From the Jungle to a Park:
Harmonizing Annotations across Languages

Daniel Zeman

Charles University in Prague

From the Jungle to a Park:
Harmonizing Annotations across Languages

Daniel Zeman

Based on joint work with many great people, including

Philip Resnik, Alexandr Rosen, Zdeněk Žabokrtský, Martin Popel, Loganathan Ramasamy, David Mareček, Rudolf Rosa, Jan Štěpánek, Jan Hajič, Joakim Nivre, Chris Manning, Ryan McDonald, Slav Petrov, Filip Ginter, Sampo Pyysalo, Reut Tsarfaty, Yoav Goldberg, Natalia Silveira, Tim Dozat, and many others…

The research has been supported by the grant GA15-10472S.

Too Few or Too Many?

- In 2000, dependency trees were quite rare.

Too Few or Too Many?

- CoNLL 2006: dependency treebanks for 13 languages.
- What about the remaining 6987?

Too Few or ~~Too Many~~?

- Min. 83 treebanks for 51 languages
- An impenetrable jungle of annotation styles!
- Still there are about 6949 languages out in the desert…

Outline

- Cross-language learning (historical motivation)
- Normalization: morphology
- Normalization: dependencies
- Cross-language learning (current work)

Cross-Language Parser Adaptation

- 2006 with Philip Resnik (University of Maryland)
- *Delexicalized parsing*

Cross-Language Parser Adaptation

- 2006 with Philip Resnik (University of Maryland)
- *Delexicalized parsing*
- Gained on popularity now
  - McDonald, Petrov & Hall (EMNLP 2011)
  - Oskar Täckström (dissertation 2013)
  - Loganathan Ramasamy (dissertation 2014)
  - Rosa & Žabokrtský (IWPT 2015)

Parser Adaptation

- Idea:
  - Related languages L1 and L2
    - L1 treebank and morphology
    - L2 morphology
  - Train parser on L1 morphological features
  - Apply the parser to L2

- We took:
  - L1 = Danish [da]
  - L2 = Swedish [sv]

Danish – Swedish Setup

- CoNLL 2006 treebanks (dependencies)
  - Danish Dependency Treebank
  - Swedish Talbanken05
- Two constituency parsers:
  - “Charniak”
  - “Brown” (Charniak N-best parser + Johnson reranker)
- Other resources
  - JRC-Acquis parallel corpus
  - Hajič tagger for Swedish (PAROLE tagset)

Most Frequent da/sv Words

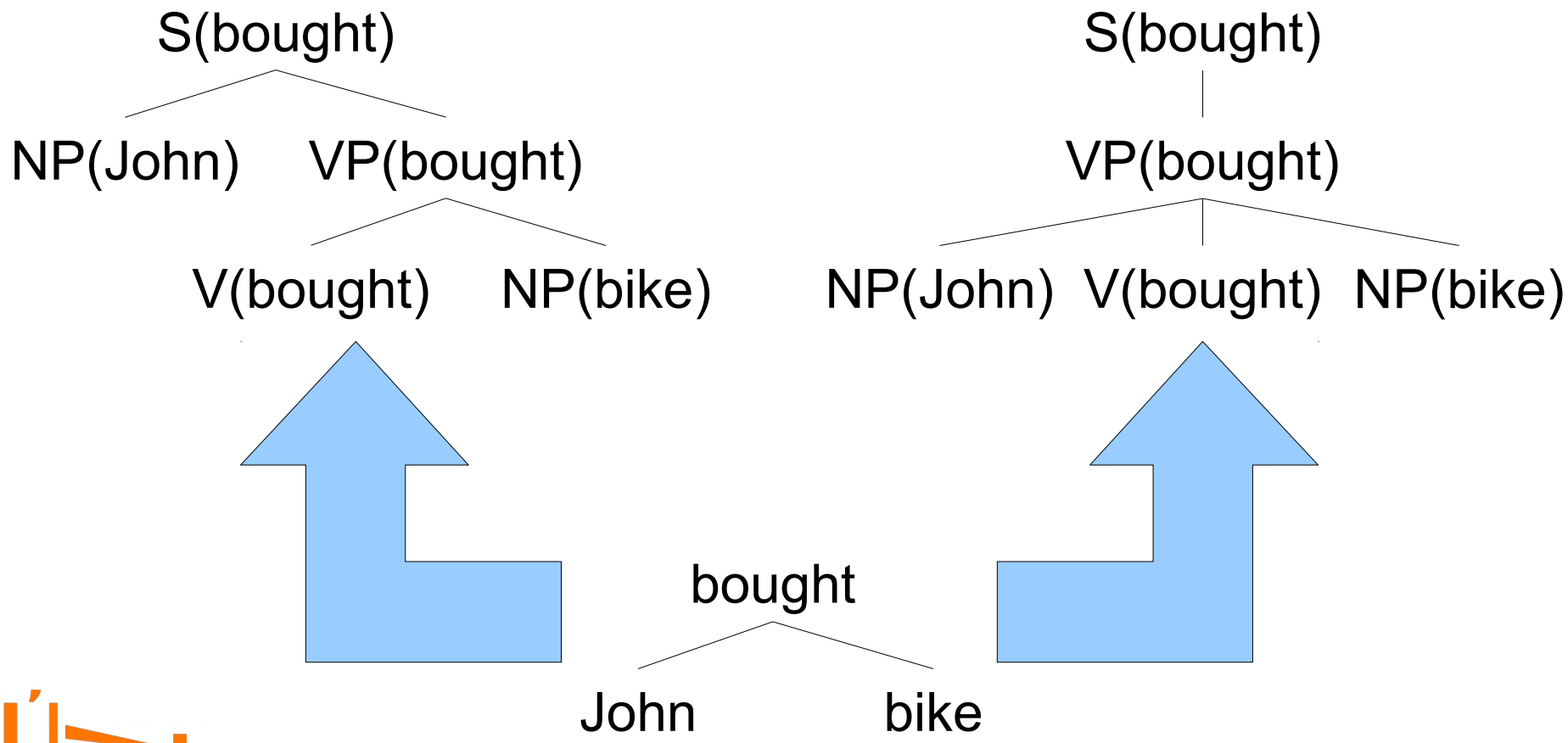
• *i*	0.024	• *och*	0.027
• *og*	0.024	• *att*	0.027
• *at*	0.021	• *i*	0.021
• *er*	0.017	• *är*	0.018
• *en*	0.014	• *som*	0.017
• *til*	0.013	• *en*	0.015
• *af*	0.013	• *det*	0.013
• *det*	0.012	• *av*	0.012
• *på*	0.012	• *på*	0.011

