Multilingual Natural Language Processing: Week 5

Daniel Zeman, Rudolf Rosa, Ondřej Bojar
zeman@ufal.mff.cuni.cz
http://ufal.mff.cuni.cz/courses/npfl120
Part-of-Speech Tagset Conversion

- See also NPFL094 (Computational Morphology and Syntax) in Winter
  - There: focus on linguistic diversity
  - Here: focus on
    - Technical aspects
    - Different expressivity
    - Different granularity
Why Convert Tags?

- For a tool that uses tags (parser)
  - The meaning of the tags is significant (they are not just strings)
  - Or the tool has been trained on a particular tagset

- For a linguist who works with corpora
  - Reduce need to learn new tags
How to Convert Tags?

- Look at source tags only
How to Convert Tags?

- Look at source tags only
  - Conversion tailored to a pair of tagsets
How to Convert Tags?

- Look at source tags only
  - Conversion tailored to a pair of tagsets
  - Reusable “interlingua” (Interset, Universal Dependencies)
How to Convert Tags?

- Look at source tags only
  - Conversion tailored to a pair of tagsets
  - Reusable “interlingua” (Interset, Universal Dependencies)

- Look at source tags + words
How to Convert Tags?

- Look at source tags only
  - Conversion tailored to a pair of tagsets
  - Reusable “interlingua” (*Interset, Universal Dependencies*)

- Look at source tags + words

- Look at source tags + words + context
Related Work

- EAGLES, PAROLE, MULTEXT
  - Rather wanted to standardize tags
  - Not to work with the tags that are already there
  - Very euro-centric
Related Work

- **EAGLES, PAROLE, MULTEXT**
  - Rather wanted to standardize tags
  - Not to work with the tags that are already there
  - Very euro-centric

- **IIIT Hyderabad: all Indian languages**
  - Indo-Aryan
  - Dravidian
  - English!
Related Work

- EAGLES, PAROLE, MULTEXT
  - Rather wanted to standardize tags
  - Not to work with the tags that are already there
  - Very euro-centric

- IIIT Hyderabad: all Indian languages
  - Indo-Aryan
  - Dravidian
  - English!

- Gold Ontology
  - Defines linguistic terms
  - The same term may denote different things in different languages
Related Work

- **EAGLES, PAROLE, MULTEXT**
  - Rather wanted to standardize tags
  - Not to work with the tags that are already there
  - Very euro-centric

- **IIIT Hyderabad: all Indian languages**
  - Indo-Aryan
  - Dravidian
  - English!

- **Gold Ontology**
  - Defines linguistic terms
  - The same term may denote different things in different languages

- **Interset, Google UPOS, Universal Dependencies**
Related Work

- EAGLES, PAROLE, MULTEXT
  - Rather wanted to standardize tags
  - Not to work with the tags that are already there
  - Very euro-centric
- IIIT Hyderabad: all Indian languages
  - Indo-Aryan
  - Dravidian
  - English!
- Gold Ontology
  - Defines linguistic terms
  - The same term may denote different things in different languages
- Interset, Google UPOS, Universal Dependencies
- Papers claiming that universal tagset does not exist
Josef následující jejímuž stě jsem nejméně v aby jen ejhle noor ,
Prague Tags for Czech

NNMS1------A----- NMS1A
AGFS3------A----- AVGFS3A
P12S3FS3------- P2FSZS3
C4XP3--------2 CGXP3-2
VB-S--1P-AA--- VPS1A
Dg--------3A----- DG3A
RR--6-------- R6
J,X--3-------- JVX3
TT------------- T
II------------- I
X@------------- NOMORPH
Z:------------- ZIP
Prague Tags for CoNLL 2006 Shared Task

NNMS1-----A----  N  N  Gen=M|Num=S|Cas=1...
AGFS3-------A----  A  G  Gen=F|Num=S|Cas=3...
P1ZS3FS3--------  P 1  Gen=Z|Num=S|Cas=3...
ClXP3--------------2  C 1  Gen=X|Num=P|Cas=3...
VB-S----1P-AA-----  V B  Num=S|Per=1|Ten=P...
Dg---------3A------  D g  Gra=3|Neg=A
RR--6----------  R R  Cas=6
J,-X---3--------  J ,  Num=X|Per=3
TT-----------------  T T  _
II---------------  I I  _
X@-----------------  X @  _
Z:-----------------  Z :  _
## Multext East

