Interset: Reusable Tagset Conversion

Daniel Zeman, Rudolf Rosa

March 20, 2020
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

Interset: Reusable Tagset Conversion
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
P1ZS3FS3------ PSEFSZS3
ClXP3--------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

<table>
<thead>
<tr>
<th>Tag</th>
<th>Gen</th>
<th>Num</th>
<th>Per</th>
<th>Ten</th>
<th>Cas</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNMS1</td>
<td>N</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td>Gen=M, Num=S, Cas=1...</td>
</tr>
<tr>
<td>AGFS3</td>
<td>A</td>
<td>G</td>
<td></td>
<td></td>
<td></td>
<td>Gen=F, Num=S, Cas=3...</td>
</tr>
<tr>
<td>P1ZS3FS3</td>
<td>P</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>Gen=Z, Num=S, Cas=3...</td>
</tr>
<tr>
<td>ClXP3</td>
<td>C</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>Gen=X, Num=P, Cas=3...</td>
</tr>
<tr>
<td>VB-S---1P-AA---</td>
<td>V</td>
<td>B</td>
<td>Num=S, Per=1, Ten=P...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dg-------3A----</td>
<td>D</td>
<td>g</td>
<td>Gra=3, Neg=A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RR--6-------</td>
<td>R</td>
<td>R</td>
<td>Cas=6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>J,-X---3-------</td>
<td>J,</td>
<td>Num=X, Per=3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TT---------</td>
<td>T</td>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>II---------</td>
<td>I</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X@---------</td>
<td>X</td>
<td>@</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z:---------</td>
<td>Z</td>
<td>:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Interset: Reusable Tagset Conversion**
NNMS1-----A-----  Ncmsny
AGFS3-----A-----  Afpfsd
P1ZS3FS3-------- Pr3mdsfnayn
ClXP3------------2 Mcmn3y
VB-S---1P-AA---- Vmip1smany
Dg-------------3A---- Rgs
RR--6------------ Sps1
J,-X---3------- Css3
TT--------------- Q
II--------------- I
X©--------------- X
Z:---------------
Majka Tagset from Brno

