Parallel corpora and tools

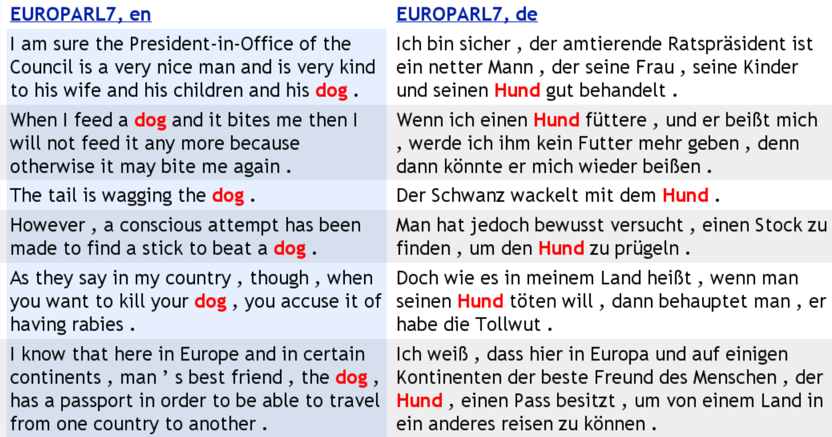
Jan Michelfeit

Lexical 文 Computing

jan.michelfeit@sketchengine.co.uk

6th Sketch Engine Workshop
Herstmonceux, August 10, 2015

Multilingual data



EUROPARL7, en	EUROPARL7, de
I am sure the President-in-Office of the Council is a very nice man and is very kind to his wife and his children and his **dog** .	Ich bin sicher , der amtierende Ratspräsident ist ein netter Mann , der seine Frau , seine Kinder und seinen **Hund** gut behandelt .
When I feed a **dog** and it bites me then I will not feed it any more because otherwise it may bite me again .	Wenn ich einen **Hund** füttere , und er beißt mich , werde ich ihm kein Futter mehr geben , denn dann könnte er mich wieder beißen .
The tail is wagging the **dog** .	Der Schwanz wackelt mit dem **Hund** .
However , a conscious attempt has been made to find a stick to beat a **dog** .	Man hat jedoch bewusst versucht , einen Stock zu finden , um den **Hund** zu prügeln .
As they say in my country , though , when you want to kill your **dog** , you accuse it of having rabies .	Doch wie es in meinem Land heißt , wenn man seinen **Hund** töten will , dann behauptet man , er habe die Tollwut .
I know that here in Europe and in certain continents , man ' s best friend , the **dog** , has a passport in order to be able to travel from one country to another .	Ich weiß , dass hier in Europa und auf einigen Kontinenten der beste Freund des Menschen , der **Hund** , einen Pass besitzt , um von einem Land in ein anderes reisen zu können .

Implementation

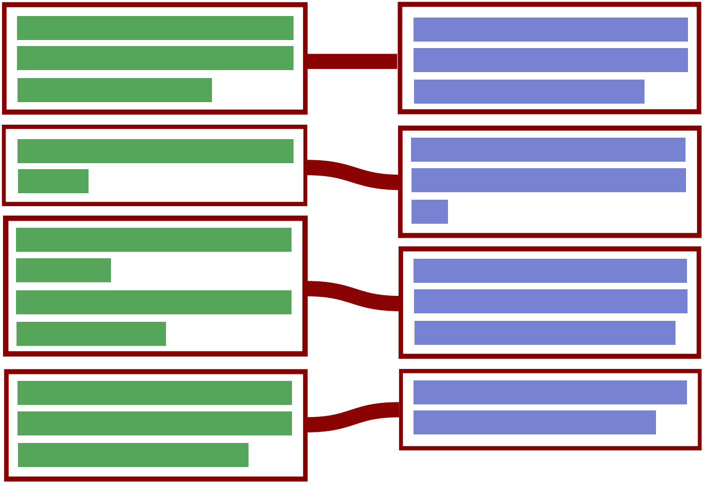
1 : 1 alignment

- aligned using a special structure (`<align>`)
- simple to understand

1 : 1 alignment

- aligned using a special structure (`<align>`)
- simple to understand
- but everything must be aligned (no gaps)
- and not really useful for more than two languages

