Processing of EU Multilingual Corpora

M.T. Carrasco Benitez

Propor 2016 - Corpora Workshop

Tomar, 13 July 2016

European Commission

Processing of EU Multilingual Corpora

M.T. Carrasco Benitez

Propor 2016 - Corpora Workshop

Tomar, 13 July 2016

Good translation is expensive

Estimated translation cost: 44 billions euros

All EU institutions/bodies 1952 - 2016

One must re-use

Industrial translation

24 languages

High volume

Non-literary

Big multilingual data

Millions of files

Hundred of millions of segments

Machine processing

Unrealistic manually

Input data

Each collection is different

Packaging

Path and file naming

File formats

No single solution

Toolbox

Objectives

Align all texts - log un-extracted

Align 24 languages - not bilingual

Publish data and programs

Easy to repeat the processing

N-language 1

Bilingual: individual point of view

N-lingual: industrial point of view

Most alignment are bilingual

PT-EN PT-FR : EN-FR

N-language 2

Better for programatic comparison

Clustering of languages

Bilingual bad

Much harder, worth the effort

Processing approach

Align at file level

Keep original structure

Metadata might be in the path

Create map of linguistic sets

Linguistic sets

```
set N
  en datahome/foo/bar/123.txt
  fr datahome/something/else
  pt datahome/more/here/myfile
```

JSON file

Official Journal of the EU

2004-2011

Serie: L - C - CA

Formex4 - XML

Format: best case

Non-parallel XML - repairing

Other formats might be harder

mux

Multilingual Data Toolbox

Toolbox: no universal solution

Set of Linux commands

Mostly in Python

Any XML parallel sets

mux - Approach

Stepwise

Intermediate output

Collection particularities

Common aspects

mux - Install

Unzip anywhere mux.zip - directory mux

source /foo/mux/msetup

Or insert in "~/.bashrc"

Three environment variables

muxhome PATH PYTHONPATH

mux - Run

conf file

```
collection=oj
dir_data_ori=/foo/mydata
```

mrun - daemonised

mux - mrun

mopen: open data packages

map: create set map - JSON

extract: extract

meter: measuring, statistics, etc

mux - Data output

Files: stripped files
Sets: three presentations
Segments: three presentations - Moses style

mux - Auxiliary output

Maps

Statistics

Logs: journal, stdout, stderr, debug, dump

mux - Comparative lengths

Central value

Language independent - not English

Corpus specific ratios

Multilingual production

Authoring, translation, publishing

Consider data reuse

Generation of Multilingual Parallel Documents

dragoman.org/propor