NNMS1-----A----- k1gMnSc1eA
AGFS3-----A----- k2gFnSc3eA
P1ZS3FS3------- k3gUnSc3p3hFxR
ClXP3--------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** = Coordinating conjunction
- **CD** = Concrete digits
- **DT** = Determiner
- **EX** = existential there
- **FW** = Foreign word
- **IN** = Preposition or subordinating conjunction
- **JJ** = Adjective
- **JJR** = Adjective, comparative
- **JJS** = Adjective, superlative
- **LS** = List marker
- **MD** = Modal verb
- **NN** = Noun, singular or mass
- **NNS** = Noun, plural
- **NNP** = Proper noun, singular
- **NNPS** = Proper noun, plural
- **PDT** = Prepositional phrase modifier
- **POS** = Possessive ending
- **PRP** = Personal pronoun
- **PRP$** = Possessive pronoun
- **RB** = Adverb
- **RBR** = Adverb, comparative
- **RBS** = Adverb, superlative
- **RP** = Particle
- **SYM** = Symbol
- **TO** = To
- **UH** = Interjection...
- **VB** = Verb, base form
- **VBD** = Verb, past tense
- **VBG** = Verb, gerund or present participle
- **VBN** = Verb, past participle
- **VBP** = Verb, 3rd person singular present
- **VBZ** = Verb, 3rd person singular present
- **WDT** = Wh-determiner
- **WP** = Wh-pronominal marker
- **WP$** = Wh-pronominal marker, possessive
- **WRB** = Wh-adverb
- **.** = Period
- **,** = Comma
- **: , : $ # ` ```" LRBR - "RRBR -
<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABL</td>
<td></td>
</tr>
<tr>
<td>ABN</td>
<td></td>
</tr>
<tr>
<td>ABX</td>
<td></td>
</tr>
<tr>
<td>AP</td>
<td></td>
</tr>
<tr>
<td>AP$</td>
<td></td>
</tr>
<tr>
<td>AP+AP</td>
<td></td>
</tr>
<tr>
<td>AT</td>
<td></td>
</tr>
<tr>
<td>BE</td>
<td></td>
</tr>
<tr>
<td>BED</td>
<td></td>
</tr>
<tr>
<td>BED*</td>
<td></td>
</tr>
<tr>
<td>BEDZ</td>
<td></td>
</tr>
<tr>
<td>BEDZ*</td>
<td></td>
</tr>
<tr>
<td>BEG</td>
<td></td>
</tr>
<tr>
<td>BEM</td>
<td></td>
</tr>
<tr>
<td>BEM*</td>
<td></td>
</tr>
<tr>
<td>BEN</td>
<td></td>
</tr>
<tr>
<td>BER*</td>
<td></td>
</tr>
<tr>
<td>BEZ</td>
<td></td>
</tr>
<tr>
<td>BEZ*</td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td></td>
</tr>
<tr>
<td>CD</td>
<td></td>
</tr>
<tr>
<td>CD$</td>
<td></td>
</tr>
<tr>
<td>CS</td>
<td></td>
</tr>
<tr>
<td>DO</td>
<td></td>
</tr>
<tr>
<td>DO*</td>
<td></td>
</tr>
<tr>
<td>DO+PPSS</td>
<td></td>
</tr>
<tr>
<td>DOD</td>
<td></td>
</tr>
<tr>
<td>DOD*</td>
<td></td>
</tr>
<tr>
<td>DOZ</td>
<td></td>
</tr>
<tr>
<td>DOZ*</td>
<td></td>
</tr>
<tr>
<td>DT</td>
<td></td>
</tr>
<tr>
<td>DT$</td>
<td></td>
</tr>
<tr>
<td>DT+BEZ</td>
<td></td>
</tr>
<tr>
<td>DT+MD</td>
<td></td>
</tr>
<tr>
<td>DTI</td>
<td></td>
</tr>
<tr>
<td>DTS</td>
<td></td>
</tr>
<tr>
<td>DTS+BEZ</td>
<td></td>
</tr>
<tr>
<td>DTX</td>
<td></td>
</tr>
<tr>
<td>EX</td>
<td></td>
</tr>
<tr>
<td>EX+BEZ</td>
<td></td>
</tr>
<tr>
<td>EX+HVD</td>
<td></td>
</tr>
<tr>
<td>EX+HVZ</td>
<td></td>
</tr>
<tr>
<td>EX+MD</td>
<td></td>
</tr>
<tr>
<td>FW-*</td>
<td></td>
</tr>
<tr>
<td>FW-AT</td>
<td></td>
</tr>
<tr>
<td>FW-AT+NN</td>
<td></td>
</tr>
<tr>
<td>FW-BE</td>
<td></td>
</tr>
<tr>
<td>FW-BER</td>
<td></td>
</tr>
<tr>
<td>FW-BEZ</td>
<td></td>
</tr>
<tr>
<td>FW-CC</td>
<td></td>
</tr>
<tr>
<td>FW-CD</td>
<td></td>
</tr>
<tr>
<td>FW-CS</td>
<td></td>
</tr>
<tr>
<td>FW-DT</td>
<td></td>
</tr>
<tr>
<td>FW-DT+BEZ</td>
<td></td>
</tr>
<tr>
<td>FW-DTS</td>
<td></td>
</tr>
<tr>
<td>FW-HV</td>
<td></td>
</tr>
<tr>
<td>FW-IN</td>
<td></td>
</tr>
<tr>
<td>FW-IN+AT</td>
<td></td>
</tr>
<tr>
<td>FW-IN+NN</td>
<td></td>
</tr>
<tr>
<td>FW-IN+NP</td>
<td></td>
</tr>
<tr>
<td>FW-JJ</td>
<td></td>
</tr>
<tr>
<td>FW-JJR</td>
<td></td>
</tr>
<tr>
<td>FW-JJT</td>
<td></td>
</tr>
<tr>
<td>FW-NN</td>
<td></td>
</tr>
<tr>
<td>FW-NN$</td>
<td></td>
</tr>
<tr>
<td>FW-NNS</td>
<td></td>
</tr>
<tr>
<td>FW-NP</td>
<td></td>
</tr>
<tr>
<td>FW-NPS</td>
<td></td>
</tr>
<tr>
<td>FW-NR</td>
<td></td>
</tr>
<tr>
<td>FW-OD</td>
<td></td>
</tr>
<tr>
<td>FW-PN</td>
<td></td>
</tr>
<tr>
<td>FW-PP$</td>
<td></td>
</tr>
<tr>
<td>FW-PPL</td>
<td></td>
</tr>
<tr>
<td>FW-PPL+VBZ</td>
<td></td>
</tr>
<tr>
<td>FW-PPO</td>
<td></td>
</tr>
<tr>
<td>FW-PPO+IN</td>
<td></td>
</tr>
<tr>
<td>FW-PPS</td>
<td></td>
</tr>
<tr>
<td>FW-PPSS</td>
<td></td>
</tr>
<tr>
<td>FW-PPSS+HV</td>
<td></td>
</tr>
<tr>
<td>FW-QL</td>
<td></td>
</tr>
<tr>
<td>FW-RB</td>
<td></td>
</tr>
<tr>
<td>FW-RB+CC</td>
<td></td>
</tr>
<tr>
<td>FW-TO+VB</td>
<td></td>
</tr>
<tr>
<td>FW-UH</td>
<td></td>
</tr>
<tr>
<td>FW-VB</td>
<td></td>
</tr>
<tr>
<td>SynTagRus Tags for Russian</td>
<td></td>
</tr>
<tr>
<td>---------------------------</td>
<td></td>
</tr>
<tr>
<td><strong>S ЕД МУЖ ИМ</strong></td>
<td><strong>NNMS1------А-----</strong></td>
</tr>
<tr>
<td><strong>S МН РОД ОД</strong></td>
<td><strong>PSXXXP3----------</strong></td>
</tr>
<tr>
<td><strong>А МН ИМ</strong></td>
<td><strong>AAXP1-----1А-----</strong></td>
</tr>
<tr>
<td><strong>NUM ВИН</strong></td>
<td><strong>ClXX4------------</strong></td>
</tr>
<tr>
<td><strong>V НЕСОВ ИЗЪЯВ НЕПРОШ МН З-Л</strong></td>
<td><strong>VB-P---3P-AA---</strong></td>
</tr>
<tr>
<td><strong>ADV СПАВ</strong></td>
<td><strong>Dg----------2А-----</strong></td>
</tr>
<tr>
<td><strong>PR</strong></td>
<td><strong>RR--6-----------</strong></td>
</tr>
<tr>
<td><strong>CONJ</strong></td>
<td><strong>J^-------------</strong></td>
</tr>
<tr>
<td><strong>PART</strong></td>
<td><strong>TT-------------</strong></td>
</tr>
<tr>
<td><strong>INTJ</strong></td>
<td><strong>II-------------</strong></td>
</tr>
</tbody>
</table>
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.

