Multilingual and Parallel Corpora

Corpus basics (ctd.)

Amir Zeldes

amir.zeldes@georgetown.edu

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)

Tokenization - examples

```
<w id="5.1">Harry</w>
  <w id="5.2">Potter</w>
  <w id="5.3">!!!</w>
  <time id="T6E" value="00:01:25,480" />
  <time id="T7S" value="00:01:25,800" />
  <w id="5.4">now</w>
  <w id="5.5">,</w>
  <w id="5.6">you'</w>
  <w id="5.7">ve</w>
  <w id="5.8">done</w>
  <w id="5.9">it</w>
  <w id="5.10">!</w>
```

XCES

Tokenization - examples

Harry
Potter
!!!
now
,
you
've
done
it
!

TreeTagger

Chinese

- <s id="6">
  <time id="T4S" value="00:01:16,800" />
  <w id="6.1">你</w>
  <w id="6.2">說</w>
  <w id="6.3">如果</w>
  <w id="6.4">。</w>
  <w id="6.5">。</w>
  <w id="6.6">。</w>
  <time id="T4E" value="00:01:20,500" />
  </s>
  <s id="7">
  <time id="T5S" value="00:01:20,500" />
  <w id="7.1">哈</w>
  <w id="7.2">利</w>
  <w id="7.3">波特</w>
  <w id="7.4">！</w>
  <time id="T5E" value="00:01:25,500" />
  </s>

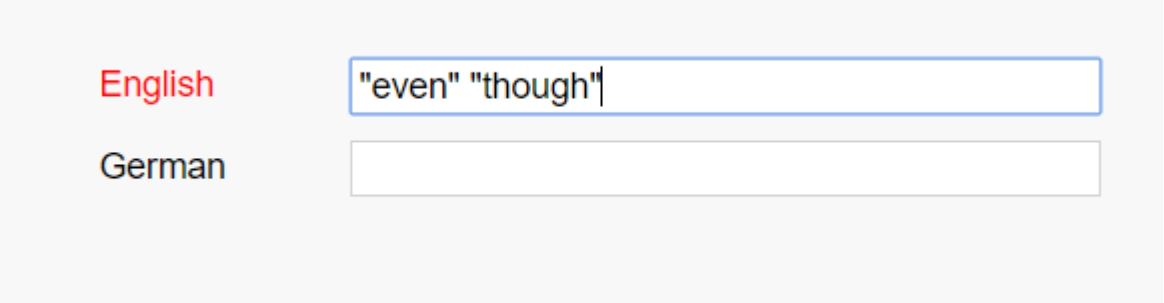
Japanese

- Four conventions (see Tanaka et al. 2016)
  - SWU – minimal morphemes
  - MWU – similar to English, inflections belong to words
  - LWU – no compound splitting, complex auxiliaries
  - Bunsetsu (also include post-positions, nominalizers..)

	魚フライを食べたかもしれないペルシャ猫 *"the Persian cat that may have eaten fried fish"*										
SUW	魚 NOUN *fish*	フライ NOUN *fry*	を ADP -ACC	食べ VERB *eat*	た AUX -PAST	か PART	も ADP	しれ VERB *know*	ない AUX -NEG	ペルシャ PROPN *Persia*	猫 NOUN *cat*
LUW	魚フライ NOUN *fried fish*		を ADP -ACC	食べ VERB *eat*	た AUX -PAST	かもしれない AUX *may*				ペルシャ猫 NOUN *Persian cat*	

Searching for multiple tokens

- Tokens are kept separate in each language:



- We'll see how to search with possible intervening tokens in a bit

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.*

Traditional parts of speech

- Perpetuated in Latin grammar, subsequent grammars of many languages
- Good starting point, especially for (Indo-)European languages
- But many things don't really fit in:
  - *watch out* (out =?? Adverb? "Particle"?)
  - *the man I met yesterday's dog* ('s=?? Genitive marker??)

Part of speech tagging

Part of speech (or **POS**) tagging is:

- Used as input for many NLP tasks
  - Syntactic parsing
  - Word sense disambiguation
  - Machine translation
  - ...
- Essential for distinguishing linguistic constructions:
  - *... if they bar/**VERB** the use of cellphones*
  - *... after which they went to the bar/**NOUN** , where*

Tag sets for English

Common in the US:

- Penn Treebank Tagset (PTB) — 36 Tags
- Extended PTB — 56 Tags

Common in the UK:

- CLAWS 5 — 62 Tags
- CLAWS 7 — 137 Tags

Older, notable mention:

- Brown tag set — 85 Tags

The PTB tag set (variations exist!)

CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person sg. present
VBZ	Verb, 3rd person sg. present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

Key variants for English

- Most English corpora use either PTB or **TreeTagger** tags:
  - VB – the verb be (and VBZ, VBN...)
  - VH – the verb have (VHZ, VHN...)
  - VV – any lexical verb (VVZ, VVN...)
  - PP for PRP personal pronoun (they...)
  - PP$ for PRP$ possessive pronoun (my...)
  - NP for NNP proper noun (Diane...)

