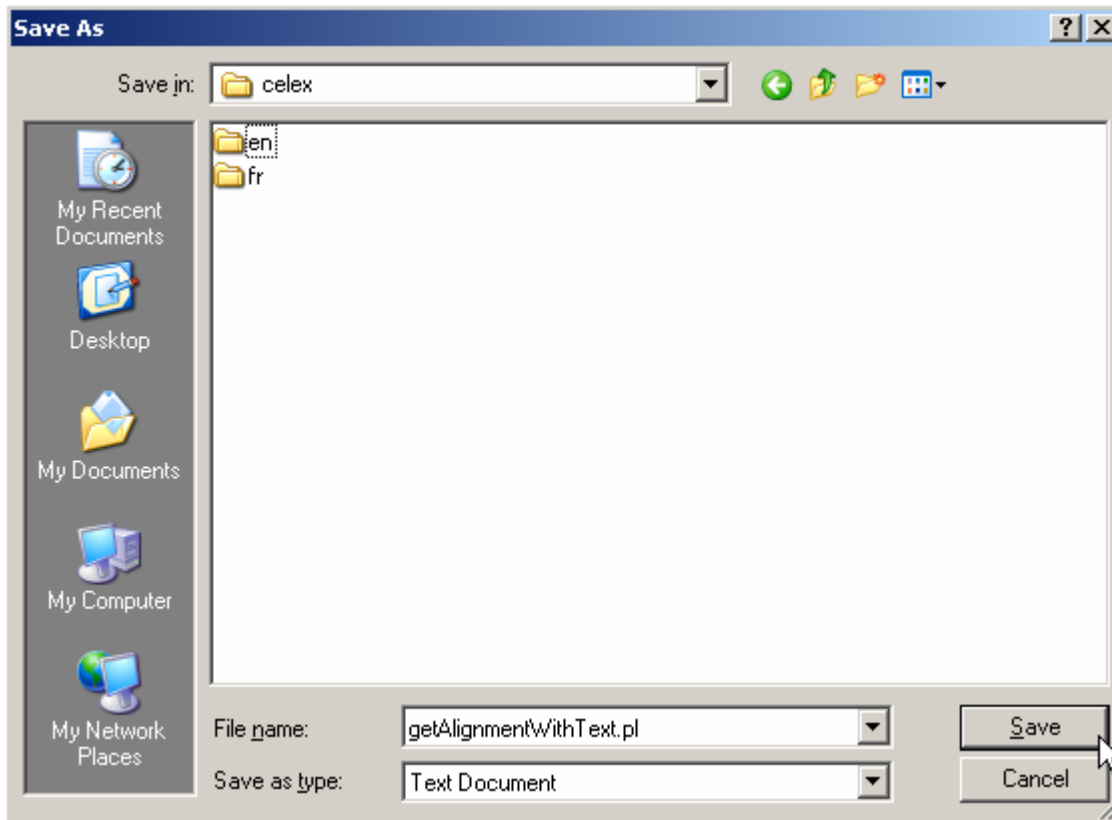
Adobe Y! Search Upgr

Google G Go W

Index of JRC-Acquis/alignm

- [jrcHeader-align.html](#)
- [jrcHeader-align.xml](#)
- [statall_alignments.txt](#)
- [getAlignmentWithText.pl](#)
- [jrc-cs-da.xml.gz](#)
- [jrc-cs-de.xml.gz](#)
- [jrc-cs-el.xml.gz](#)
- [jrc-cs-en.xml.gz](#)
- [jrc-cs-es.xml.gz](#)
- [jrc-cs-et.xml.gz](#)
- [jrc-cs-fi.xml.gz](#)
- [jrc-cs-fr.xml.gz](#)
- [jrc-cs-hu.xml.gz](#)
- [jrc-cs-it.xml.gz](#)
- [jrc-cs-lt.xml.gz](#)
- [jrc-cs-lv.xml.gz](#)

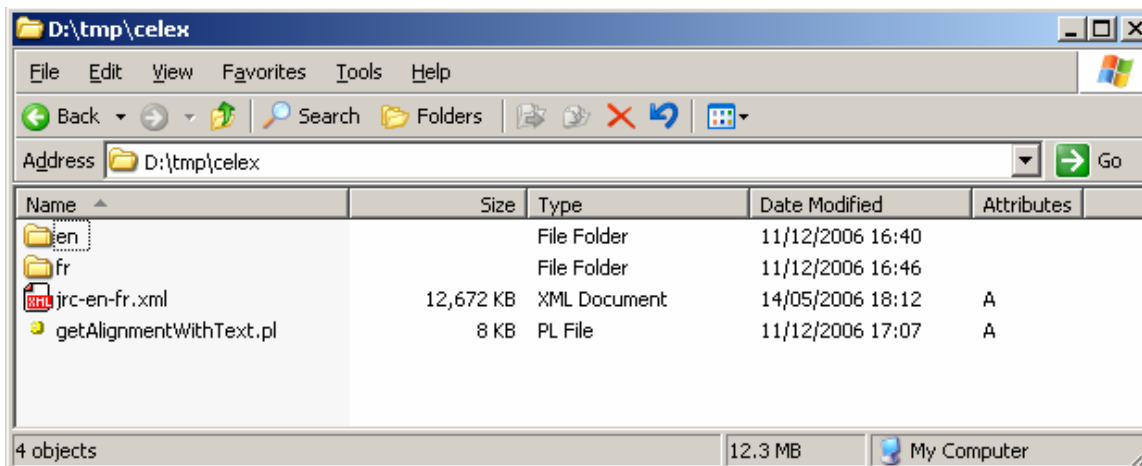
Open
Open in New Window
Save Target As...
Print Target
Cut
Copy
Copy Shortcut
Paste
Add to Favorites...
Convert link target to Adobe PDF
Convert link target to existing PDF
Edit with Altova XMLSpy