<token_Arabic>
  <voc>wabiAlfAlwjp</voc>
  <pos>wa/CONJ+bi/PREP+AlfAlwjp/NOUN_PROP</pos>
</token_Arabic>

<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 vous
P07 = 把 bǎ, 將 jiāng
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 nominativ)
AQP0PN0S ... politiska
MC00G0S ... fyras (gt. gen.)
V@IPAS ... har
AP00N0S ... oberoende
RGOS ... inte
PF@000@S ... sig
MAMBA and PAROLE Tagsets for Swedish

<table>
<thead>
<tr>
<th>Tagset</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td>noun</td>
<td>konfliktarna (substantiv utrum pluralis bestämd nominativ)</td>
</tr>
<tr>
<td>PN</td>
<td>proper noun</td>
<td></td>
</tr>
<tr>
<td>VN</td>
<td>gerund</td>
<td></td>
</tr>
<tr>
<td>AJ</td>
<td>adjective</td>
<td>politiska</td>
</tr>
<tr>
<td>AV</td>
<td>verbs</td>
<td></td>
</tr>
<tr>
<td>BV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>QV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HV</td>
<td>the verb</td>
<td>har</td>
</tr>
<tr>
<td>I?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IQ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IR</td>
<td>punctuation</td>
<td>oberoende</td>
</tr>
<tr>
<td>IS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IT</td>
<td></td>
<td>inte</td>
</tr>
<tr>
<td>IU</td>
<td></td>
<td>sig</td>
</tr>
<tr>
<td>IR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IU</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IU</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Intersect: Reusable Tagset Conversion

