CZECH NATIONAL

CORPUS

CZECH NATIONAL
CORPUS

Introduction to Text Corpora and Their Applications

Corpus characteristics and design

Lucie Chlumská, Ph.D.

lucie.chlumska@korpus.cz

OUTLINE:

1. LECTURE

- compiling a corpus
- tokenization, segmantation, lemmatization & tagging
- various types of annotation: morphological, syntactic, semantic
- types of corpora

2. SEMINAR

- reading (Biber et al.): corpus annotation
- what types of annotation are there...? what are the pros and cons?

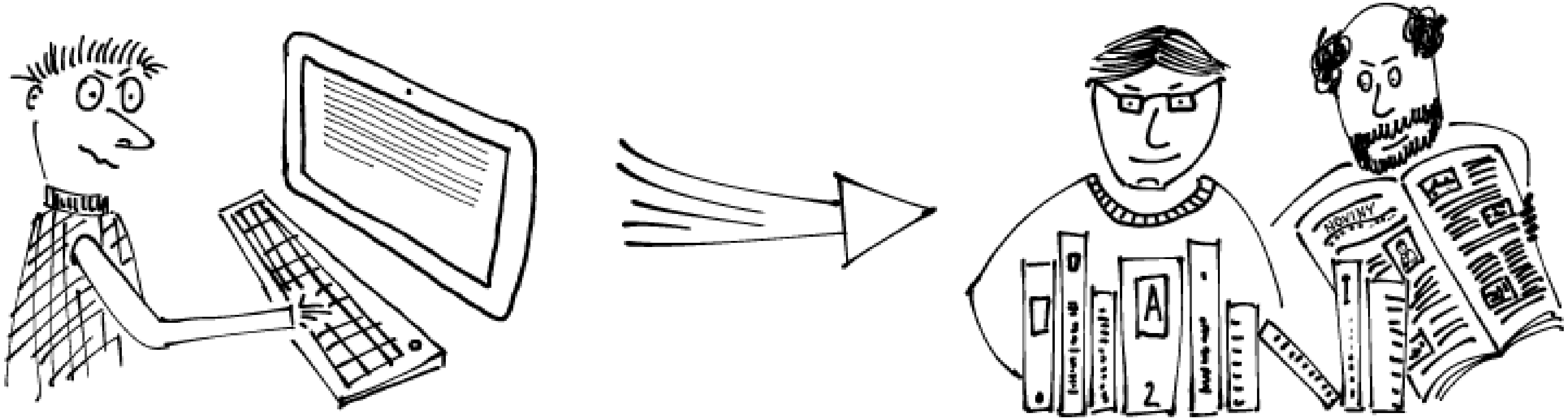
LECTURE

Compiling a corpus in a nutshell

1. Getting a text

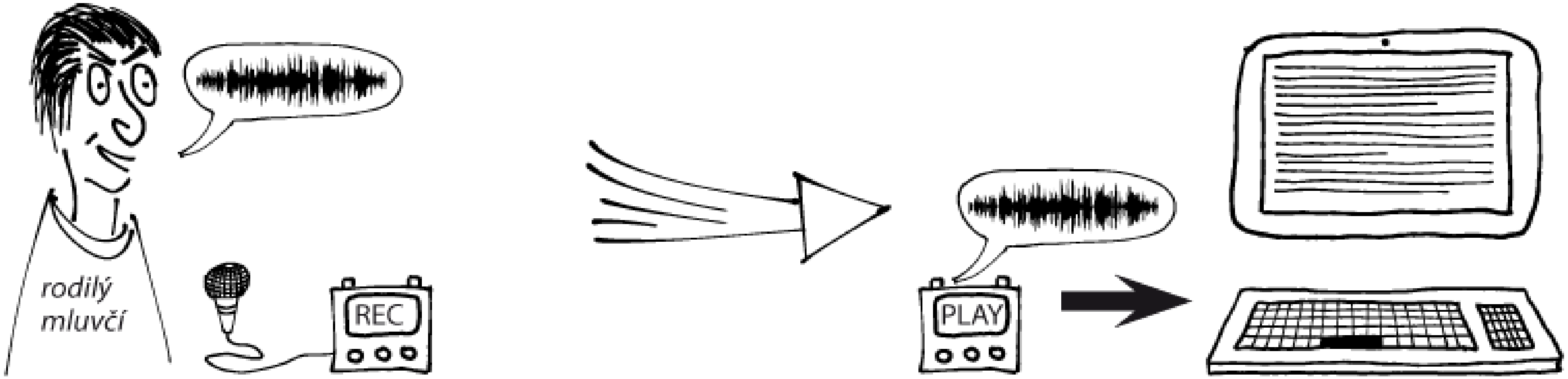
text authors

publishers

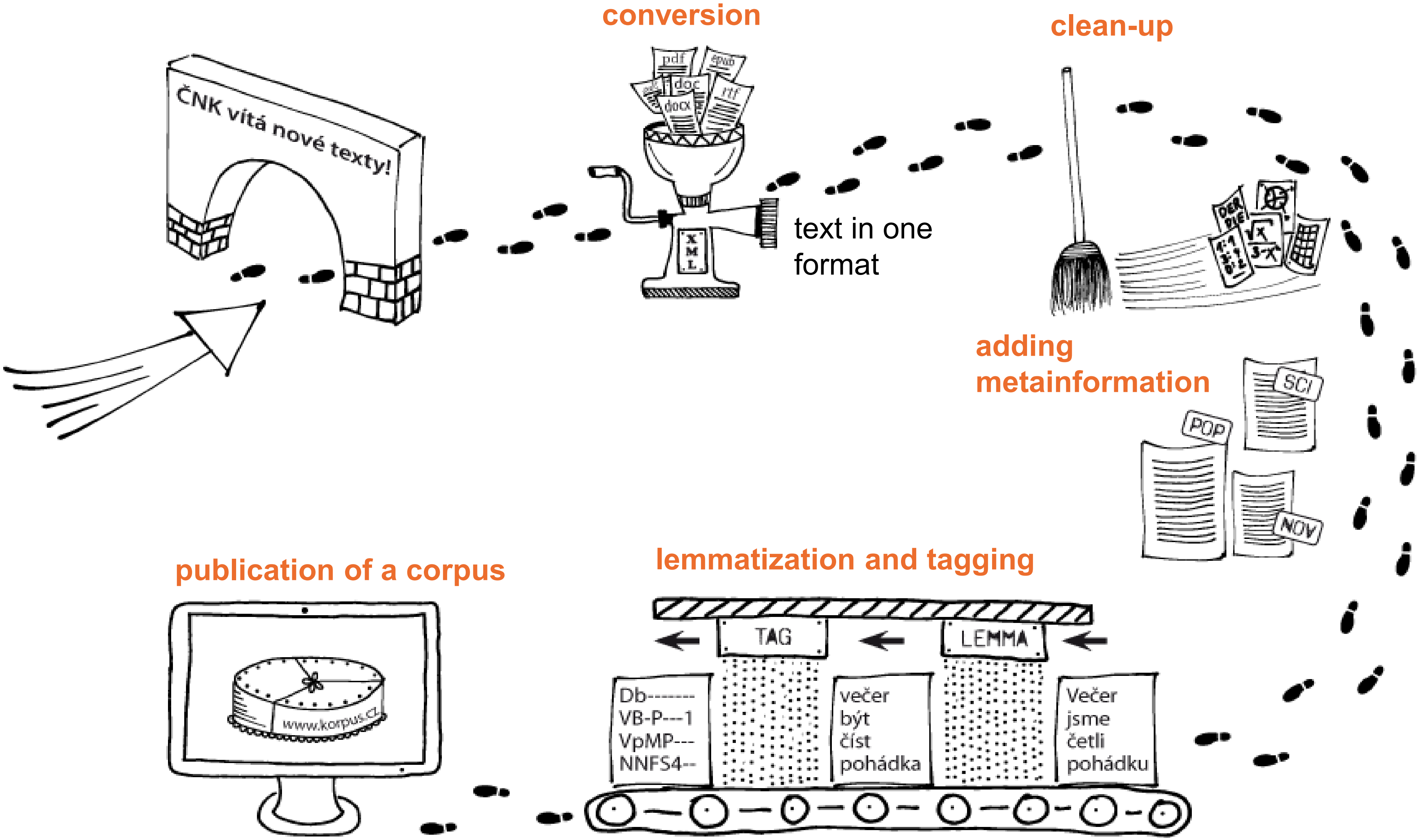


spoken language recording

transcription of recordings



2. Text processing (CNC)



conversion

text in one format

clean-up

adding metainformation

lemmatization and tagging

publication of a corpus

Text processing

What happens to the raw text?