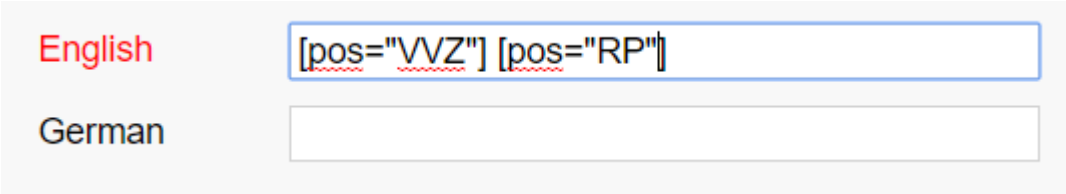
This happens for every language…

German tagsets (see Rapp & Lezius 2001):

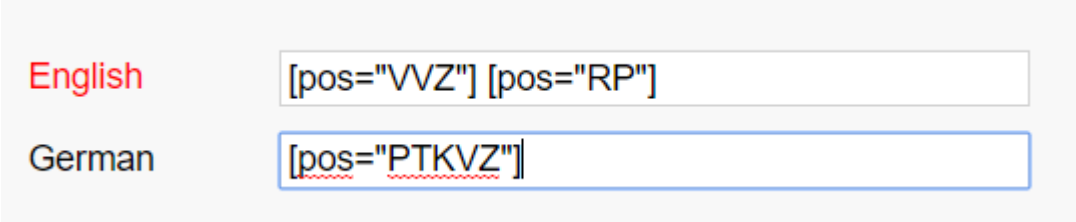
IBM Heidelberg	689	33
Münster	143	54
STTS (Stuttgart/Tübingen Tag Set)		**50**
ISSCO (Geneva)		56
Morphy (Paderborn)	500	52
…		

Searching for POS tags

- To search for pos tags use [pos="……"]

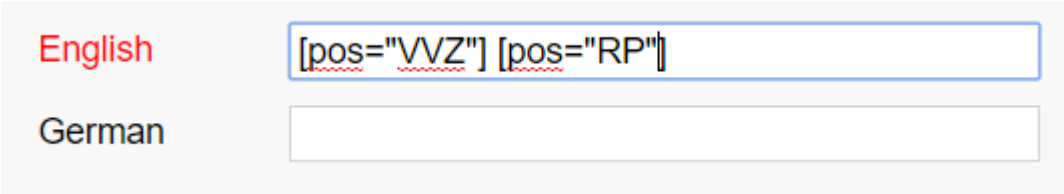


- You can combine searches across languages

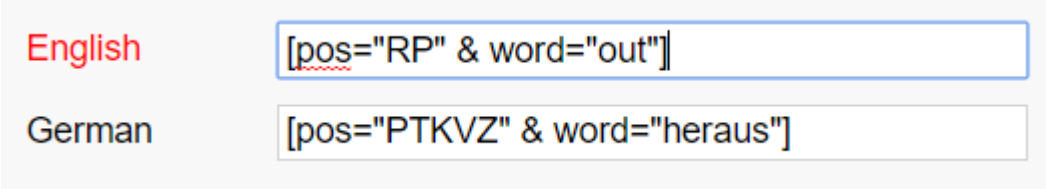


Searching for POS tags

- To combine POS and token searches, use:



- Similarly across languages:



midwife , with lilac ribbons in her cap , came **out** of Anna 's boudoir . She approached Karenin , and

Im Salon war niemand ; aus Annas Wohnzimmer kam auf das Geräusch seiner Schritte die Hebamme in einer Haube mit lila Bändern heraus .

Errors

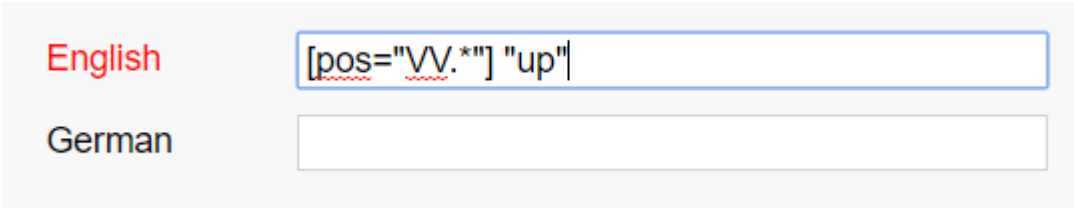
- Automatic taggers make mistakes!
- Very few parallel corpora are manually tagged
- Potential for ambiguity:
  - *like to **shop/VB***
  - *into his **shop/NN***
- Many words not in the lexicon (think of foreign words!)
  - *the **dancerliness/??** of that creation is experiential*
  - *had some alpha releases of **gnutls/??** called 0.2.x*
  - *I think this by **200:34:&%/??*** (what to do?)

