Multilingual and Parallel Corpora

Corpus basics

Amir Zeldes

amir.zeldes@georgetown.edu

Summer School: Researching translanguaging

- Translation and Translanguaging: Investigating Linguistic and Cultural Transformations in Superdiverse Wards in Four UK Cities
- 5 day summer school in Birmingham, 19th – 23rd June 2017
- Application deadline: **Feb 10**
  - http://www.birmingham.ac.uk/generic/tlang/events/index.aspx

What do we get from studies like these?

- Could we have come to these conclusions without corpus data?
- What is the alternative?
- When is it sufficient?

Corpora as empirical data

- Corpora give **empirical data** with few a priori hypotheses or assumptions
- The main alternative to a corpus is **introspection:**
  - Rationalist approach
  - Ability to describe intuitions
  - Give reasons
  - Negative claims

- There are also other types of empirical data:
  - Psycholinguistic experiments
  - Questionnaires
  - … (when do we need what?)

Intuition – exercise

- English declarative word order is generally Subject-Object-Verb
- Can you think of counter examples?
  - SOV?
  - VSO?
  - VOS?
  - VO?
  - …

An example

- English VO
  - The Welshman said Huck had good spots in him , and the widow said : " You can depend on it . That 's the Lord 's mark . He don 't leave it off . He never does . **Puts it** somewhere on every creature that comes from his hands . "
- How would this be translated into your second language?

An example

- English word order is SVO
  - The Welshman said Huck had good spots in him , and the widow said : " You can depend on it . That 's the Lord 's mark . He don 't leave it off . He never does . **Puts** it somewhere on every creature that comes from his hands . "
  - … Er trägt des Herren Mal an sich . _Er_ wird ihn nie verlassen . Er tut's nie . **Er vergißt keine Kreatur , die von ihm stammt**
- Translation effect?

Another example

- What do English clefts translate in Spanish?
  - ***It was silence that*** *spoke for the lovers at that moment*
  - ***El silencio fue allí el que*** habló por los dos amantes

  (Don Quijote)

- Cleft <> Cleft

Another example

- How do you translate a cleft into your language?
  - It was X that Y…
- What sentences get translated as clefts?
- Translate this:
  - *It was they who terrified her*

Another example

- What do English clefts translate in Spanish?
  - ***It was they who*** *terrified her*
  - ***Ellos,*** *señor, la sobresaltaron, como has dicho (Don Quijote)*

- Consistent mapping?
  - Reasons?
  - Tendencies? (quantitative!)
  - Meaning?

Introspection – yes or no?

- Maybe jein (deu) / la‘am (ara)
- Advantages:
  - Always available
  - Easy to vary parameters
- Disadvantages:
  - Highly subjective
  - We only choose cases we happen to imagine
  - Lack of real / random contexts
  - No quantitative data

Next

- Crash course – corpus basics:
  - More about using corpora and **annotations**
  - Tokenization
  - Lemmatization
  - Part of speech tags
- Following topics:
  - Alignment and search in parallel corpora
  - Introduction to translation studies

Processing and searching in corpora

- Our results will depend on how we analyze ourr corpus:
  - **Digitization** (how does a translation of a paper novel become a computer file?)
  - Segmentation, a.k.a. **tokenization**
  - Labeling, a.k.a. **annotation**
- For parallel corpora in particular, issues arise in:
  - **Harmonization** of schemes – can impact comparability, in parallel and comparable data
  - **Alignment** (big topic, bi-text only)

Preprocessing

- Many steps between novel and searchable corpus
- All involve (linguistic!) decisions
- Many procedures are automatable, but also error prone (think about OCR!)
- General options:
  - Manual vs. automatic
  - Statistical / rule-based / hybrid
  - Heavy / light on linguistic knowledge

Tokenization

- Token
  - Smallest unit of analysis
  - Working definition: An orthographic unit surrounded by white-space or punctuation
  - ≈ graphemic word