- conversion to SGML/XML format
- tokenization: divides text into words, i.e. usually strings of characters surrounded by spaces (issue: *can’t* etc.)
- segmentation: end of sentences recognition (issue: abbreviations)

  <s> *This is a simple clause*. </s>

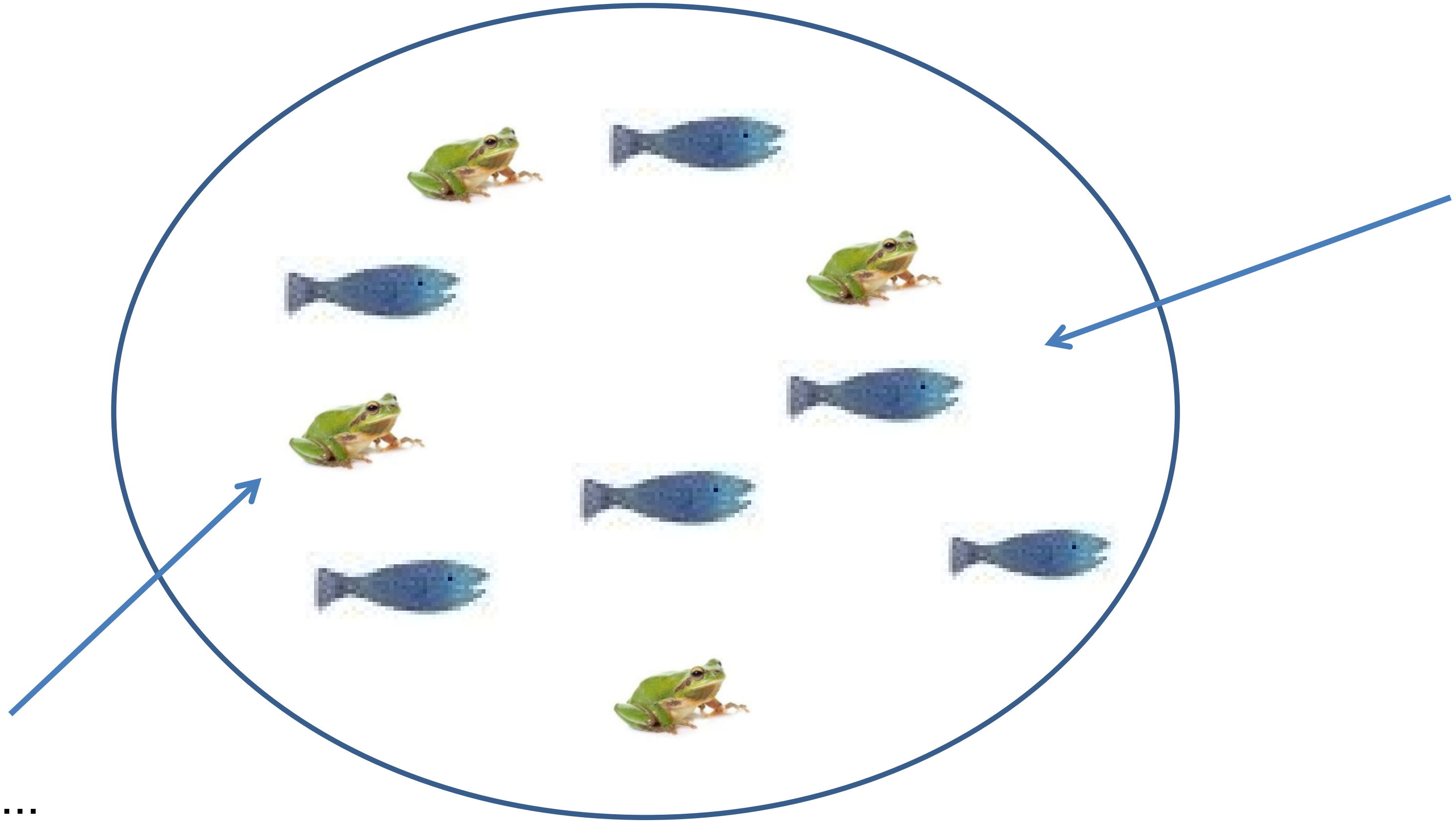
- morphological analysis (tagging and lemmatization)

