# Bitext

Multilinguisme — lecture n°1

Guillaume Wisniewski
guillaume.wisniewski@linguist.univ-paris-diderot.fr
September 2019

Université de Paris & LLF



## Definition

### Bitext

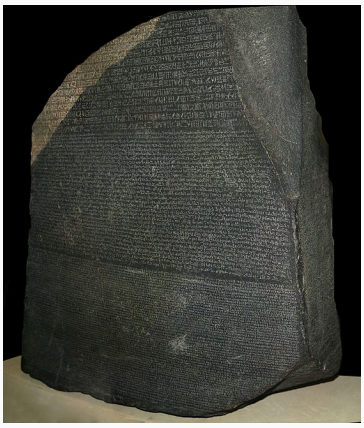
- originally: translation studies [Harris, 1988]
- documents ⊕ translation in another language

### Now...

- pair of texts...
- ... with some translational equivalence
- two kinds of bitexts:
  1. parallel texts = document + its translation
     ↪ *Candide* and its translation in English
  2. comparable corpora = documents in different languages about the same domain but necessarily translation of each other
     ↪ all the tweets about FIFA World Cup
- note: in translation studies parallel text = comparable corpora



## Example (1)



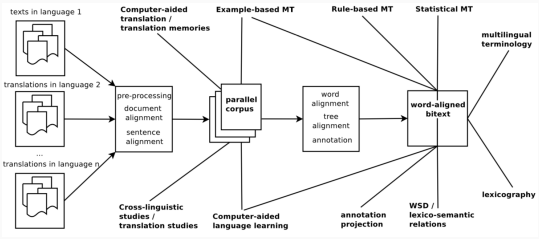
The Rosetta Stone



## Example (2)




## Parallel corpora & their applications






## How to create a parallel corpus...

## Source for Translations (1)



## Source for Translations (1)

- obvious answer: ask people to translate texts (manual annotation)



## Source for Translations (1)

- obvious answer: ask people to translate texts (manual annotation)
- corpus size to train a state-of-the-art system:

| | # parallel sentences |
|---|---|
| fr-en | 180.4M |
| fr-es | 99.5M |
| fr-eu | 2.2M |
| fr-nap | 49 |

source: http://opus.nlpl.eu/



## Source for Translations (1)

- obvious answer: ask people to translate texts (manual annotation)
- corpus size to train a state-of-the-art system:

| | # parallel sentences |
|---|---|
| fr-en | 180.4M |
| fr-es | 99.5M |
| fr-eu | 2.2M |
| fr-nap | 49 |

source: http://opus.nlpl.eu/

↪ manual translation is not an option



## Source for Translations (2)



## Source for Translations (2)

1. international organization (ONU, European Parliament, Hansard, ...)
   ↪ most of their documents are freely available & translated into several languages
2. software manuals (in particular: free software)
3. classical books
4. subtitles

## Source for comparable data



## Source for comparable data

Twitter

↪ using hashtag to identify tweets about the same topic

Wikipedia

↪ using interlanguage links



## Setup to build parallel corpora

1. data collection: identify source of documents & fetch them
2. pre-processing: extract text from documents (OCR, remove formatting or XML tags)
3. document alignment: identify translation(s) of a given document
4. sentence alignment: align paragraph and sentence for each parallel document pair

⇒ a lot of hand-crafted rules and heuristics except the last step



## Documents Alignment: difficulties

What to do with:

| | |
|---|---|
| 00_70221.isv | Swedish (sv) |
| 03_70821.pl | Polish (pl) |
| 03_70221.isf | Swedish and Finish (sf) |
| 03_70421.ien.ny_utg_970627 | English (en), new edition ("ny utgåva") |
| 02_60422.idenyutg.960626 | German (de), new edition |
| 08_80505.ity | German (ty = tyska) |
| 08_80505.iho | Dutch (ho = holländska) |
| 01_60123.pnl | Dutch (nl) |
| 12B_help_sd_sve.rtf | Swedish (sve=svenska) |
| 12B_help_sp_eng.rtf | English (eng) |
| om10aen.01 | English (en) |

parallel Scania corpus [Sang, 1996]