<table>
<thead>
<tr>
<th>NNMS1-------A----</th>
<th>Ncmsny</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGFS3-------A----</td>
<td>Afpfsd</td>
</tr>
<tr>
<td>P1ZS3FS3-------</td>
<td>Pr3mdsfnayn</td>
</tr>
<tr>
<td>ClXP3----------2</td>
<td>Mcmn3y</td>
</tr>
<tr>
<td>VB-S---1P-AA---</td>
<td>Vmiplssmanyn</td>
</tr>
<tr>
<td>Dg----------3A----</td>
<td>Rgs</td>
</tr>
<tr>
<td>RR--6----------</td>
<td>Sps1</td>
</tr>
<tr>
<td>J,-X---3-------</td>
<td>Css3</td>
</tr>
<tr>
<td>TT-------------</td>
<td>Q</td>
</tr>
<tr>
<td>II-------------</td>
<td>I</td>
</tr>
<tr>
<td>X@-------------</td>
<td>X</td>
</tr>
<tr>
<td>Z:-------------</td>
<td></td>
</tr>
</tbody>
</table>
Majka Tagset from Brno

NNMS1-------A----- k1gMnSc1eA
AGFS3-------A----- k2gFnSc3eA
P1ZS3FS3-------- k3gUnSc3p3hFxR
C1XP3----------2 k4gXnPc3xC
VB-S---1P-AA--- k5gXnSp1mIaIeA
Dg---------3A---- k6d3eAxD
RR--6---------- k7c6
J,-X---3------- k8p3xS
TT------------- k9
II------------- k0
X@-------------
Z:--------------
Penn Treebank Tags for English

CC  CD  DT  EX  FW  IN  JJ  JJR  JJS  LS  MD  NN  NNS  NNP  NNPS  PDT  POS  PRP  PRP$  RB  RBR  RBS  RP  SYM  TO  UH  VB  VBD  VBG  VBN  VBP  VBZ  WDT  WP  WP$  WRB  .  ,  :  $  #  `  '  –  _  –

- **EX** = existential *there*
- **FW** = foreign word
- **IN** = preposition or subordinating conjunction
- **TO** = *to*
- **UH** = interjection…
Brown Corpus Tags for English

SynTagRus Tags for Russian

S ЕД МУЖ ИМ NNMS1------A-----
S МН РОД ОД PSXXXXXP3--------
A МН ИМ AAXP1------1A-----
NUM ВИН C1XX4--------------
V НЕСОВ ИЗЪЯВ НЕПРОШ МН 3-Л VB-P---3P-AA---
ADV СПАВ Dg--------2A-----
PR RR--6------------
CONJ J-^---------------
PART TT--------------
INTJ II--------------
Stuttgart-Tübingen Tagset for German

ADJA ADJD ADV APPR APPRART APPO APZR ART CARD FM ITJ KOUI KOUS KON KOKOM NN NE PDS PDAT PIS PIAT PIDAT PPER PPOSS PPOSAT PRELS PRELAT PRF PWS PWAT PWAV PAV PTKZU PTKNEG PTKVZ PTKANT PTKA TRUNC VVFIN VVIMP VVINF VVIZU VVPP VAFIN VAIMP VAINF VAPP VMFIN VMINF VMPP XY $, $. $ (  

- Like in Penn TB: parts of speech only, but slightly more fine-grained
- No morphology (German has gender, number, case, degree, person…)
- “Substantive” vs. “attributive” pronouns (S vs. AT)  
- Adposition = Präposition, Postposition, Zirkumposition
Ambition: common tagset for all Indian languages (IE and Dravidian!)
No morphology (although the languages are rich on morphology)
- Hierarchical tagset, morphology can be added at the end
- And they “do not want to decrease tagging accuracy” (!)
Cloned from Penn tagset and modified
- New categories, e.g. postposition, “quotative”
- Removed traces of morphology, e.g. plural, comparative, superlative
Tagging is interwined with tokenization.

```xml
<token_Arabic>وبالفالوجة</token_Arabic>
  <voc>wabiAlfAlwjp</voc>
  <pos>wa/CONJ+bi/PREP+AlfAlwjp/NOUN_PROP</pos>
</token_Arabic>
<token_Arabic>مثال</token_Arabic>
  <voc>mivAlu</voc>
  <pos>mivAl/NOUN+u/CASE_DEF NOM</pos>
</token_Arabic>
```
ElixirFM (PADT) Arabic Tags by Ota Smrž