  1) assigning all possible interpretations to the word

  2) disambiguation > stochastic (statistic) or rule-based

Precision and recall or efficient fishing

our goal: to catch all the fish and no frogs

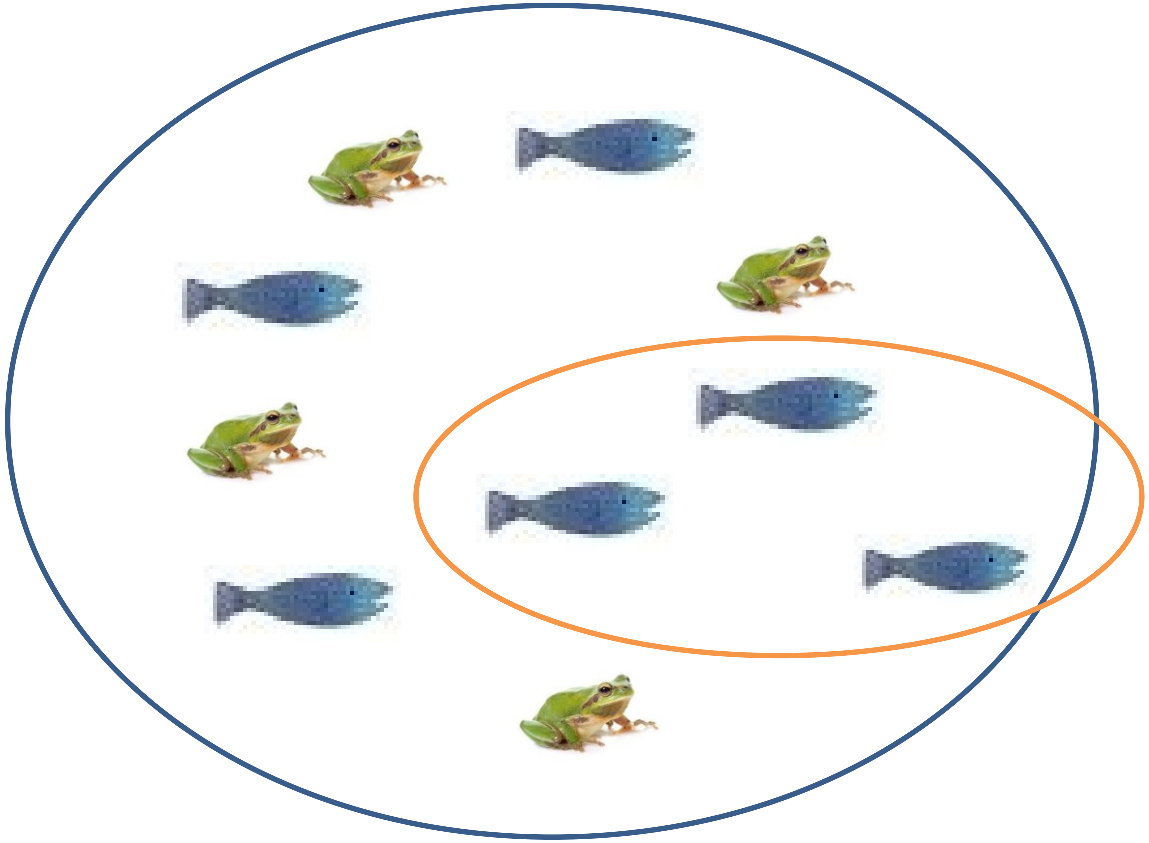


fish ☺

frogs...

Precision and recall

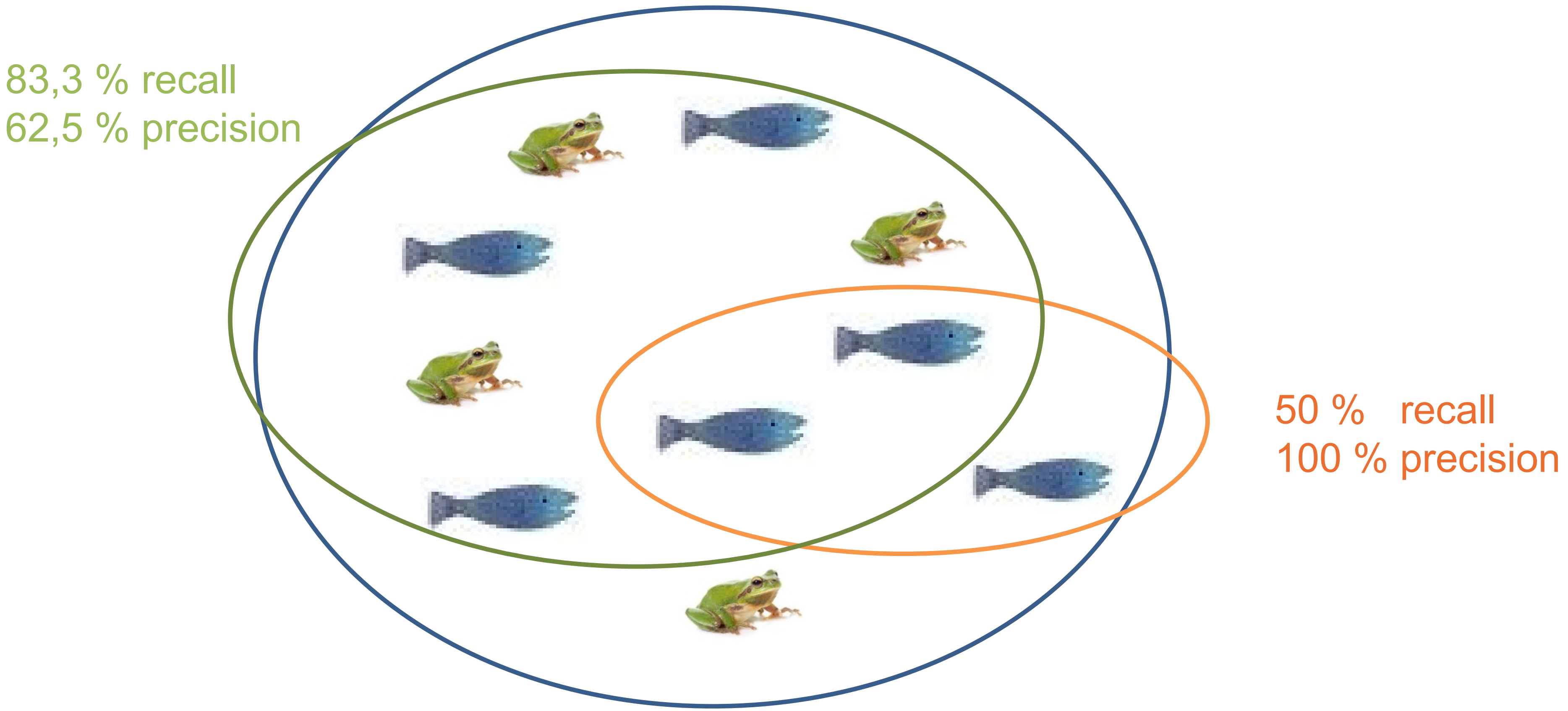
100 % recall: to find all the fish (plus some frogs)
100 % precision: to find only the fish (and no frogs)