JRC-Acquis Aligned Example

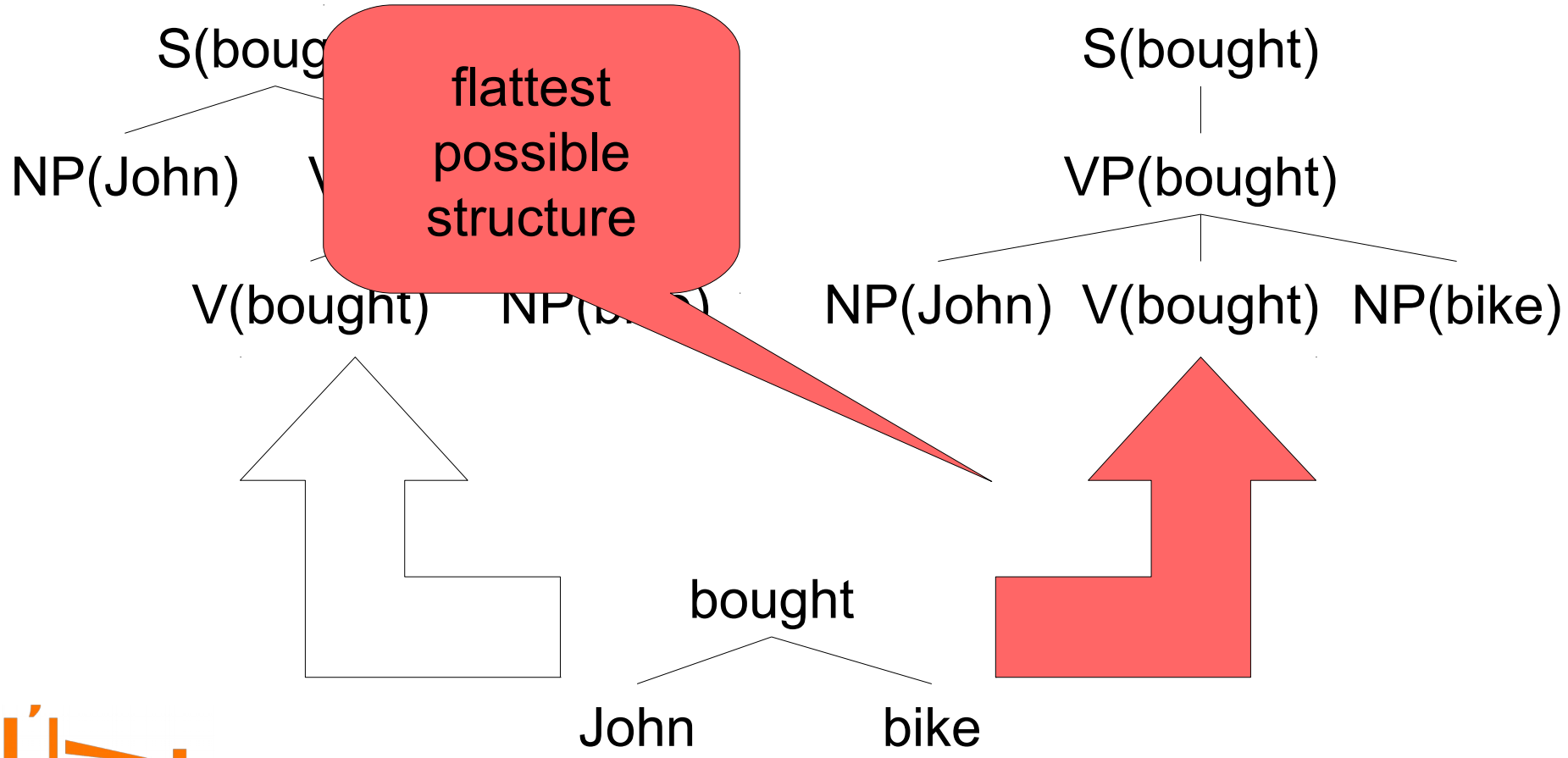
- *Enhver kontraherende part kan opsige denne konvention ved skriftlig henvendelse til depositaren.*

- *En fördragsslutande part får säga upp denna konvention genom skriftlig notifikation till depositarien.*

Treebank Preparation



Treebank Preparation



Treebank Preparation

- DA / SV tagset converted to the Penn Treebank tags
- Nonterminal labels:
  - derived from POS tags
  - then translated to the Penn set of nonterminals

- Make the parser feel it works with the Penn TB
- (Although it could have been configured to use other sets of labels.)

Treebank Normalization

Danish

- DET governs ADJ, ADJ governs NOUN
- NUM governs NOUN
- GEN governs NOM
  Ruslands vej
  Russia's way
- COORD: last member on conjunction, everything else on first member

Swedish

- NOUN governs both DET and ADJ
- NOUN governs NUM
- NOM governs GEN
  års inkomster
  year's income
- COORD: member on previous member, commas and conjs on next member

Treebank Normalization

- A few heuristics
- Transform Danish to the Swedish tree style

- Concrete annotation style does not matter!
  - This was only for testing
  - Hypothesis: no Swedish treebank
  - Than any style is good

Parsing Danish Treebank

- CoNLL test: 322 sents, 5852 words
- CoNLL training: 5190 sents, 94386 words
  - 4900 sents my training
  - 290 sents my devtest
- Following are results on CoNLL test

Parser	F = 2PR/(P+R)
Charniak	78.16
Brown	78.24

Parsing Swedish Treebank

- CoNLL test: 389 sents, 5656 words
- CoNLL training: 11042 sents, 191467 words
  - 10700 sents my training
  - 342 sents my devtest
- Following are results on CoNLL test

Parser	F = 2PR/(P+R)
Charniak	77.81
Brown	78.74

Parsing Swedish with Danish Parser

- Trained on Danish training data
- Parse Swedish test data
- No morphology tweaking so far!
  - Most words are UNKNOWN
- Following are results on CoNLL test

Parser	F = 2PR/(P+R)
Charniak	43.28
Brown	41.84

Delexicalized Parsing

- What if we feed the parser with tags instead of words?
  - *Ændringer i listen i bilaget offentliggøres og meddeles på samme måde.*
  - `NNS IN NN IN NN VB CC VB IN DT NN`
  - `NNS IN NN MD VB CC VB IN DT NN`
  - *Förändringar i förteckningen skall offentliggöras och meddelas på samma sätt.*

Delexicalized Parsing

- What if we feed the parser with tags instead of words?
  - *Ændringer i listen i bilaget offentliggøres og meddeles på samme måde.*
  - `((NNS (IN NN (IN NN))) ((VB CC VB) (IN (DT NN))))`
  - `((NNS (IN NN)) ((MD (VB CC VB)) (IN (DT NN))))`
  - *Förändringar i förteckningen skall offentliggöras och meddelas på samma sätt.*

Delexicalized Parsing

- Trained on Danish training data (tags only)
- Parse Swedish test data (tags from Hajič tagger)
- Restuff Swedish trees with original words
- All data in hybrid Swedish-Danish Hajič-like tagset
  - ("words" = sv/da tags, "tags" = Penn tags)

Parser	da-da	sv-sv	da-sv
Charniak	79.62	76.07	65.50
Brown	80.20	77.01	66.40

Glosses

- JRC-Acquis is a parallel corpus
  - more than 430,000 sentences
- Giza++ & lexical weighting generate a da-sv glossary
- Always use highest-weighted gloss
- Translate Swedish word-by-word to Danish
- Many words are no longer unknown!