An example

- We separate all punctuation characters from adjacent words

  "I happen to think it was right," he said flatly.

```
"   I     happen    to    think   it   was right
,   "     he        said  flatly  .
```

  - 10 "words" – but if we separate those, material is left over
  - 14 tokens – minimal units of analysis
  - impossible to represent words and retain text otherwise

Tokenization guidelines

- Are space-delimited words really what we want to work with?
- Some problematic examples:
  - Number orthography: *20 000* or *(202) 687.5956*
  - Complex names, abbreviations: New York, U.S., etc. /et cetera, ...
  - Contractions: it's, I'll, we've
- How can we deal with these?
  - Human-readable guidelines – linguistic decisions are made here!
  - Automatic preprocessing: lists, regular expressions (more later)

Tokenization – problems

- Often hard to compare across languages
- German phrasal verbs (parliament proceedings – anbauen ~ cultivate, grows)
  - *Dass man in der Landwirtschaft auf die Freigabe des Anbaus wartet und Hanf **anbauen** will ,*
  - *Wird Hanf auf stillgelegten Flächen **angebaut** , gibt es anstelle der Beihilfe …*
  - *George Washington , liebe Kolleginnen und Kollegen , und Thomas Jefferson **bauten** auf ihren Plantagen Cannabis **an** .*
  - ➢ syntaktische Analyse nötig

Tokenization

- Sometimes we have multiple ‘words’ in tokens
  - **German:** ***beim, zum, siehste***
  - **French:** ***du, au***
  - **Spanish:** ***ocultarlo***
  - …
  - → Lists, regular expressions, heuristics
- Comparability across languages?
  - **Russian:** reflexives ‘fused’ ***мыться*** – wash oneself
  - **Polish:** reflexives separate ***myć się*** – wash oneself

Consequences?

Sentence segmentation

- Automatically disambiguating '.' (period) non-trivial, but important for later steps:
  - Sentence end (potential alignment region)
  - Abbreviation (does the period belong to the token? What happens if it's sentence final?)
    Ev. Elisabeth-Krankenhaus, etc., George W. Bush
  - in numbers
    Eng. 1.50, 1,000, 12:30 but *Deu. 5. Versuch, 3.100 Teilnehmer, 7.00 Uhr Ortszeit*
  - Word Alignment can be tricky:
    Deu. E.coli-Bakterien : Eng. E. coli bacteria

Tokenization - conclusion

- Word semgentation is non-trivial
- Decisions and errors happen already at this level
- Consequences for all following steps
- Language specific knowledge is required
- Sometimes it’s better to have a consistent ‘dumb’ guideline, rather than a clever but potentially inconsistent one

Part of speech tagging

- Tagging means adding
  (in principle arbitrary) linguistic categories to
  (in principle arbitrary) textual units

- Often tagging is short for ‘part-of-speech tagging’
  (POS tagging)

Traditional parts of speech

- Go back in Western thought to Dionysius Thrax (170-90 BCE)
- Credited with writing the *τέχνη γραμματική*, the Greek *"Art of Grammar"* (though earlier grammars exist, e.g. Pāṇini's)
- Basic description of 8 POS categories for Greek:
  - *Noun*
  - *Verb*
  - *Participle*
  - *Article*
  - *Pronoun*
  - *Preposition*
  - *Adverb*
  - *Conjunction*

  (but not adjective, interjection...)

Τοῦ δὲ λόγου μέρη ἐστὶν ***ὀκτώ*** *· ὄνομα, ῥῆμα, μετοχή, ἄρθρον, ἀντωνυμία, πρόθεσις, ἐπίρρημα, σύνδεσμος.*

Reading

- By next week, read:
  - Jakobson 1959: On Linguistic Aspects of Translation
  - Munday 2008 (30-34): Towards Contemporary Translation Theory

- We will finish corpus basics on Wednesday and start learning about translation theory from Monday