N--------1D  NNXX1------A-----
Z--------1-  NNXX1------A-----
A-------FP2D  AAFP2------1A-----
S-------3MP1- PPMP1--3---------
VIS------- VcXX---XP-AA----
Rocling / Sinica Tagset for Chinese

Na = common noun
Nb = proper noun
Nc = location noun
Nd = time noun
Nf = classifier
Nh = pronoun
Ne = determiner or cardinal number
Ng = postposition
P = preposition
P01 = 為 wèi, 承蒙 chéngméng, 深為 shēnwèi
P02 = 被 bèi
P03 = 為了 wèile, 為 wèi
P04 = 給 gěi
P06 = 由 yóu
P07 = 把 bǎ, 將 jiāng
...
P66 = 為 wèi
PAROLE Danish and Swedish

NCCPU==I ... *historikere*

NCNPU==D ... *Charta_77-folkene*

ANP (CN) PU= (DI) U ... *russiske*

AC---U==-- ... *5.000*

VADR=-----A- ... *har*

VAPR=(SP) (CN) (DI) A-U ... *gældende*

RGU ... *af*

PP3 (CN) (SP) U-YU ... *sig*

NCUPN@DS ... *konflikterna*

(substantiv utrum pluralis bestämd num

AQP0PN0S ... *politiska*

MC00G0S ... *fyras (gt. gen.)*

V@IPAS ... *har*

AP000N0S ... *oberoende*

RG0S ... *inte*

PF@000@S ... *sig*
MAMBA and PAROLE Tagsets for Swedish

NN … noun
PN … proper noun
VN … gerund
AJ … adjective
AV BV FV GV HV KV
MV QV SP SV VV WV … verbs
HV … the verb *hava*
I? IC IG IK IP IQ
IR IS IT IU … punctuation

NCUPN@DS … *konflikterna* (substantiv utrum pluralis bestämd nominativ)
AQP0PN0S … *politiska*
V@IPAS … *har*
AP000N0S … *oberoende*
RG0S … *inte*
PF@000@S … *sig*
<table>
<thead>
<tr>
<th>pos</th>
<th>noun</th>
<th>adj</th>
<th>num</th>
<th>verb</th>
<th>adv</th>
<th>prep</th>
<th>conj</th>
<th>part</th>
<th>int</th>
<th>punc</th>
</tr>
</thead>
<tbody>
<tr>
<td>subpos</td>
<td>prop</td>
<td>class</td>
<td>pdt</td>
<td>det</td>
<td>art</td>
<td>digit</td>
<td>roman</td>
<td>card</td>
<td>ord</td>
<td></td>
</tr>
<tr>
<td>prontype</td>
<td>prs</td>
<td>rcp</td>
<td>int</td>
<td>rel</td>
<td>dem</td>
<td>neg</td>
<td>ind</td>
<td>tot</td>
<td></td>
<td></td>
</tr>
<tr>
<td>punctype</td>
<td>peri</td>
<td>quest</td>
<td>excl</td>
<td>quot</td>
<td>brck</td>
<td>comm</td>
<td>colo</td>
<td>semi</td>
<td>dash</td>
<td>symb</td>
</tr>
<tr>
<td>puncside</td>
<td>ini</td>
<td>fin</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>synpos</td>
<td>subst</td>
<td>attr</td>
<td>adv</td>
<td>pred</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>poss</td>
<td>poss</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>reflex</td>
<td>reflex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>definiteness</td>
<td>ind</td>
<td>def</td>
<td>red</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gender</td>
<td>masc</td>
<td>fem</td>
<td>com</td>
<td>neut</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>animateness</td>
<td>anim</td>
<td>inan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>number</td>
<td>sing</td>
<td>dual</td>
<td>plu</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>case</td>
<td>nom</td>
<td>gen</td>
<td>dat</td>
<td>acc</td>
<td>voc</td>
<td>loc</td>
<td>ins</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>prepcase</td>
<td>npr</td>
<td>pre</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>degree</td>
<td>pos</td>
<td>com</td>
<td>sup</td>
<td>abs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>person</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>politeness</td>
<td>inf</td>
<td>pol</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>possgender</td>
<td>masc</td>
<td>fem</td>
<td>com</td>
<td>neut</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>posnumber</td>
<td>sing</td>
<td>dual</td>
<td>plu</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>subcat</td>
<td>intr</td>
<td>tran</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>verbform</td>
<td>fin</td>
<td>inf</td>
<td>sup</td>
<td>part</td>
<td>trans</td>
<td>ger</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mood</td>
<td>ind</td>
<td>imp</td>
<td>cond</td>
<td>sub</td>
<td>jus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tense</td>
<td>past</td>
<td>pres</td>
<td>fut</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>subtense</td>
<td>aor</td>
<td>imp</td>
<td>ppq</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>aspect</td>
<td>imp</td>
<td>perf</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>voice</td>
<td>act</td>
<td>pass</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>foreign</td>
<td>foreign</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>abbr</td>
<td>abbr</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hyph</td>
<td>hyph</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>style</td>
<td>arch</td>
<td>form</td>
<td>nom</td>
<td>coll</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>typo</td>
<td>typo</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>variant</td>
<td>short</td>
<td>long</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>tagset</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>other</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