Excerpt from sv-da Glossary

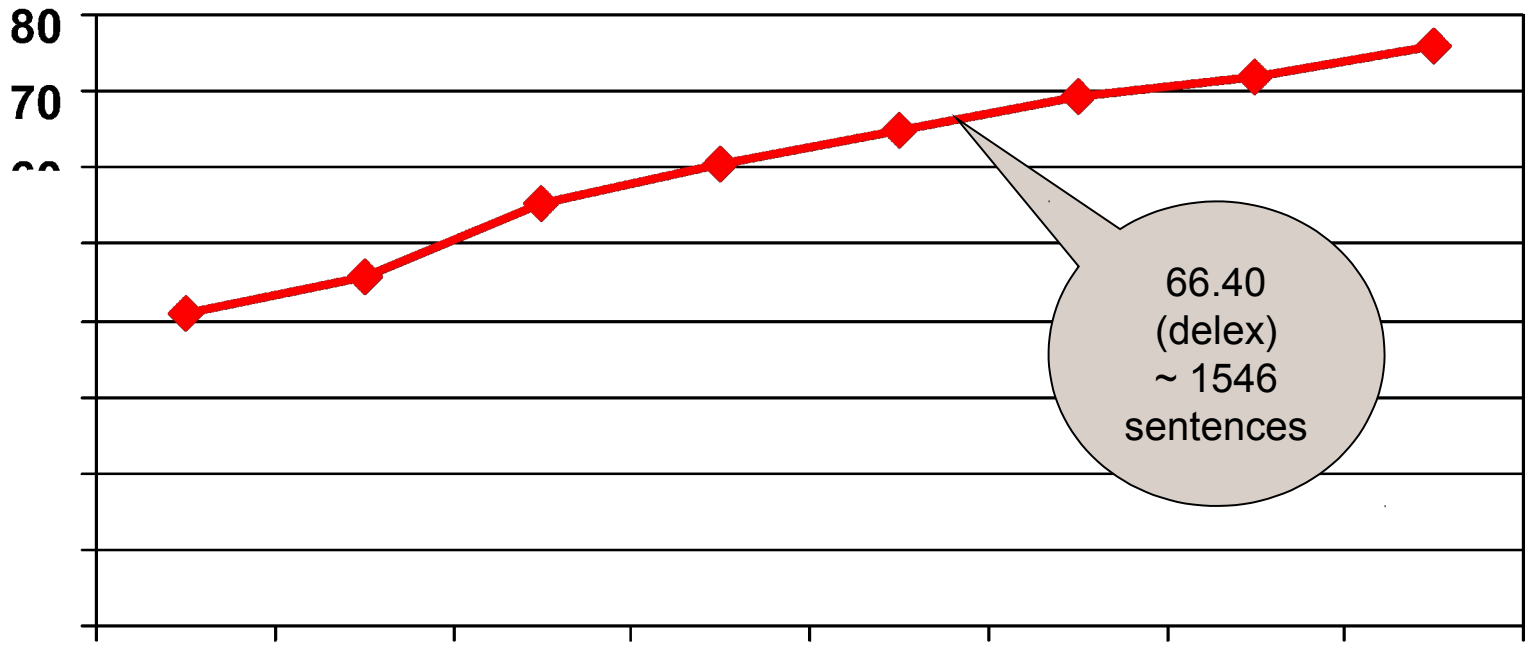
- *behandlingsaktörer*
- *behandlingsanläggning*
- *behandlingsanläggningar*
- *behandlingsanläggningen*
- *behandlingsdatum*
- *behandlingsformer*
- *behandlingsfrister*
- *behandlingsförfaranden*
- *behandlingsförsök*
- *behandlingsindikation*
- *behäftad*
- *behandlingsvirksomheder*
- *behandlingsanlæg*
- *behandlingsvirksomheders*
- *behandlingsanlægget*
- *datøn*
- *behandlingsmuligheder*
- *frister*
- *behandlingsprocedurer*
- *befolkningsforsøg*
- *indikation*
- *behæftet*

Glossed parsing

- Trained on Danish training data
- Translate Swedish test data to Danish
- Parse it using Danish-trained model
- Restuff trees with Swedish and evaluate
- Following are results on CoNLL test

Parser	F = 2PR/(P+R)
Charniak	63.40
Brown	61.50

How big a Swedish treebank can produce the same results?



Outline

- Cross-language learning (historical motivation)
- Normalization: morphology
- Normalization: dependencies
- Cross-language learning (current work)

Tagset Mapping: Interset

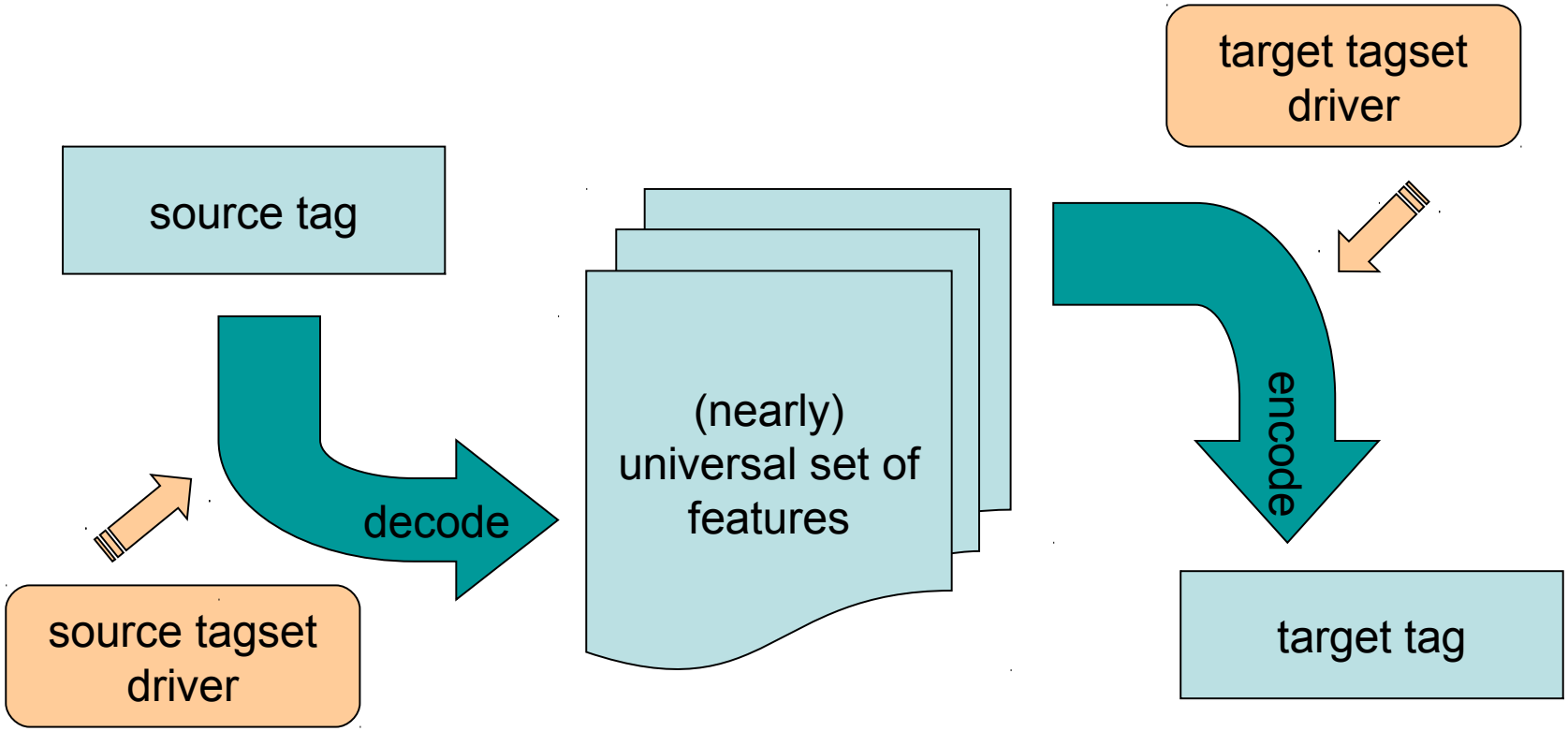
- Already mentioned: da/sv → Penn
- We want to preserve features that
  - are present in both [da] and [sv]
  - are not present in Penn
- This is **CRUCIAL:**
  - unmapped tags are unknown words again
- Mapping tags is always hard even for the same language
- Languages can be similar, approaches way different!

Tagset Discrepancy Examples

- No determiners in [da], pronouns instead
- Subject/object pronoun forms in [sv] (cf. [en] *he/him*), nominative vs. “unmarked” case in [da]
- Masculine gender in [sv] (pronouns)
- Numerals are adjectives in [da]

- Supine in [sv] – probably the only difference truly caused by the language

Interset



reusable!

Limitations

- Universal features (the “interlingua”)
  - should be linguistically adequate
  - built bottom-up: new features/values added when needed
  - “marginal” phenomena may be ignored?
- Tagset conversion
  - motivated rather technically than linguistically
    - (why would a linguist use a Swedish tagset for Danish?)
  - we may lose information (if target tagset cannot encode it)
  - we do not add information (Interset is not a tagger!)