1 : 1 alignment



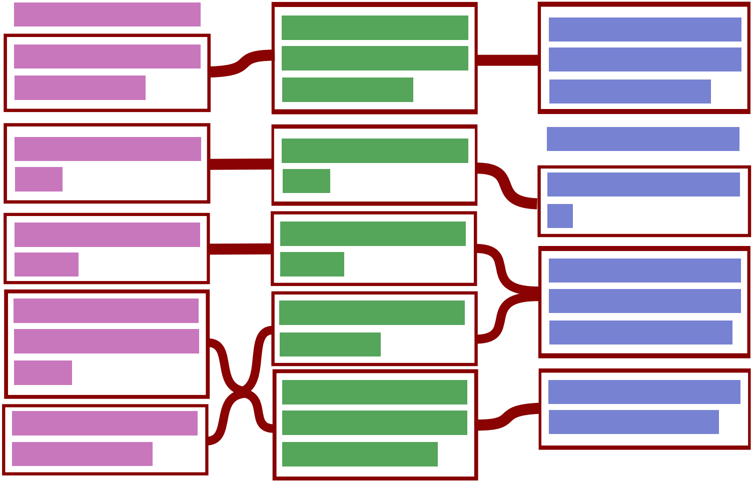
$m : n$ alignment

- m sentences to n sentences
- needs a mapping file for every pair of corpora
- gaps are allowed
- one corpus for each language, aligned to each other

m : n alignment

- m sentences to n sentences
- needs a mapping file for every pair of corpora
- gaps are allowed
- one corpus for each language, aligned to each other
- mapping must still be sequential

m : n alignment



Preloaded corpora

Preloaded corpora

- Old 1 : 1 corpora: ~~Europarl3, OPUS~~
- Europarl7 (2M English sentences, spoken, by Philipp Koehn)
- DGT (4M English sentences, translation memory)
- OPUS2 (130M English sentences)

Preloaded corpora

- Old 1 : 1 corpora: ~~Europarl3, OPUS~~
- Europarl7 (2M English sentences, spoken, by Philipp Koehn)
- DGT (4M English sentences, translation memory)
- OPUS2 (130M English sentences):
  - EU documents
  - UN documents
  - movie subtitles
  - Tatoeba
  - software localization
  - etc.

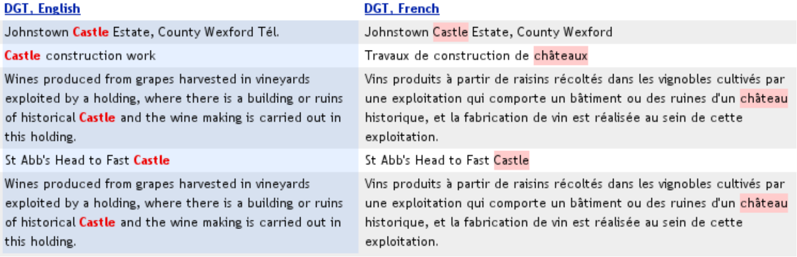
  (based on data from the OPUS project by Jörg Tiedemann)

Upcoming preloaded corpora

- DCEP (EP – EuroParl)
- JRC-Acquis
- update OPUS, make subcorpora directly visible