{ obscure_feature_1 => [0, 7, 351.2, [a",",b"] ] }
Reference:

Multi-values (disjunction)

- A tag may say: gender = masculine OR neuter
- Interset allows multiple values of a feature
- Problem: multiple combinations of values
Multi-values (disjunction)

- A tag may say: gender = masculine OR neuter
- Interset allows multiple values of a feature

Problem: multiple combinations of values
  - EITHER (gender = feminine AND number = singular)
Multi-values (disjunction)

- A tag may say: gender = masculine OR neuter
- Interset allows multiple values of a feature

Problem: **multiple combinations of values**
  - EITHER (gender = feminine AND number = singular)
  - OR (gender = neuter AND number = plural)
A tag may say: gender = masculine OR neuter

Interset allows multiple values of a feature

Problem: multiple combinations of values
  - EITHER (gender = feminine AND number = singular)
  - OR (gender = neuter AND number = plural)
  - BUT NOT (feminine plural or neuter singular)

Interset cannot represent this (only two separate feature structures)
Tagset Drivers

- A (Perl) module with the following functions:
  - `decode()` ... converts a tag to Interset
  - `encode()` ... generates a tag from Interset
  - `list()` ... lists known tags in the tagset (optional)
Not Everything Fits in the Target Tagset

- Throw away information that cannot be represented
- Warning! May generate “unexpected” tag
  - Swedish knows: `noun, gender=com|neut`
Not Everything Fits in the Target Tagset

- Throw away information that cannot be represented
- Warning! May generate “unexpected” tag
  - Swedish knows: noun, gender=com|neut
  - and also: personal pronoun, gender=masc|fem|com|neut
Not Everything Fits in the Target Tagset

- Throw away information that cannot be represented
- Warning! May generate “unexpected” tag
  - Swedish knows: noun, gender=com|neut
  - and also: personal pronoun, gender=masc|fem|com|neut
  - From Czech: noun, gender=masc
Not Everything Fits in the Target Tagset

- Throw away information that cannot be represented
- Warning! May generate “unexpected” tag
  - Swedish knows: noun, gender=com|neut
  - and also: personal pronoun, gender=masc|fem|com|neut
  - From Czech: noun, gender=masc
  - Either change noun to pronoun
  - or change gender=masc to gender=com
Not Everything Fits in the Target Tagset

- Throw away information that cannot be represented
- Warning! May generate “unexpected” tag
  - Swedish knows: noun, gender=com|neut
  - and also: personal pronoun, gender=masc|fem|com|neut
  - From Czech: noun, gender=masc
  - Either change noun to pronoun
  - or change gender=masc to gender=com
  - What has higher priority?
Does It Matter?

