Interset: Reusable Tagset Conversion

Daniel Zeman, Rudolf Rosa

March 20, 2020




Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

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

Prague Tags for Czech

NNMS1-----A----	Josef
AGFS3-----A----	následující
P1ZS3FS3-------	jejímuž
ClXP3---------2	stě
VB-S---1P-AA---	jsem
Dg-------3A----	nejméně
RR--6----------	v
J,-X---3-------	aby
TT-------------	jen
II-------------	ejhle
X@-------------	noor
Z:-------------	,

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

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

```
NNMS1-----A----  Ncmsny
AGFS3-----A----  Afpfsd
P1ZS3FS3-------  Pr3mdsfnayn
ClXP3---------2  Mcmn3y
VB-S---1P-AA---  Vmip1smanyn
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 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

ABL ABN ABX AP AP$ AP+AP AT BE BED BED* BEDZ BEDZ* BEG BEM BEM* BEN BER
BER* BEZ BEZ* CC CD CD$ CS DO DO* DO+PPSS DOD DOD* DOZ DOZ* DT DT$ DT+BEZ
DT+MD DTI DTS DTS+BEZ DTX EX EX+BEZ EX+HVD EX+HVZ EX+MD FW-* FW-AT
FW-AT+NN FW-BE FW-BER FW-BEZ FW-CC FW-CD FW-CS FW-DT FW-DT+BEZ FW-DTS
FW-HV FW-IN FW-IN+AT FW-IN+NN FW-IN+NP FW-JJ FW-JJR FW-JJT FW-NN FW-NN$
FW-NNS FW-NP FW-NPS FW-NR FW-OD FW-PN FW-PP$ FW-PPL FW-PPL+VBZ FW-PPO
FW-PPO+IN FW-PPS FW-PPSS FW-PPSS+HV FW-QL FW-RB FW-RB+CC FW-TO+VB FW-UH
FW-VB...

SynTagRus Tags for Russian

S ЕД МУЖ ИМ	NNMS1-----A----
S МН РОД ОД	PSXXXXP3-------
A МН ИМ	AAXP1----1A----
NUM ВИН	ClXX4----------
V НЕСОВ ИЗЪЯВ НЕПРОШ МН 3-Л	VB-P---3P-AA---
ADV СРАВ	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

Anncorra from IIIT Hyderabad

```
NN NST NNP PRP DEM VM VAUX JJ RB PSP RP CC WQ QF QC QO CL INTF INJ NEG UT
SYM *C RDP ECH UNK
```

- 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

Arabic by Tim Buckwalter

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
...

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 nominativ)

`AQP0PNOS` … *politiska*
`MC00GOS` … *fyras* (gt. gen.)
`V@IPAS` … *har*
`AP000NOS` … *oberoende*
`RGOS` … *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)

`AQPOPNOS` … *politiska*

`V@IPAS` … *har*

`APOOONOS` … *oberoende*
`RGOS` … *inte*
`PF@OOO@S` … *sig*

pos	noun	adj	num	verb	adv	prep	conj	part	int	punc	
subpos	prop	class	pdt	det	art	digit	roman	card	ord	...	
prontype	prs	rcp	int	rel	dem	neg	ind	tot			
punctype	peri	qest	excl	quot	brck	comm	colo	semi	dash	symb	root
puncside	ini	fin									
synpos	subst	attr	adv	pred							
poss	poss										
reflex	reflex										
negativeness	pos	neg									
definiteness	ind	def	red								
gender	masc	fem	com	neut							
animateness	anim	inan									
number	sing	dual	plu								
case	nom	gen	dat	acc	voc	loc	ins				
prepcase	npr	pre									
degree	pos	com	sup	abs							
person	1	2	3								
politeness	inf	pol									
possgender	masc	fem	com	neut							
possnumber	sing	dual	plu								
subcat	intr	tran									
verbform	fin	inf	sup	part	trans	ger					
mood	ind	imp	cnd	sub	jus						
tense	past	pres	fut								
subtense	aor	imp	ppq								
aspect	imp	perf									
voice	act	pass									
foreign	foreign										
abbr	abbr										
hyph	hyph										
style	arch	form	norm	coll							
typo	typo										
variant	short	long	0	1	2	3	4	5	6	7	8	9

Interset

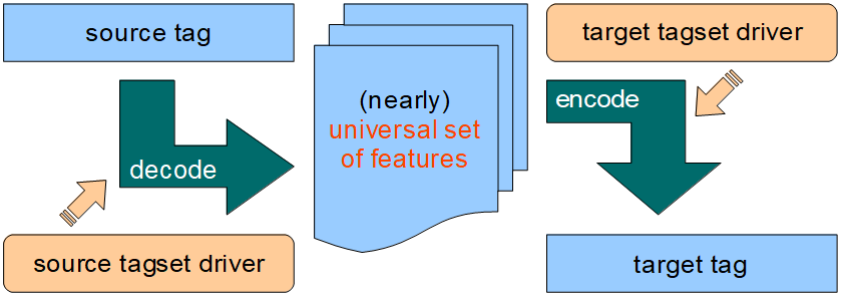
- Reference:
  - Daniel Zeman. 2008. Reusable Tagset Conversion Using Tagset Drivers. In Proceedings of LREC.
  - Daniel Zeman, Philip Resnik: Cross-Language Parser Adaptation between Related Languages. In: Proceedings of IJCNLP 2008 Workshop on NLP for Less Privileged Languages. Hajdarábád, Indie, 2008.

- CPAN Perl libraries:
  - `cpanm install Lingua::Interset`