Translation candidate highlighting

- using statistical dictionaries generated from parallel corpora

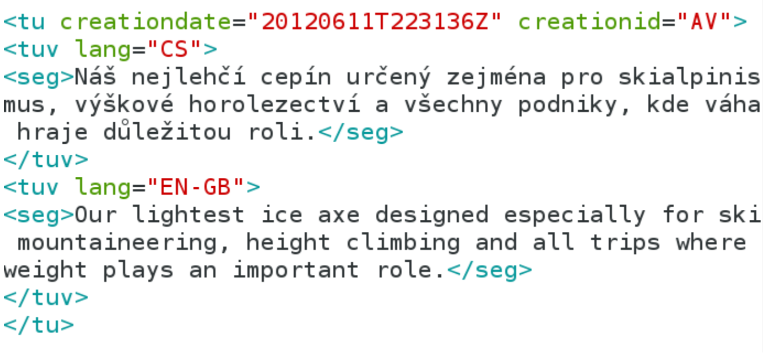


DGT, English	DGT, French
Johnstown **Castle** Estate, County Wexford Tél.	Johnstown Castle Estate, County Wexford
Castle construction work	Travaux de construction de châteaux
Wines produced from grapes harvested in vineyards exploited by a holding, where there is a building or ruins of historical **Castle** and the wine making is carried out in this holding.	Vins produits à partir de raisins récoltés dans les vignobles cultivés par une exploitation qui comporte un bâtiment ou des ruines d'un château historique, et la fabrication de vin est réalisée au sein de cette exploitation.
St Abb's Head to Fast **Castle**	St Abb's Head to Fast Castle
Wines produced from grapes harvested in vineyards exploited by a holding, where there is a building or ruins of historical **Castle** and the wine making is carried out in this holding.	Vins produits à partir de raisins récoltés dans les vignobles cultivés par une exploitation qui comporte un bâtiment ou des ruines d'un château historique, et la fabrication de vin est réalisée au sein de cette exploitation.

User corpora

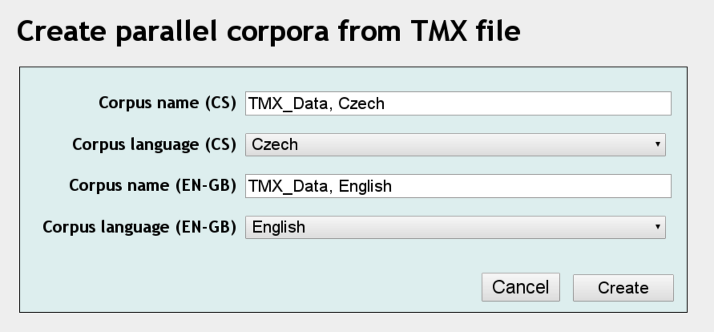
TMX

- TMX is a good input format for 1 : 1 bilingual data



```
<tu creationdate="20120611T223136Z" creationid="AV">
<tuv lang="CS">
<seg>Náš nejlehčí cepín určený zejména pro skialpinis
mus, výškové horolezectví a všechny podniky, kde váha
 hraje důležitou roli.</seg>
</tuv>
<tuv lang="EN-GB">
<seg>Our lightest ice axe designed especially for ski
 mountaineering, height climbing and all trips where
weight plays an important role.</seg>
</tuv>
</tu>
```

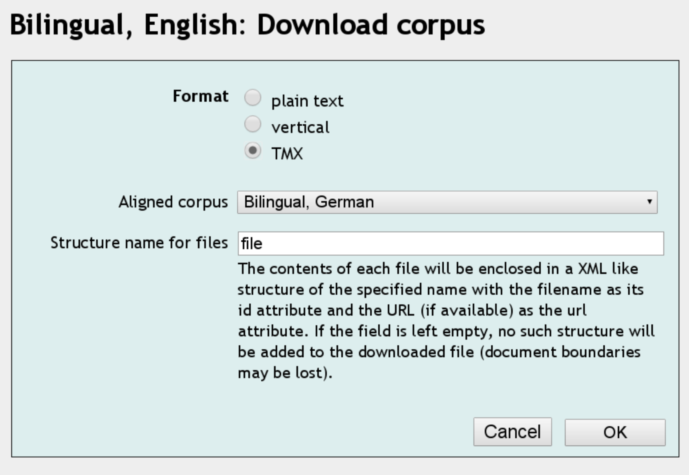
TMX import



Create parallel corpora from TMX file

Corpus name (CS)	TMX_Data, Czech
Corpus language (CS)	Czech
Corpus name (EN-GB)	TMX_Data, English
Corpus language (EN-GB)	English

TMX export



Bilingual, English: Download corpus

Format ○ plain text
○ vertical
● TMX

Aligned corpus: Bilingual, German

Structure name for files: file

The contents of each file will be enclosed in a XML like structure of the specified name with the filename as its id attribute and the URL (if available) as the url attribute. If the field is left empty, no such structure will be added to the downloaded file (document boundaries may be lost).

The future

- automatic sentence-alignment of non-aligned (or document-aligned) data
- manual adjustment of aligned chunks
- import from Excel spreadsheets (i.e. parallel corpora for normal people ;)