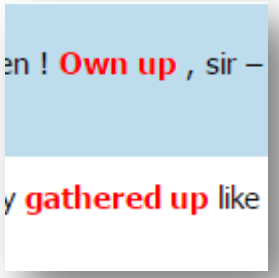
Errors

- Errors also vary by language:
  - *nach 13 Jahren Kohl/NN* [Mannheimer Korpus, TreeTagger]
    - *After 13 years of Kohl (~cabbage)*
  - *ich schmiede mir ne/FM schicke rubinklinge*
    - *I forge myself a (~ French not) nice ruby blade*
  - *meine neue waffe/FM*
    - *my new weapon (lower case!)*
      [www.worldofgothic.de, TreeTagger]

Using wildcards

- More complex searches use wildcards – **regular expressions**
- Possible to avoid some errors (at the cost of more results)
- Any lexical verb form:



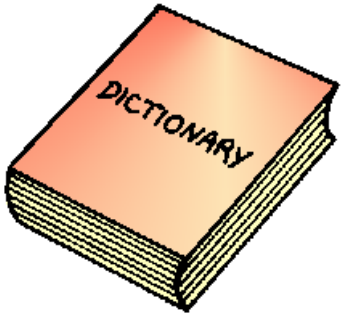


Lemmatization and stemming

- For many purposes we would like to ignore inflection:
  - *elected, elect, elects, electing*...
  - ➢ Investigate semantics of a term
  - ➢ Inflections can be very elaborate in many languages

- There are two different approaches to abstracting away from inflection:
  - Lemmatization – recover "dictionary entry" of word
  - Stemming – find lexical stem (may not be a word at all)

Lemmatization



- The **dictionary entry** of a word, carries no inflectional morphology:
  - Remove tense and person suffixes: *walked, been, sees*
  - Remove plural and possessive endings: *trees, America's*
    (but: if *'s* isn't already a separate token – it is in PTB)
  - In some cases, the entry has a completely different form (suppletion): *is* → *be*, *better* → *good*

Lemmatization

- However: many **derivationally** related words have different dictionary entries:
  - *elects, electing, elected…* → *elect*
  - *election, elections* → *election*
- Derivational morphology is not stripped in lemmatization
- Another way of thinking about it: lemmatization does not change the part-of-speech (*elect*: verb, *election*: noun)

Lemmatization

- Lemmatization conventions are language specific!
  - How do your languages lemmatize?
    - Basic form chosen for each category
    - What are the regular categories?
  - What are some problematic/irregular areas?
  - How do they affect language comparison?

Stemming

- Still, in some cases we want to know that something is about *election-related things*
- All we care about is the stem *elect-*
- All of the following lemmas have the same **stem**
  - *lemma="**election**" – forms: elections, election*
  - *lemma="**elect**" – forms: elects, elected, electing*
  - *lemma="**electoral**" – forms: electoral*
  - ...
- What exactly is covered by stemming depends on the guidelines we define (or tools we use)

Lemmatization and stemming

- Lemmatization and stemming are primarily useful for
  - lexicography
  - sentiment analysis
  - topic modelling
  - sociolinguistic research
  - machine translation (lemma/stem if word unknown)
  - ...
- Not as central for work on English as they are for work on strongly inflecting languages (e.g. Slavic, Finnish…)

Searching for lemmas

- Not all corpora are lemmatized ☹
  - In cases like these we need to use clever wild card searches
  - We do not have time for an extensive introduction to regular expressions…
  - But we can cover the basics
  - More information at:
    - http://www.regular-expressions.info/

The most useful **operators**

- Operators are special symbols in regular expressions:

`.`	(dot) – any character: `d.g` matches *dog, dig, dug*
*	(star) – previous letter any number of times: `of*` matches *o, of, off, offf, offffff ….*
+	(plus) – previous letter at least once: `of+` matches *of, off, offf, offffff ….*

- *Bonus question:* why did this work? `[pos="VV.*"]`

The most useful **operators**

- The ? operator makes the previous character optional:

```
[word="honou?r"]
```

- How would you find tomorrow spelled with and without a hyphen? (lemma doesn't cover this!)

```
[word="to-?morrow"]
```

➢ Caution: search is case sensitive! We won't find "Tomorrow"

Disjunction – *a* or *b*: (a|b)

- We may want to search for two alternative forms:

```
[lemma="(lady|gentleman)"]
[word="(slew|slayed)"]
```

- This also works **within** a pattern:

```
[word="be(tter|st)"]
```

- What is the shortest expression to find present and past participles of the verb *be*?

```
[word="be(ing|en)"]
```

Much more…

- There are other operators and tricks to find complex patterns
- We will learn a few more as we go along
- There is also a cheat sheet here:
  - https://corpling.uis.georgetown.edu/cqp/doc/CQPweb_query_syntax.pdf

Summary

- Understanding annotations is crucial for all subsequent work with a corpus
- Decisions are made at every step:
  - Tokenization
  - Sentence segmentation
  - POS tagging
  - Lemmatization/stemming
- Each step affects the next, and potentially introduces **errors!**

Reading

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

- We will start with a discussion of these texts and dive into translation theory for a bit!