50 % recall
100 % precision

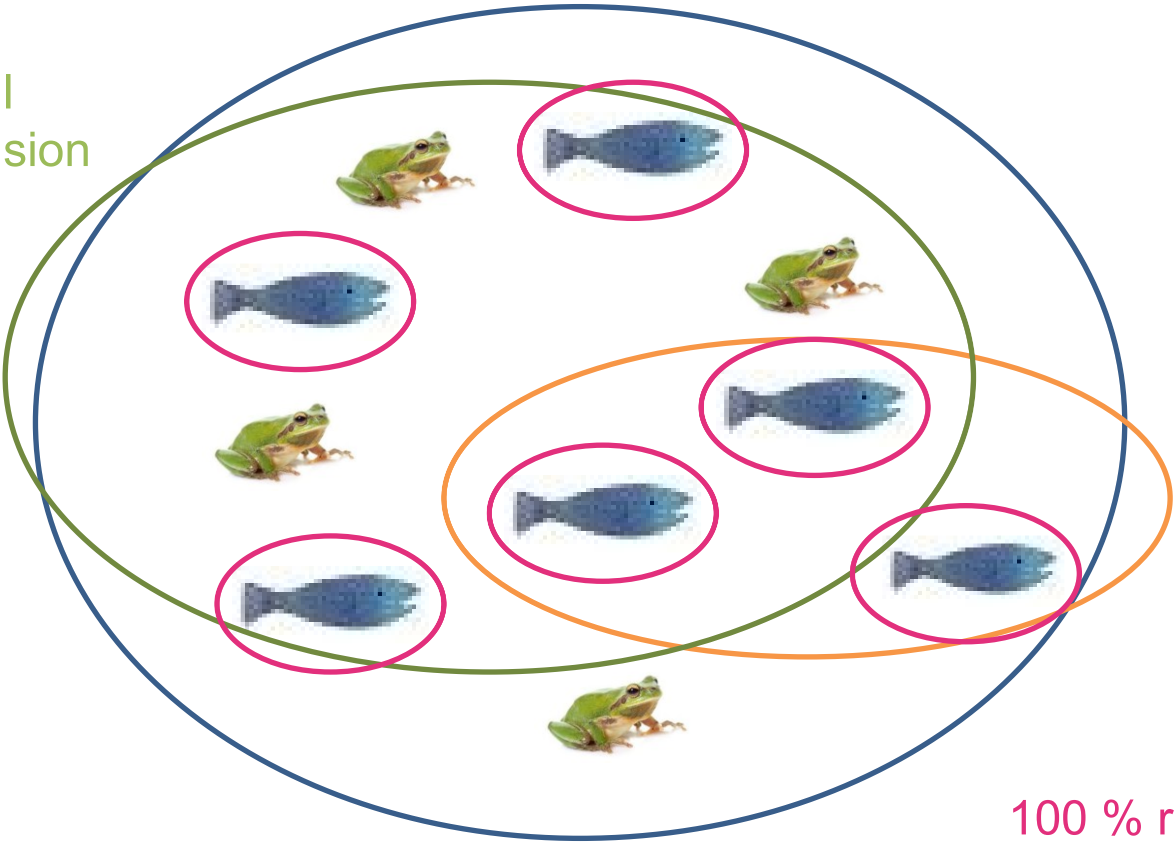
Precision and recall

100 % recall: to find all the fish (plus some frogs)
100 % precision: to find only the fish (and no frogs)