- Atomic tagsets (Penn): no choice
- Positional tagsets can encode “impossible” combinations, e.g. a plural accusative adverb
- What is our goal?
Does It Matter?

- Atomic tagsets (Penn): no choice
- Positional tagsets can encode “impossible” combinations, e.g. a plural accusative adverb

What is our goal?

- Just querying attributes? ⇒ Preserve as much info as possible!
Does It Matter?

- Atomic tagsets (Penn): no choice
- Positional tagsets can encode “impossible” combinations, e.g. a plural accusative adverb

What is our goal?
- Just querying attributes? ⇒ Preserve as much info as possible!
- Use a pre-trained black-box tool? ⇒ Don’t give it data that it doesn’t expect!
Enforcing Defaults

- Need the list of known target tags

- Centrally for all tagsets:
  - Priorities of features
  - For every feature value, ordered list of substitutes
    - Typically, empty value is the best substitute
    - But: number = dual is better substituted by plural!

```

[ 
  [ "null" ],
  "number" => 
  [ 
    [ "sing" ],
    [ "dual", "plu" ],
    [ "plu" ]
  ],
  "possnumber" => 

0 → sing, dual, plur; sing → 0, dual, plur
```
Enforcing Defaults

- Decode all known target tags
- Construct trie for known feature-value combinations
- Follow path in trie when encoding
- If a value is not allowed, find the best substitute

(It is more complex when multi-values come into play.)
Substitution Trie

Seminář ÚFAL, Homí Míšecky, 9.2.2009
Google Universal Part-of-Speech Tags

Google Universal Part-of-Speech Tags

- Just the POS category. No morphology
- For many tools this is enough
Google Universal Part-of-Speech Tags

- Just the POS category. No morphology
- For many tools this is enough
- Good idea
- But it must be applied well!
Google Universal Part-of-Speech Tags

- Just the POS category. No morphology
- For many tools this is enough
- Good idea
- But it must be applied well!

- **pronoun** → **PRON**
  - determiners, numerals, adverbs
Google Universal Part-of-Speech Tags

- Just the POS category. No morphology
- For many tools this is enough
- Good idea
- But it must be applied well!

- **pronoun** → **PRON**
  - determiners, numerals, adverbs
- similar for numerals in Danish
- similar for nominal/adjectival verb forms
Lemma-based Re-tagging

```perl
my $lemma = $node->lemma();
# Fix Interset features of pronominal words.
if ($node->is_pronominal())
{
    # Indefinite pronouns and determiners cannot be distinguished by
    if ($lemma =~ m/^((ně|lec|ledas?|kde|bůhví|kdoví|nevím|málo|sotva)|^\$)/)
    {
        $node->iset()->set('pos', 'noun');
    }
    elsif ($lemma =~ m/^(jaký|který)|(jaký|který)$|^\(každý|všechen|sá)/)
    {
        $node->iset()->set('pos', 'adj');
    }
    # Pronouns čí, něčí, čísí, číkoli, ledačí, kdečí, bůhvíčí, nevímčí
    elsif ($lemma =~ m/^((ně|lec|ledas?|kde|bůhví|kdoví|nevím|ni)?čí|čí)/)
    {
        $node->iset()->set('pos', 'adj');
    }
    $node->iset()->set('poss', 'poss');
}
```

# Pronoun (determiner) "sám" is difficult to classify in the trad
Universal Dependencies: UPOS and Features

- UPOS = extended version of Google universal tags

- Features = extended Interset
  - (now it is the target representation rather than something intermediate)
  - “Universal” feature + set of values
  - Language-specific value of universal feature
  - Language-specific (or treebank-specific) feature + set of values
A Grain of Salt: Even UD Can Be Used Inconsistently!

- https://lindat.mff.cuni.cz/services/pmltq/
  - Find two UD treebanks of related languages
  - Where the “same word” does not get the same UPOS category
A Grain of Salt: Even UD Can Be Used Inconsistently!

- https://lindat.mff.cuni.cz/services/pmltq/
  - Find two UD treebanks of related languages
  - Where the “same word” does not get the same UPOS category