if you do not have Perl installed, you can get a version from various sources, one of them is ActivePerl available for free at <http://www.activestate.com/store/activeperl/>

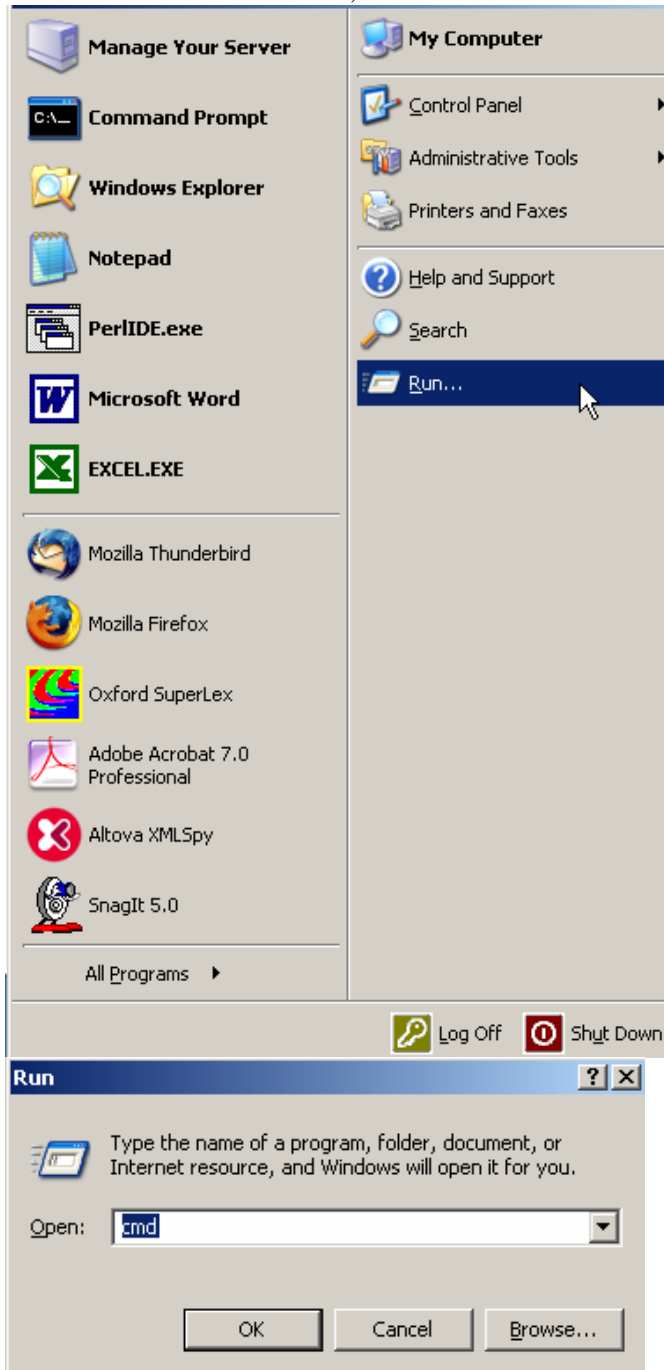
Produce the aligned corpus

If everything is right you should see the following directory:



At that point you are able to launch the Perl script:

Click on the “Start” button, then choose “run...”



you have to “go” in the directory you stored all the files and then execute the perl script with the alignment file as parameter:

```
C:\WINDOWS\system32\cmd.exe - perl getAlignmentWithText.pl jrc-en-fr.xml
Microsoft Windows [Version 5.2.3790]
(C) Copyright 1985-2003 Microsoft Corp.

C:\Documents and Settings\poulibr>d:
D:\>cd \tmp\celex
D:\tmp\celex>perl getAlignmentWithText.pl jrc-en-fr.xml > alignedCorpus-en-fr.xml
1
```

then the file alignedCorpus-en-fr.xml will be created in the same directory (it takes about 5 minutes)

you can verify roughly the content of the aligned corpus using “more”:

```
more alignedCorpus-en-fr.xml

...
<link type="1-1" xtargets="13;13">
<s1>HAVE AGREED as follows:</s1>
<s2>SONT CONVENUS de ce qui suit:</s2>
</link>
      <link type="1-1" xtargets="14;14">
<s1>Article 1</s1>
<s2>Article premier</s2>
</link>
...
```