Interset: current state

pos	noun	adj	num	verb	adv	adp	conj	part	int	punc	sym
nountype	com	prop	class								
prontype	prn	prs	rcp	art	int	rel	exc	dem	neg	ind	tot
punctype	peri	qest	excl	quot	brck	comm	colo	semi	dash	root	
puncside	ini	fin									
morphpos	noun	adj	pron	num	adv	mix	def				
poss	poss										
reflex	reflex										
negativeness	pos	neg									
definiteness	ind	def	red	com							
gender	masc	fem	com	neut							
animateness	anim	inan	nhum								
number	sing	dual	plur	ptan	coll						
case	nom	gen	dat	acc	voc	loc	ins	abl	del	par	dis	ess	…
prepcase	npr	pre									
degree	pos	com	sup	abs	dim						
person	1	2	3								
politeness	inf	pol									
possgender	masc	fem	com	neut							
possnumber	sing	dual	plur								
possednumber	sing	dual	plur								
subcat	intr	tran									
verbform	fin	inf	sup	part	trans	ger	gdv				
mood	ind	imp	cnd	sub	jus	pot	opt	des	nec	qot	
tense	past	pres	fut	aor	imp	nar	pqp				
aspect	imp	perf	pro	prog							
voice	act	pass	mid	rcp	cau	int					
foreign	foreign	fscript	tscript								
abbr	abbr										
hyph	hyph										
style	arch	form	norm	coll	rare	poet	vrnc	slng	expr	derg	vulg
typo	typo										
variant	short	long	0	1	2	3	4	5	6	7	8	9
tagset	*cs::pdt*											
other	*{ obscure_feature_1 => [0, 7,351.2, [„a“, „b“]] }*											

absperson	absnumber	abspoliteness	echo
ergperson	ergnumber	ergpoliteness	erggender
datperson	datnumber	datpoliteness	datgender
possperson			

nametype
adjtype
numtype
verbtype
advtype
adpostype
conjtype
parttype
numvalue
numform
position

60 features
349 values

Disjunctive Values

- Tag says that gender is *masc* or *neut*.
- Interset stores list of alternative values.
- We cannot represent alternative combinations of values, for example:
  - either *feminine singular*,
  - or *neuter plural*,
  - but not *feminine plural* or *neuter singular*

Does It Fit in Target Tagset?

- We fill only representable features
- The rest will be lost
- WARNING: it may be “representable” but still alien!
  - Swedish knows: pos = noun & gender = com | neut
  - And also: prontype = prs & gender = masc | fem | com | neut
  - Czech input: pos = noun & gender = masc

  - Keep the “alien” combination in Swedish?

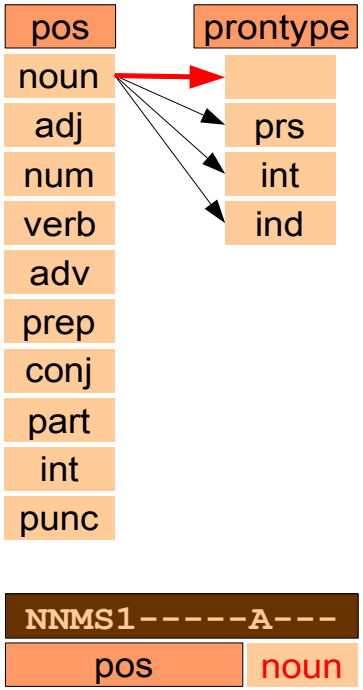
Alien Tags in Target Tagset

- What is the goal of the conversion?
  - Corpus query etc. => keep alien tags
  - Blackbox tool => avoid data that it does not expect

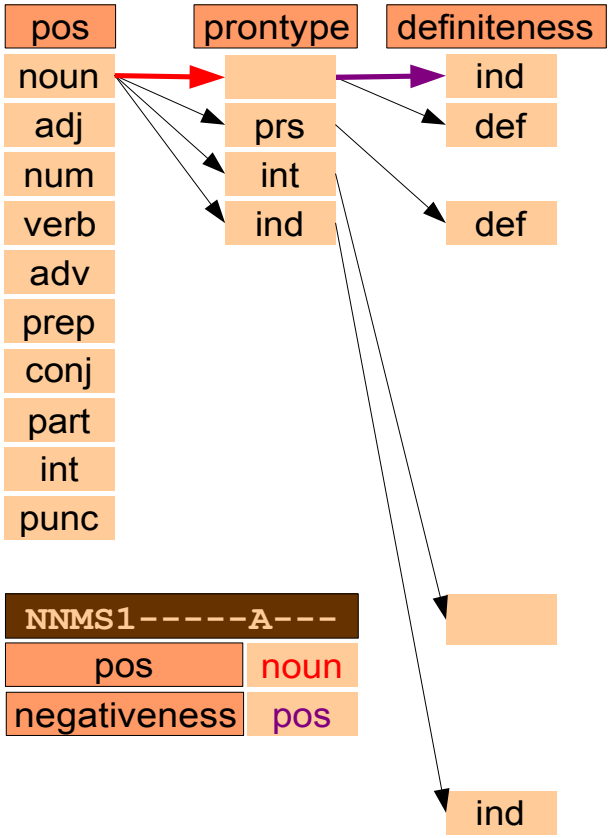
- Atomic tagsets (Penn): no choice
- Structured tags (features encoded separately): impossible combinations can be represented

- How do we avoid them?