## Documents Alignment: difficulties

What to do with:

| | |
|---|---|
| 00_70221.isv | Swedish (sv) |
| 03_70821.pl | Polish (pl) |
| 03_70221.isf | Swedish and Finish (sf) |
| 03_70421.ien.ny_utg_970627 | English (en), new edition ("ny utgåva") |
| 02_60422.idenyutg.960626 | German (de), new edition |
| 08_80505.ity | German (ty = tyska) |
| 08_80505.iho | Dutch (ho = holländska) |
| 01_60123.pnl | Dutch (nl) |
| 12B_help_sd_sve.rtf | Swedish (sve=svenska) |
| 12B_help_sp_eng.rtf | English (eng) |
| om10aen.01 | English (en) |

parallel Scania corpus [Sang, 1996]

↪ basic assumption: similar names ⇒ translations

↪ in a perfect world: same filename ⊕ ISO-639 language code



## Document Alignment: methods

**Name mapping**

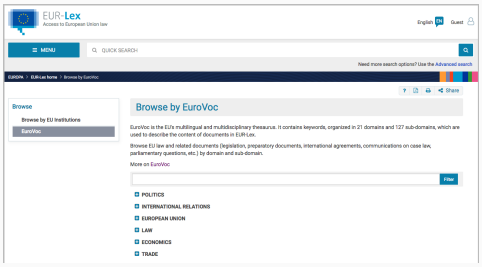
- handcrafted rules based on filenames
- tedious ⊕ source of errors

**Content-Based similarity [Patry and Langlais, 2005]**

- documents represented by a set of basic features (numbers, named entities, punctuations) ⊕ cosine similarity
- alignment score do we find the 'translation' of source words in the target documents?
- structure similarity
- looking for *hapax*
- using multilingual thesaurus (e.g. EuroVoc)

⇒ classifier

## EuroVoc






## Mining the Web

Same problem & solution



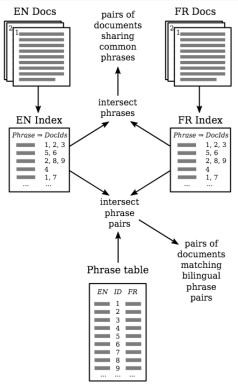
## A solved task?

- WMT'16 shared task on document alignment [Buck and Koehn, 2016]
- 13 participants (mainly academic)
- evaluation: % of the aligned pages in the test set that are found (recall)
- results:

| | Recall (%) |
|---|---|
| baseline (url matching) | 59.8% |
| best system | 95.0% |
| average (all non-bogus systems) | 85.3% |



## The best system [Gomes and Pereira Lopes, 2016]




- use a phrase table
  - ↪ by-product of the training of a phrase-based MT system
  - ↪ translation of *n*-grams
- intuition: parallel documents share more bilingual phrase pairs than non-parallel documents
- bootstrap / chicken-egg problem: only possible for resource-rich languages



## A case study: the JW300 corpus

## Context

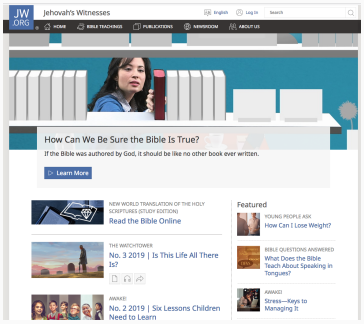
**Why this corpus?**

- very recent work (ACL'19)
- good illustration
- important resource

**What is it?**

- Parallel corpus
- 300 languages
- $\simeq$ 100,000 sentences per language
- 54 376 pair of languages

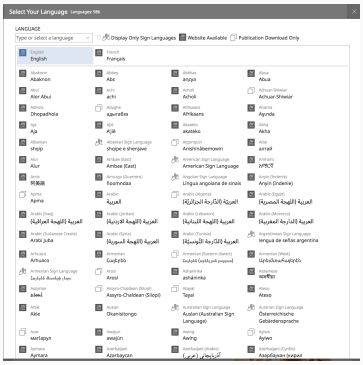
## The documents : jw.org




- "texts stem from a religious society"
- cover a "immense" range of topic
- mainly translations from English



## The documents : jw.org




- "texts stem from a religious society"
- cover a "immense" range of topic
- mainly translations from English

17

## Parallel corpus creation

### Data collection

- observation: articles carry unique identifiers...
- ... identifiers span across language

### Curation

- conversion HTML → plain text
- `Polyglot` sentence splitting & tokenization
  - ↪ rule based approach
  - ↪ use rules of the closest language
- sentence alignment