Precision and recall

100 % recall: to find all the fish (plus some frogs)
100 % precision: to find only the fish (and no frogs)



83,3 % recall
62,5 % precision

50 % recall
100 % precision

100 % recall
100 % precision

Precision and recall in annotation

- after the mofphological analysis > 100 % recall, but low precision

EN: *love* – noun, verb, adjective (*love affair*)?

CS: *jí* – pronoun, verb (+ all the flective characteristics)

- after the disambiguation, the precision gets higher
  - based on rules, context and language typology...

try Czech tagging: http://utkl.ff.cuni.cz/desamb-1/#

try English tagging: http://ucrel.lancs.ac.uk/claws/trial.html

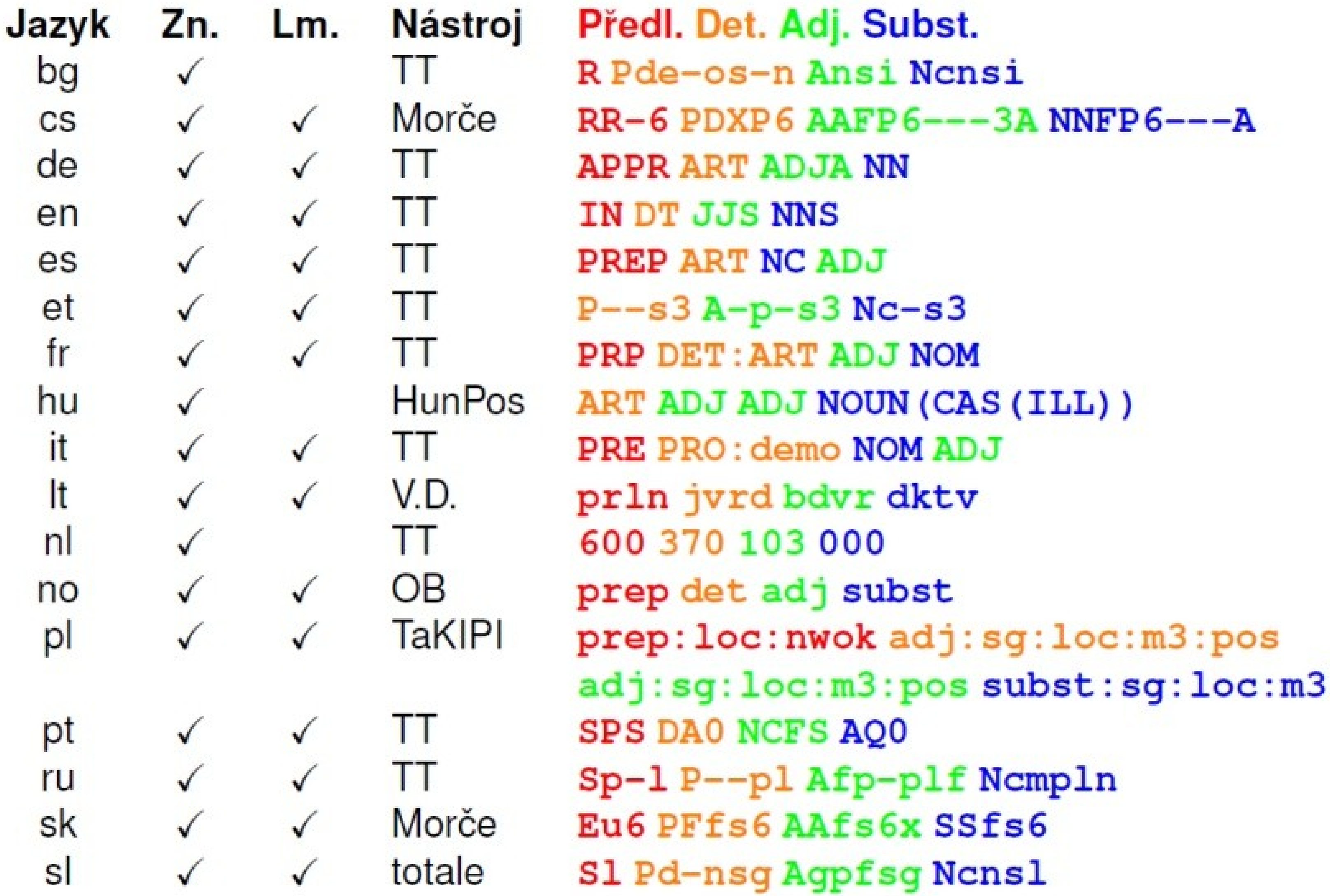
Lemmatization

- each word in corpus is assigned a lemma = basic form, headword
- especially useful for flective languages
- average Czech word has 13 different forms (due to the inflection)