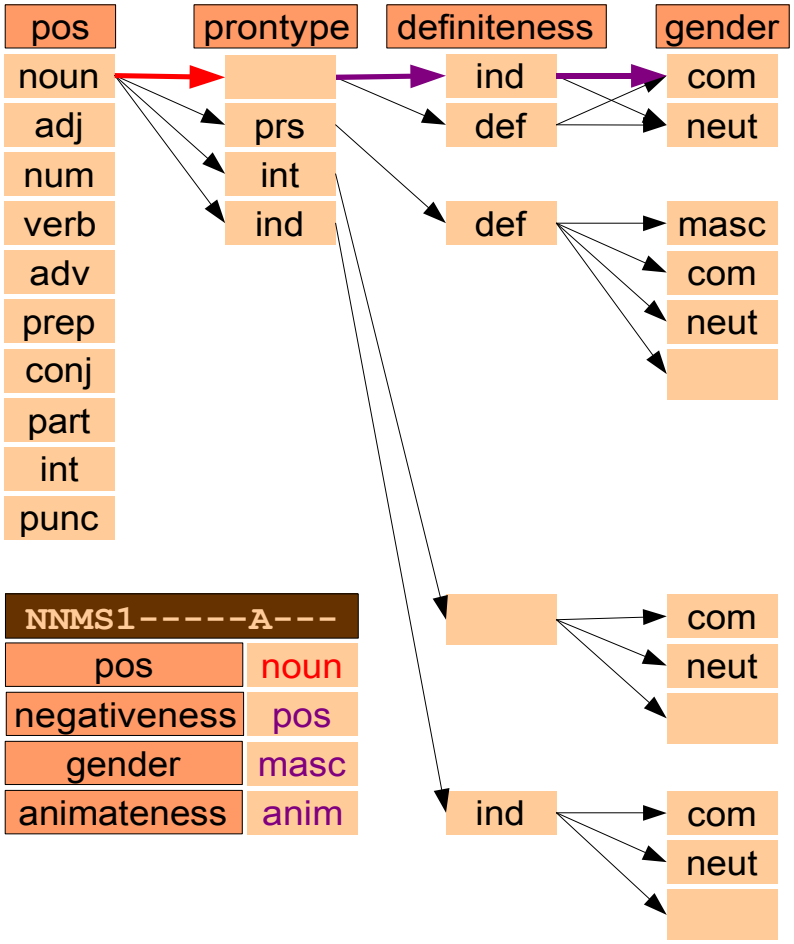
Example: cs → sv



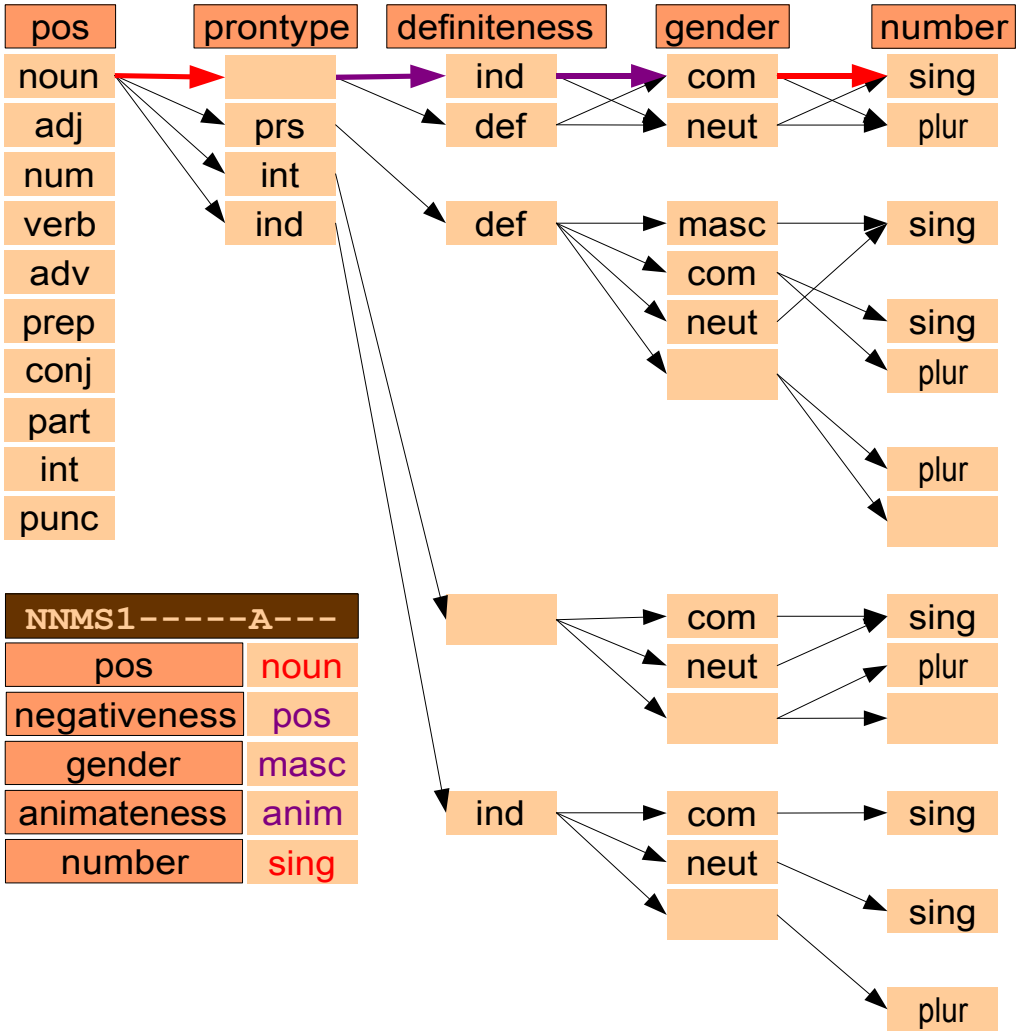
Example: cs → sv



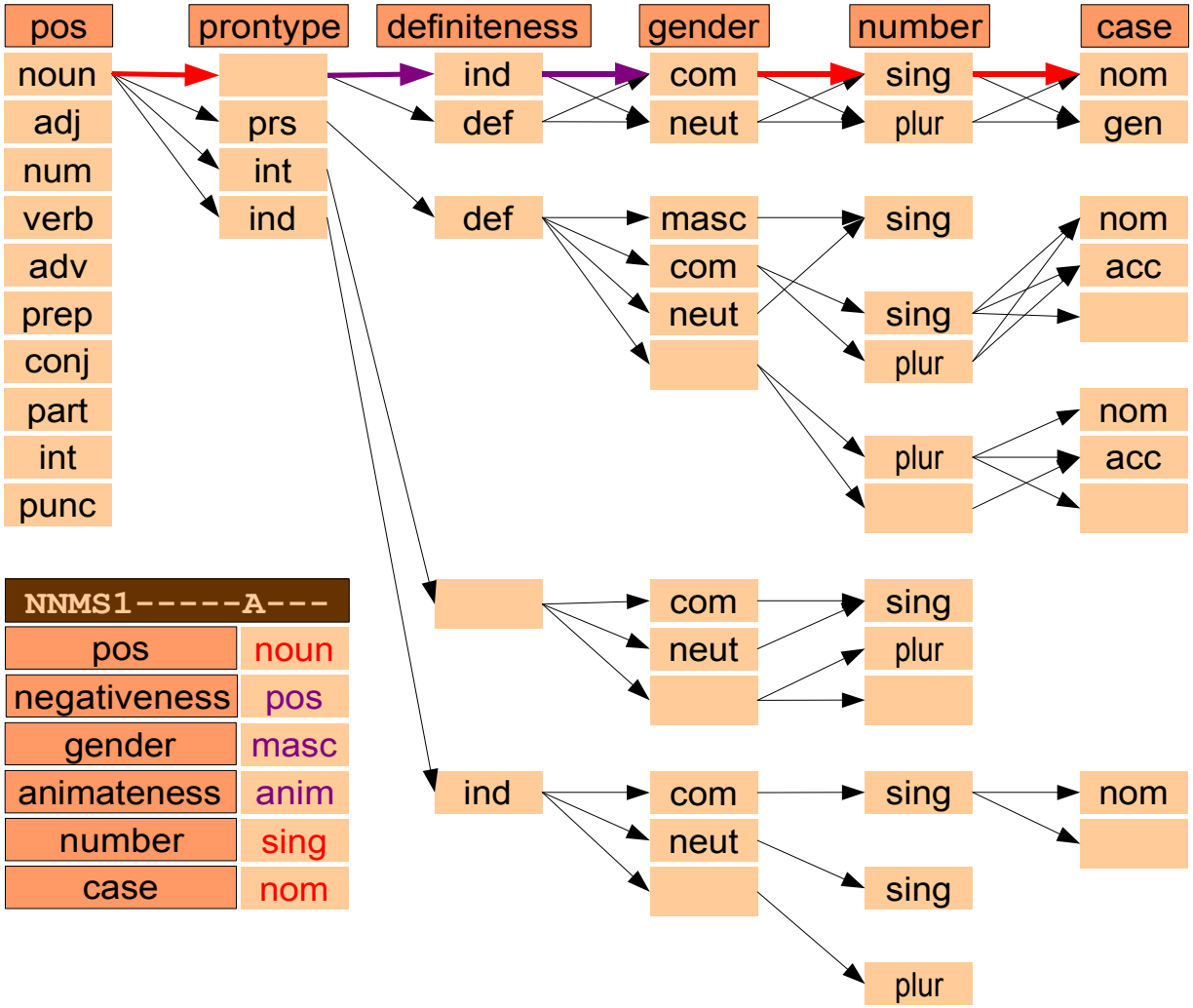
Example: cs → sv



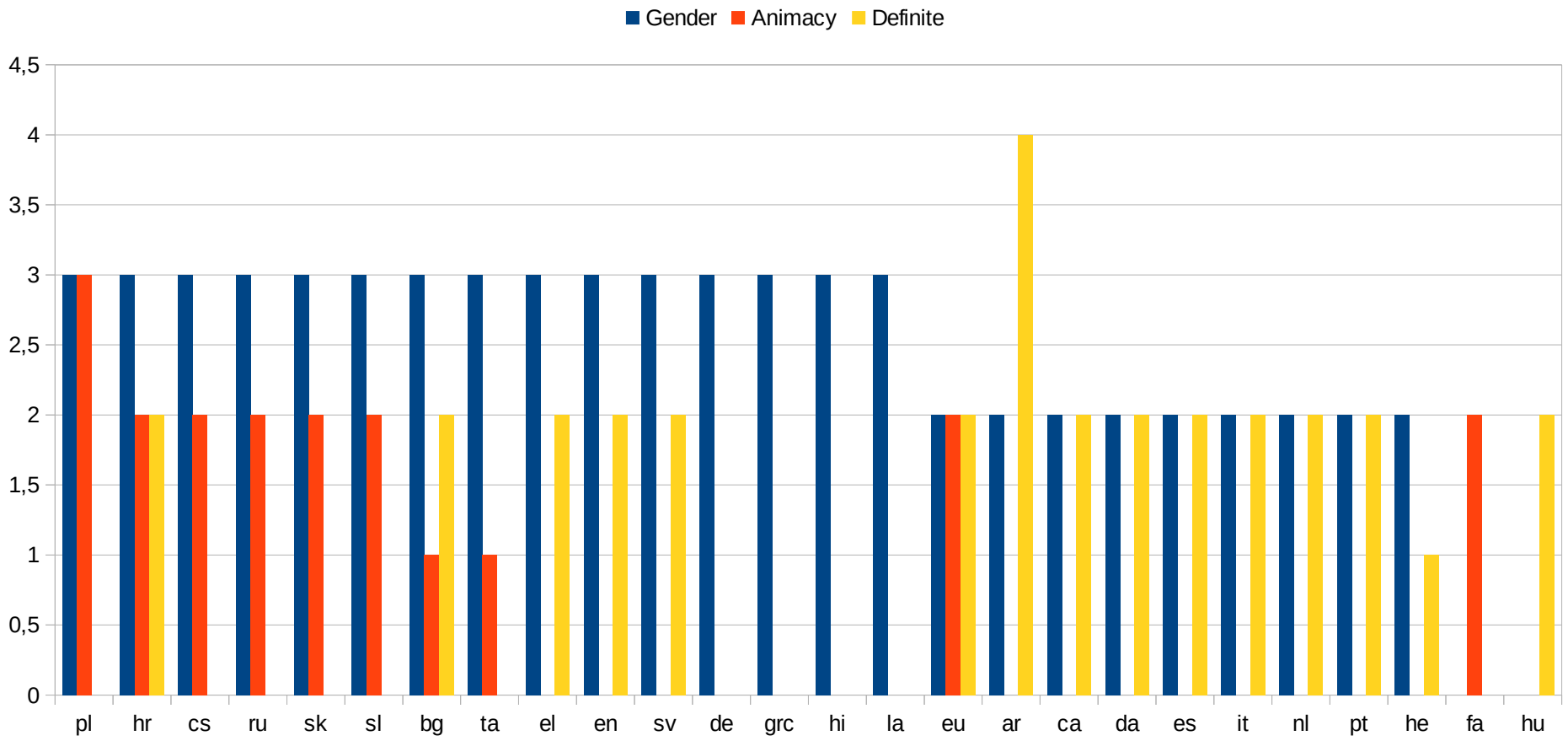
Example: cs → sv



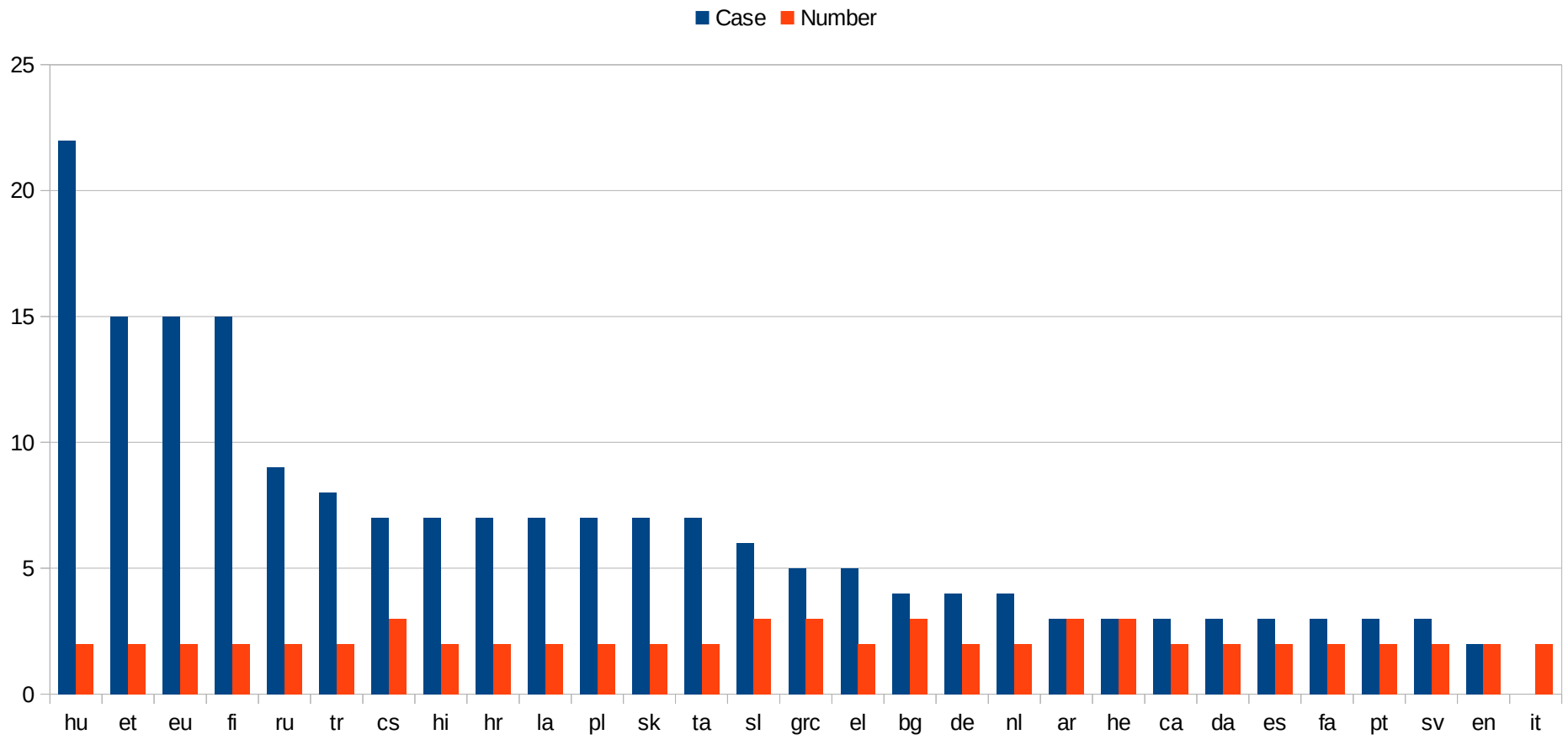
Example: cs → sv



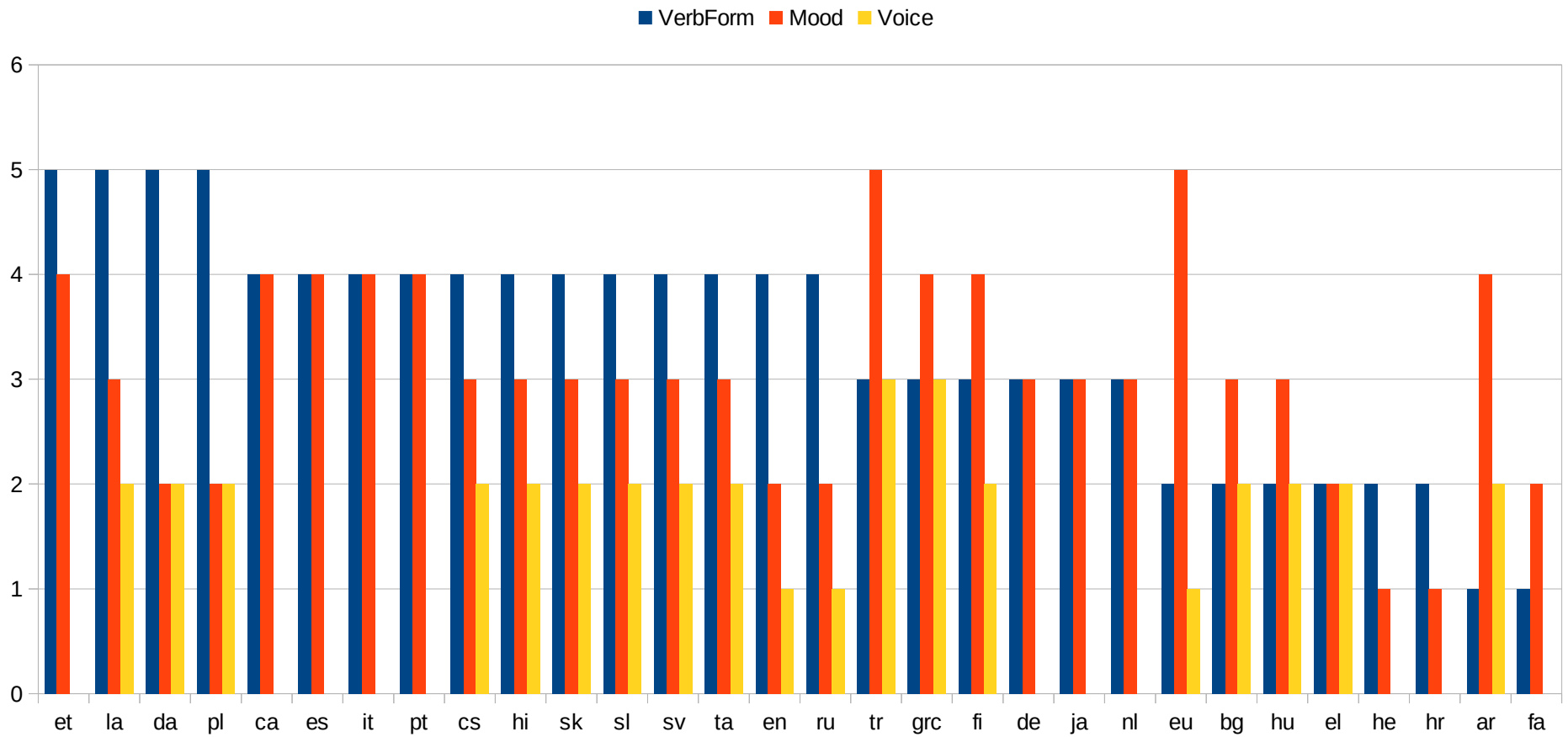
Gender & Animacy & Definite



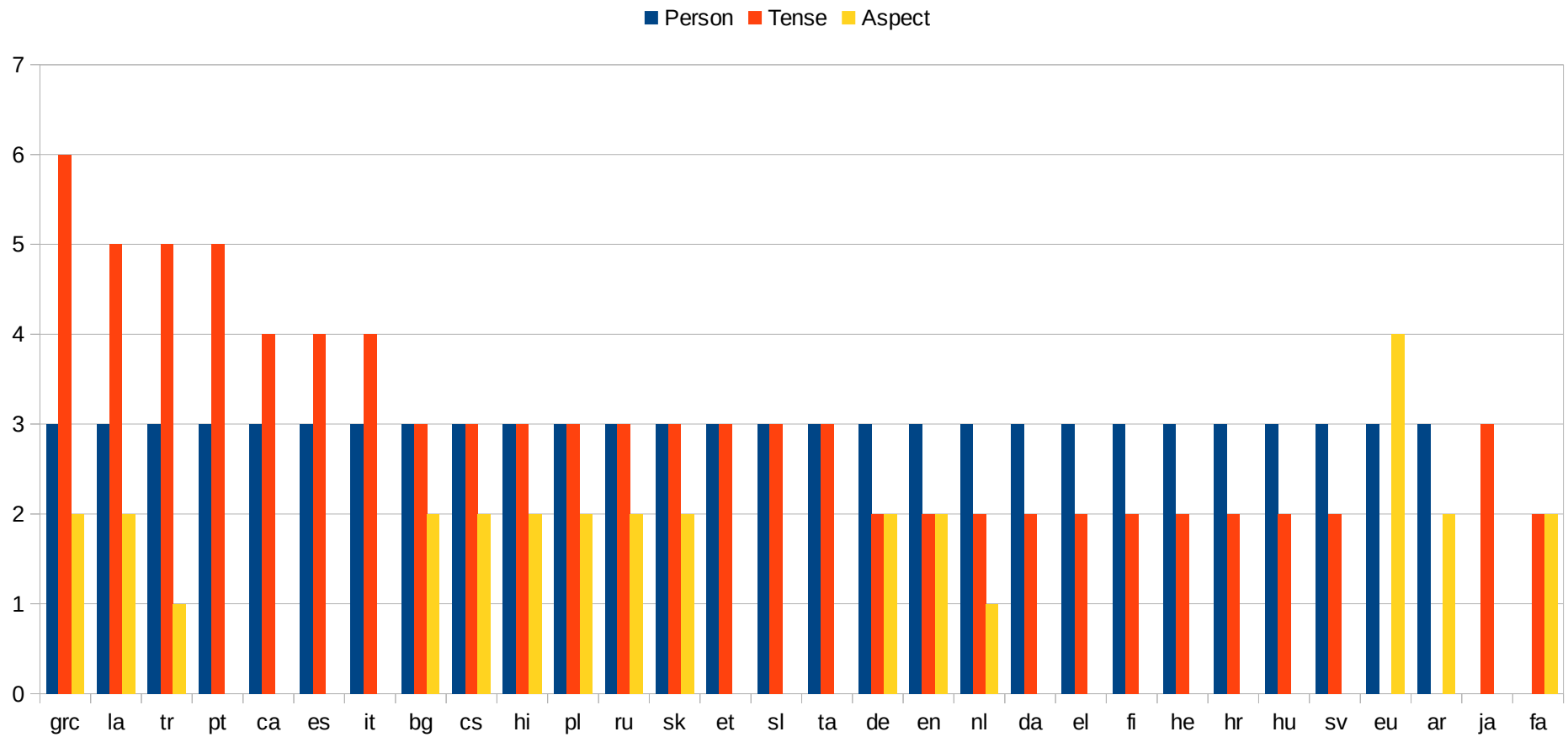
Case & Number


Case
Number
25
20
15
10
5
0
hu
et
eu
fi
ru
tr
cs
hi
hr
la
pl
sk
ta
sl
grc
el
bg
de
nl
ar
he
ca
da
es
fa
pt
sv
en
it

VerbForm & Mood & Voice



Person & Tense & Aspect



Lingua::Interset

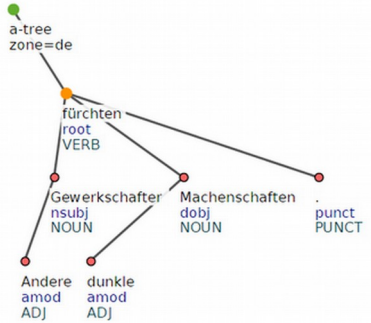
- Interset is a Perl library, available from CPAN:
  - **`cpanm Lingua::Interset`**
- Currently covers **60** tagsets of **37 languages**
- Conversion between any two tagsets:
  - simple Perl script (a few lines of code)

Universal Features

- October 2014: Universal Dependencies guidelines
- Universal POS tags
  - originally 12 Google tags, extended to 17 UPOS tags
- Universal Features