```
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?

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? $\Rightarrow$ 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? $\Rightarrow$ Preserve as much info as possible!
- Use a pre-trained black-box tool? $\Rightarrow$ 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!

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

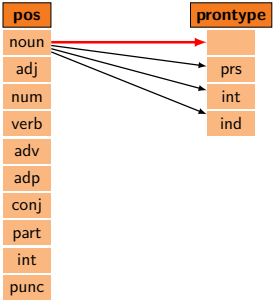
0 → sing, dual, tri, pauc, ...
sing → 0, dual, tri, pauc, ...
dual → plur, 0, sing, tri, ...
tri → plur, 0, sing, dual, ...
pauc → plur, 0, sing, ...
grpa → plur, 0, sing, ...
plur → 0, sing, dual, tri, ...
grpl → plur, 0, sing, ...
inv → 0, sing, dual, tri, ...

Enforcing Defaults

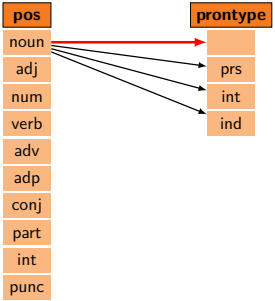
- Decode all known target tags
- Construct trie for known feature-value combinations
- Follow path in trie when encoding
- If a value is not allowed, find the best substitute

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

Substitution Trie

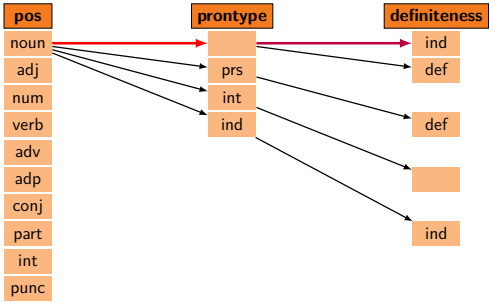


Substitution Trie



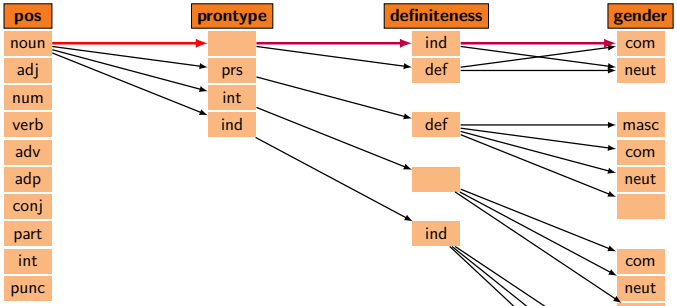
NNMS1-----A----	
pos	**noun**
polarity	**pos**