- lemmatization issues:
  - CZ: *nemluvně* > *mluvně*, *Ho-Či-Min* (all Czech words), *česko-polský*, *jak* – pronoun or animal?
  - EN: homonymous *lie*, *bark*, possesive ‘s

POS-tagging

- different languages > different tagsets!



Jazyk	Zn.	Lm.	Nástroj	Předl. Det. Adj. Subst.
bg	✓		TT	R Pde-os-n Ansi Ncnsi
cs	✓	✓	Morče	RR-6 PDXP6 AAFP6---3A NNFP6---A
de	✓	✓	TT	APPR ART ADJA NN
en	✓	✓	TT	IN DT JJS NNS
es	✓	✓	TT	PREP ART NC ADJ
et	✓	✓	TT	P--s3 A-p-s3 Nc-s3
fr	✓	✓	TT	PRP DET:ART ADJ NOM
hu	✓		HunPos	ART ADJ ADJ NOUN(CAS(ILL))
it	✓	✓	TT	PRE PRO:demo NOM ADJ
lt	✓	✓	V.D.	prln jvrd bdvr dktv
nl	✓		TT	600 370 103 000
no	✓	✓	OB	prep det adj subst
pl	✓	✓	TaKIPI	prep:loc:nwok adj:sg:loc:m3:pos adj:sg:loc:m3:pos subst:sg:loc:m3
pt	✓	✓	TT	SPS DA0 NCFS AQ0
ru	✓	✓	TT	Sp-l P--pl Afp-plf Ncmpln
sk	✓	✓	Morče	Eu6 PFfs6 AAfs6x SSfs6
sl	✓	✓	totale	Sl Pd-nsg Agpfsg Ncnsl