19/32
<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>rep</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>gest</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>punside</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>prod</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>negativeness</td>
<td>pos</td>
<td>neg</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>end</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>norm</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>
</tbody>
</table>
Reference:

CPAN Perl libraries:
- cpanm install Lingua::Interset

```perl
use Lingua::Interset::Converter;
my $c = new Lingua::Interset::Converter ('from' => 'cs::multext', 'to' => 'cs::pdt');
...
my $target_tag = $c->convert ($source_tag);
```
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
- 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?
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!

```plaintext
'number' =>
{
  'priority' => 320,
  'values' => ['sing', 'dual', 'tri', 'pauc', 'grpa', 'plur', 'replacements' =>
  [
    ['sing'],
    ['dual', 'plur'],
    ['tri', 'plur'],
    ['pauc', 'plur'],
    ['grpa', 'plur'],
    ['plur'],
    ['grpl', 'plur'],
    ['inv'],
    ['ptan', 'plur'],
    ['...'],
  ]
}
```

0 → sing, dual, tri, pauc, ...
sing → 0, dual, tri, pauc, ...
dual → plur, 0, sing, tri, ...
tri → plur, 0, sing, dual, ...
pauc → plur, 0, sing, ...
grpa → plur, 0, sing, ...
plur → 0, sing, dual, tri, ...
grpl → plur, 0, sing, ...
inv → 0, sing, dual, tri, ...
ptan → plur, 0, sing
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

pos
noun
adj
num
verb
adv
adp
conj
part
int
punc

prontype
prs
int
ind

NNMS1-----A-----
pos  noun

Interset: Reusable Tagset Conversion
Substitution Trie

- pos
  - noun
  - adj
  - num
  - verb
  - adv
  - adp
  - conj
  - part
  - int
  - punc

- prontype
  - prs
  - int
  - ind

Interset: Reusable Tagset Conversion
Substitution Trie

pos
noun
adj
num
verb
adv
adp
conj
part
int
punc

prontype
prs
int
ind

definiteness
ind
def

NNMS1-----A-----
pos
noun
polarity
pos

Interset: Reusable Tagset Conversion
Substitution Trie

- pos
  - noun
  - adj
  - num
  - verb
  - adv
  - adp
  - conj
  - part
  - int
  - punc

- prontype
  - prs
  - int
  - ind

- definiteness
  - ind
  - def

- gender
  - com
  - masc
  - neut

- number
  - sing
  - plur

NNMS1

data:
pos: noun
polarity: pos
gender: masc
animacy: anim
number: sing
Substitution Trie

NNMS ------A-----

pos noun
polarity pos
gender masc
animacy anim
number sing
case nom

Interset: Reusable Tagset Conversion
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 $\rightarrow$ 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 their PDT tag (PZ
    if($lemma =~ m/^((ně|lec|ledas?|kde|bůhví|kdoví|nevím|málo|sotva)?(kdo|cos?)(si|ko
    {
        $node->iset()->set('pos', 'noun');
    }
    elsif($lemma =~ m/^((jaký|který)|(jaký|který)$|^každý|všechn|sám|žádný|some|taký
    {
        $node->iset()->set('pos', 'adj');
    }
    # Pronouns čí, něčí, čísi, číkoli, ledačí, kdečí, bůhvíčí, nevímčí, ničí should ha
    elsif($lemma =~ m/^((ně|lec|ledas?|kde|bůhví|kdoví|nevím|ni)?čí|čí(si|koliv?))$/
    {
        $node->iset()->set('pos', 'adj');
        $node->iset()->set('poss', 'poss');
```
- **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