  - from Interset (subset), only cosmetic changes
  - 17 features (lexical and inflectional), 103 values so far
- Approximate conversion tables from Interset tagsets to UPOS + UFeatures are available
  http://universaldependencies.github.io/docs/u/feat/index.html

Outline

- Cross-language learning (historical motivation)
- Normalization: morphology
- Normalization: dependencies
- Cross-language learning (current work)



HamleDT

=

HArmonized

Multi-LanguagE

Dependency Treebank

HamleDT 1.0

- 2011: first version available, 29 treebanks
- ~ one third freely redistributable
- ~ one third easily obtainable + transformation by us
- ~ one third hard to get

- Morphology: Interset features and → Prague tags
- Syntax: Prague-style trees and labels

Google Universal Treebanks

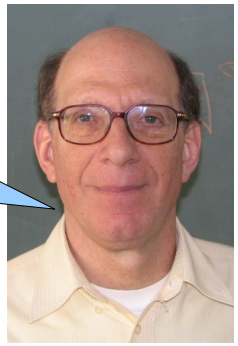
- Version 1, 2013, 6 languages
- Version 2, 2014, 11 languages
- Stanford dependencies
- Google universal POS tags

- **Another common standard?**

Google Universal Treebanks

- Version 1, 2013, 6 languages
- Version 2, 2014, 11 languages
- Stanford dependencies
- Google universal POS tags

"The nice thing about standards is that you have so many to choose from."
(Andy Tannenbaum)



HamleDT 2.0

- May 2014: version 2.0 available, 30 treebanks
- ~ one third freely redistributable
- ~ one third easily obtainable + transformation by us
- ~ one third hard to get

- Morphology: Interset features and → Google UPOS
- Syntax: added Universal Stanford Dependencies
  - Stanford and Prague were the two most widely used standards

Universal Dependencies

- Joint effort by a growing crowd of people

- Universal POS tags
- Universal Features (from Interset)
- Dependency relations (modified Stanford)
- Language-specific extensions
  - or even treebank-specific

Universal Dependencies

- The guidelines 1.0 in October 2014

- UD 1.0: 10 treebanks in January 2015
- UD 1.1: 19 treebanks in May 2015
  - All freely redistributable!
  - Some of them currently lack morphology (lemmas, features)
- Next release in November 2015

- Conversions of old data
- Newly annotated data (hu, hr, …?)

