Interset: Reusable Tagset Conversion

Daniel Zeman, Rudolf Rosa
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

Interset: Reusable Tagset Conversion
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, MULTTEXT**
  - 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*
<table>
<thead>
<tr>
<th>Tag</th>
<th>Czech</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNMS1-----A----</td>
<td>Josef</td>
</tr>
<tr>
<td>AGFS3-----A----</td>
<td>následující</td>
</tr>
<tr>
<td>P1ZS3FS3------</td>
<td>jejímuž</td>
</tr>
<tr>
<td>ClXP3-------2</td>
<td>stě</td>
</tr>
<tr>
<td>VB-S---1P-AA---</td>
<td>jsem</td>
</tr>
<tr>
<td>Dg---------3A----</td>
<td>nejméně</td>
</tr>
<tr>
<td>RR--6--------</td>
<td>v</td>
</tr>
<tr>
<td>J,--X---3------</td>
<td>aby</td>
</tr>
<tr>
<td>TT----------</td>
<td>jen</td>
</tr>
<tr>
<td>II----------</td>
<td>ejhle</td>
</tr>
<tr>
<td>X@----------</td>
<td>noor</td>
</tr>
<tr>
<td>Z:----------</td>
<td>,</td>
</tr>
<tr>
<td>Tag</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------</td>
</tr>
<tr>
<td>NNMS1</td>
<td>NMS1A</td>
</tr>
<tr>
<td>AGFS3</td>
<td>AVGFS3A</td>
</tr>
<tr>
<td>P1ZS3FS3</td>
<td>PSEFSZS3</td>
</tr>
<tr>
<td>ClXP3</td>
<td>CGXP3-2</td>
</tr>
<tr>
<td>VB-S</td>
<td>VPS1A</td>
</tr>
<tr>
<td>Dg</td>
<td>DG3A</td>
</tr>
<tr>
<td>RR</td>
<td>R6</td>
</tr>
<tr>
<td>J,-X</td>
<td>JVX3</td>
</tr>
<tr>
<td>TT</td>
<td>T</td>
</tr>
<tr>
<td>II</td>
<td>I</td>
</tr>
<tr>
<td>X@</td>
<td>NOMORPH</td>
</tr>
<tr>
<td>Z:</td>
<td>ZIP</td>
</tr>
</tbody>
</table>
## Prague Tags for CoNLL 2006 Shared Task

<table>
<thead>
<tr>
<th>Tag</th>
<th>Gen</th>
<th>Num</th>
<th>Cas</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNMS1</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>AGFS3</td>
<td>A</td>
<td>G</td>
<td>F</td>
</tr>
<tr>
<td>P1ZS3FS3</td>
<td>P</td>
<td>1</td>
<td>Z</td>
</tr>
<tr>
<td>ClXP3</td>
<td>C</td>
<td>1</td>
<td>X</td>
</tr>
<tr>
<td>VB-S</td>
<td>V</td>
<td>B</td>
<td>Num=1</td>
</tr>
<tr>
<td>Dg</td>
<td>D</td>
<td>g</td>
<td>Gra=3</td>
</tr>
<tr>
<td>RR</td>
<td>R</td>
<td>R</td>
<td>Cas=6</td>
</tr>
<tr>
<td>J,-X</td>
<td>J</td>
<td>Num=X</td>
<td>Per=3</td>
</tr>
<tr>
<td>TT</td>
<td>T</td>
<td>T</td>
<td>_</td>
</tr>
<tr>
<td>II</td>
<td>I</td>
<td>I</td>
<td>_</td>
</tr>
<tr>
<td>X@</td>
<td>X</td>
<td>@</td>
<td>_</td>
</tr>
<tr>
<td>Z:</td>
<td>Z</td>
<td>:</td>
<td>_</td>
</tr>
</tbody>
</table>
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 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 . , : $ # ` " ' -LRB- -RRB-