Substitution Trie



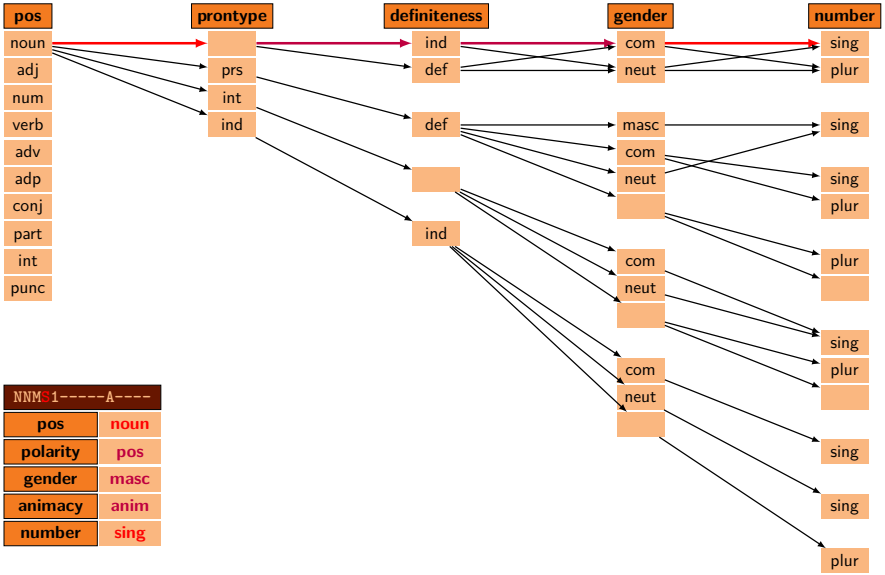
NNMS1-----A----	
pos	noun
polarity	pos

Substitution Trie

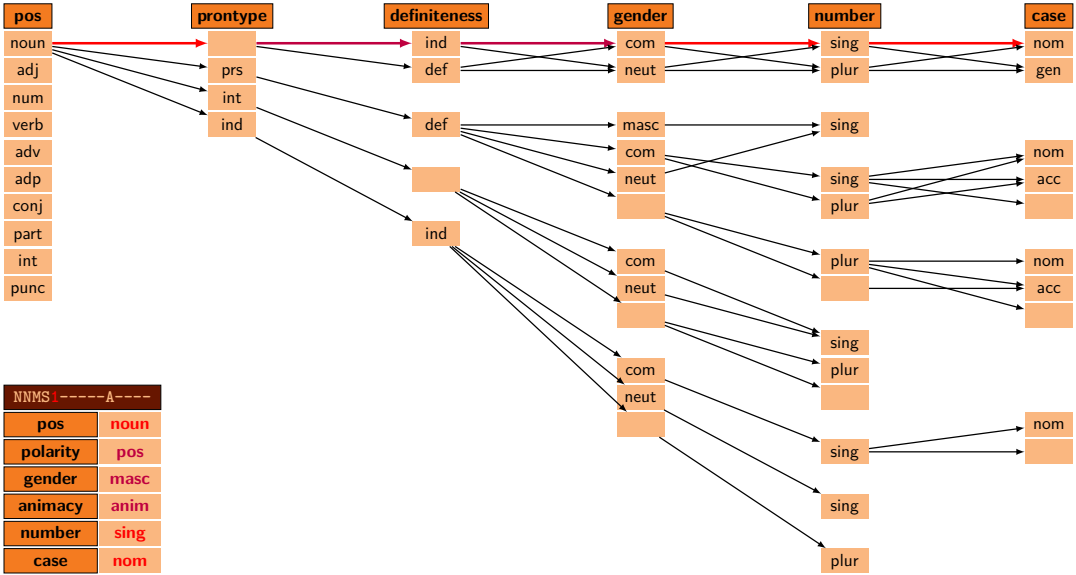


NNPS1-----A----	
pos	noun
polarity	pos
gender	masc
animacy	anim

Substitution Trie



Substitution Trie



Google Universal Part-of-Speech Tags

- Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In Proceedings of LREC.

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

```
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|všechen|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');
```

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

- `http://quest.ms.mff.cuni.cz/cgi-bin/interset/index.pl?tagset=pt::freeling`