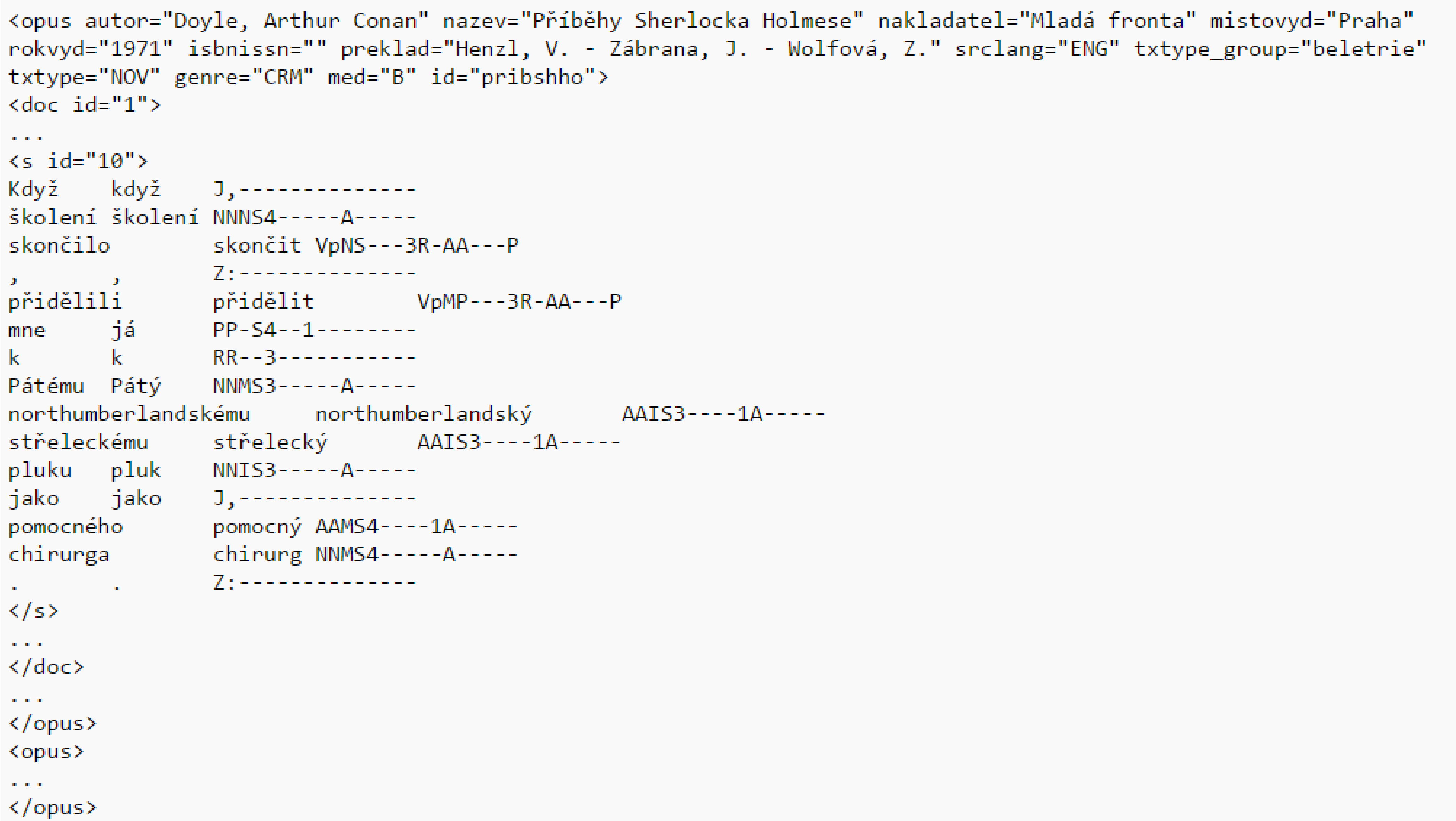
Czech v. English tags

- Czech morphological tag has currently 16 positions!

- English tag has generally 3 positions > BNC Basic (C5) Tagset
  - Each tag consists of three characters. Generally, the first two characters indicate the general part of speech, and the third character is used to indicate a subcategory. When the most general, unmarked category of a part of speech is indicated, in general the third character is 0.

E.g. AJ0 Adjective (general or positive) (e.g. good, old, beautiful)
AJC Comparative adjective (e.g. better, older)
AJS Superlative adjective (e.g. best, oldest)

„Naked“ corpus



```
<opus autor="Doyle, Arthur Conan" nazev="Příběhy Sherlocka Holmese" nakladatel="Mladá fronta" mistovyd="Praha"
rokvyd="1971" isbnissn="" preklad="Henzl, V. - Zábrana, J. - Wolfová, Z." srclang="ENG" txtype_group="beletrie"
txtype="NOV" genre="CRM" med="B" id="pribshho">
<doc id="1">
...
<s id="10">
Když    když    J,--------------
školení školení NNNS4-----A-----
skončilo        skončit VpNS---3R-AA---P
,       ,       Z:--------------
přidělili       přidělit        VpMP---3R-AA---P
mne     já      PP-S4--1--------
k       k       RR--3-----------
Pátému  Pátý    NNMS3-----A-----
northumberlandskému     northumberlandský       AAIS3----1A-----
střeleckému     střelecký       AAIS3----1A-----
pluku   pluk    NNIS3-----A-----
jako    jako    J,--------------
pomocného       pomocný AAMS4----1A-----
chirurga        chirurg NNMS4-----A-----
.       .       Z:--------------
</s>
...
</doc>
...
</opus>
<opus>
...
</opus>
```

Types of annotation and corpora

Types of corpora

time: synchronic v. diachronic v. monitor

register: spoken v. written v. multimodal

aim: representative v. specialized

language: monolingual v. bilingual v. multilingual

alignment: monolingual v. paralell

other: learner, acquisition...

Types of corpora and annotation

Presented by prof. McEnery

https://www.futurelearn.com/courses/corpus-linguistics/4/steps/69566

https://www.futurelearn.com/courses/corpus-linguistics/4/steps/69567

Let's take five now and then talk language!

CZECH NATIONAL
CORPUS

SEMINAR

Reading

- common reading:

McEnery, T., Xiao R. & Tono, Y. (2007). Corpus Annotation. In T. McEnery, R. Xiao & Y. Tono, *Corpus-Based Language Studies, an advanced resource book,* pp 30-45. NY: Routledge.

- another possible resources:

http://ucrel.lancs.ac.uk/annotation.html
http://ucrel.lancs.ac.uk/claws/
http://ucrel.lancs.ac.uk/usas/

Discussion

- What is a word? What are the segmentation problems?
- Are there any disadvantages of corpus annotation?
- What are the main issues in annotating corpora?
- How can the annotation influence the analysis?
- Why is the semantic tagging so rare?
- Why is error tagging useful?

...any other ideas, comments, suggestions?