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

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>EД МУЖ ИМ</td>
</tr>
<tr>
<td>S</td>
<td>МН РОД ОД</td>
</tr>
<tr>
<td>A</td>
<td>МН ИМ</td>
</tr>
<tr>
<td>NUM</td>
<td>ВИН</td>
</tr>
<tr>
<td>V</td>
<td>НЕСОВ ИЗЪЯВ НЕПРОШ МН З-Л</td>
</tr>
<tr>
<td>ADV</td>
<td>СРАВ</td>
</tr>
<tr>
<td>PR</td>
<td></td>
</tr>
<tr>
<td>CONJ</td>
<td></td>
</tr>
<tr>
<td>PART</td>
<td></td>
</tr>
<tr>
<td>INTJ</td>
<td></td>
</tr>
<tr>
<td>NNMS1</td>
<td>-------A-</td>
</tr>
<tr>
<td>PSXXXXP3</td>
<td>-------</td>
</tr>
<tr>
<td>AAXP1</td>
<td>-------1A-</td>
</tr>
<tr>
<td>C1XX4</td>
<td>-------</td>
</tr>
<tr>
<td>VB-P</td>
<td>-------3P-AA-</td>
</tr>
<tr>
<td>Dg</td>
<td>-------2A-</td>
</tr>
<tr>
<td>RR</td>
<td>-------6-</td>
</tr>
<tr>
<td>J~</td>
<td>-------</td>
</tr>
<tr>
<td>TT</td>
<td>-------</td>
</tr>
<tr>
<td>II</td>
<td>-------</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 yóu
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
AP000N0S ... oberoende
RG0S ... inte
PF@000@S ... sig
<table>
<thead>
<tr>
<th>Token</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td>noun</td>
<td>konflikterna</td>
</tr>
<tr>
<td>PN</td>
<td>proper noun</td>
<td>(substantiv utrum pluralis bestämd nominativ)</td>
</tr>
<tr>
<td>VN</td>
<td>gerund</td>
<td>politiska</td>
</tr>
<tr>
<td>AJ</td>
<td>adjective</td>
<td>oberoende</td>
</tr>
<tr>
<td>AV BV FV GV HV KV</td>
<td>verbs</td>
<td>inte</td>
</tr>
<tr>
<td>MV QV SP SV VV WV</td>
<td>verbs</td>
<td>sig</td>
</tr>
<tr>
<td>HV</td>
<td>the verb</td>
<td>har</td>
</tr>
<tr>
<td>I? IC IG IK IP IQ</td>
<td>punctuation</td>
<td></td>
</tr>
<tr>
<td>IR IS IT IU</td>
<td>punctuation</td>
<td></td>
</tr>
<tr>
<td>pos</td>
<td>noun</td>
<td>adj</td>
</tr>
<tr>
<td>----------</td>
<td>------</td>
<td>-----</td>
</tr>
<tr>
<td>subpos</td>
<td>prop</td>
<td>class</td>
</tr>
<tr>
<td>prontype</td>
<td>prs</td>
<td>rep</td>
</tr>
<tr>
<td>punctype</td>
<td>peri</td>
<td>gest</td>
</tr>
<tr>
<td>puncside</td>
<td>ini</td>
<td>fin</td>
</tr>
<tr>
<td>synpos</td>
<td>subst</td>
<td>attr</td>
</tr>
<tr>
<td>poss</td>
<td>poss</td>
<td></td>
</tr>
<tr>
<td>reflex</td>
<td>reflex</td>
<td></td>
</tr>
<tr>
<td>negativeness</td>
<td>pos</td>
<td>neg</td>
</tr>
<tr>
<td>definiteness</td>
<td>ind</td>
<td>def</td>
</tr>
<tr>
<td>gender</td>
<td>masc</td>
<td>fem</td>
</tr>
<tr>
<td>animateness</td>
<td>anim</td>
<td>inan</td>
</tr>
<tr>
<td>number</td>
<td>sing</td>
<td>dual</td>
</tr>
<tr>
<td>case</td>
<td>nom</td>
<td>gen</td>
</tr>
<tr>
<td>prep</td>
<td>npr</td>
<td>pre</td>
</tr>
<tr>
<td>degree</td>
<td>pos</td>
<td>com</td>
</tr>
<tr>
<td>person</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>politeness</td>
<td>inf</td>
<td>pol</td>
</tr>
<tr>
<td>possgender</td>
<td>masc</td>
<td>fem</td>
</tr>
<tr>
<td>posnumber</td>
<td>sing</td>
<td>dual</td>
</tr>
<tr>
<td>subcat</td>
<td>intr</td>
<td>tran</td>
</tr>
<tr>
<td>verbform</td>
<td>fin</td>
<td>inf</td>
</tr>
<tr>
<td>mood</td>
<td>ind</td>
<td>imp</td>
</tr>
<tr>
<td>tense</td>
<td>past</td>
<td>pres</td>
</tr>
<tr>
<td>subtense</td>
<td>aor</td>
<td>imp</td>
</tr>
<tr>
<td>aspect</td>
<td>imp</td>
<td>perf</td>
</tr>
<tr>
<td>voice</td>
<td>act</td>
<td>pass</td>
</tr>
<tr>
<td>foreign</td>
<td>foreign</td>
<td></td>
</tr>
<tr>
<td>abbr</td>
<td>abbr</td>
<td></td>
</tr>
<tr>
<td>hyph</td>
<td>hyph</td>
<td></td>
</tr>
<tr>
<td>style</td>
<td>arch</td>
<td>form</td>
</tr>
<tr>
<td>typo</td>
<td>typo</td>
<td></td>
</tr>
<tr>
<td>variant</td>
<td>short</td>
<td>long</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
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?
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!

```
'enumber' =>
{
  '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'],
    ['coll', 'sing'],
    '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

Interset: Reusable Tagset Conversion
Substitution Trie

NNMS1-----A-----

pos
pos
polarity

pos
noun

Interset: Reusable Tagset Conversion
Substitution Trie

- pos
- prontype
- definiteness
- gender

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

- prs
- int
- ind
- ind
- def
- def
- masc
- com
- neut
- com
- neut
- com
- neut

- NN
- S
- pos
- polarity
- gender
- animacy

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

Interset: Reusable Tagset Conversion
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 $\rightarrow$ 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|všechnen|sám|žádný|some|taký
    {
        $node->iset()->set('pos', 'adj');
    }
    # Pronouns čí, něčí, čísi, čikoli, ledačí, kdečí, bůhvíčí, nevímcí, níčí should ha
    elsif($lemma =~ m/^(ně|lec|ledas?|kde|bůhví|kdoví|nevím|ni)?čí|čí(isi|koliv?)$/)
    {
        $node->iset()->set('pos', 'adj');
        $node->iset()->set('poss', 'poss');
    }
```
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