Coming Soon: HamleDT 3.0

- Superset of UD 1.1 (18 languages, 19 treebanks)
- Adds 18 more languages (automatically converted using older HamleDT transformations)
- Total: 36 languages, over 40 treebanks in the UD style

http://ufal.mff.cuni.cz/hamledt

Universal Dependencies

Don’t annotate the same thing different ways!

Universal Dependencies

Don’t annotate the same thing different ways!

Don’t make different things look the same!

Universal Dependencies

Don’t annotate the same thing different ways!

Don’t make different things look the same!

Don’t annotate things that are not there!

Structural Variations

- Pre/postpositions
- Subordinate clauses
- Verb groups
- Coordination
- Apposition

We try to automatically identify these constructions and transform them to the common style.

Structural Variations

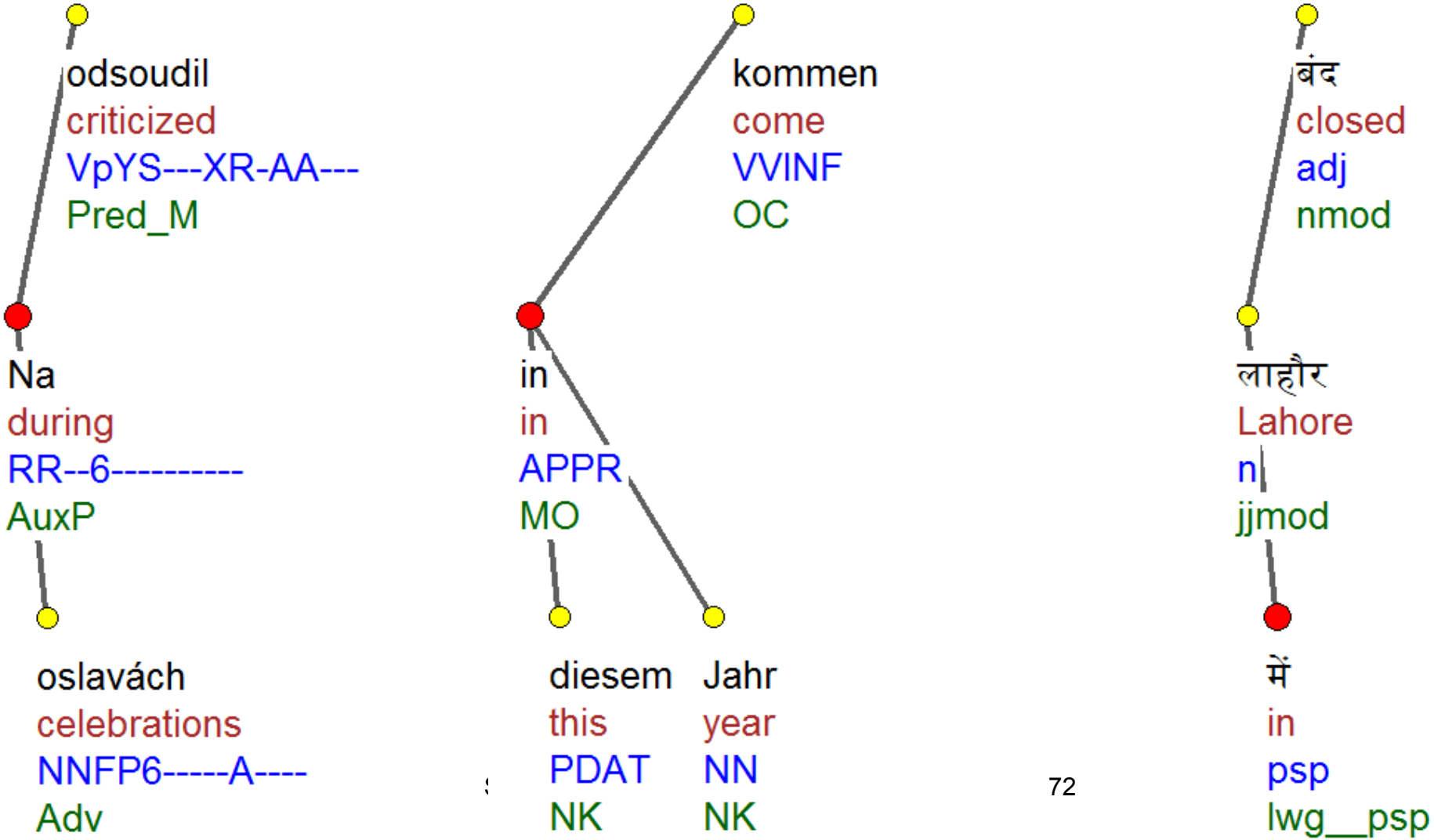
- Pre/po tpositions
- Subo
- Verb
- Coord
- Appos

Content words are heads whenever possible!

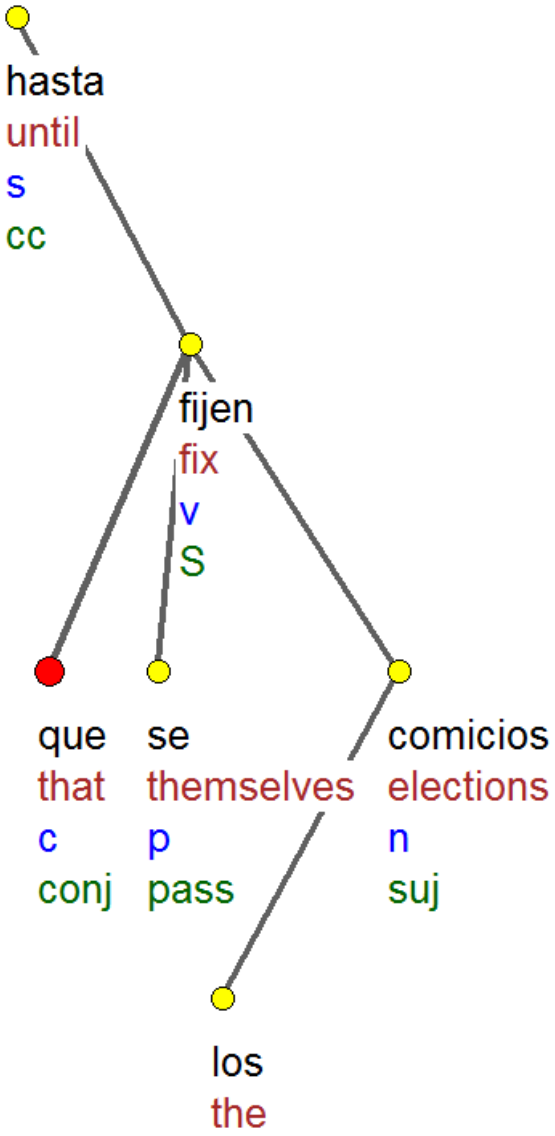
s

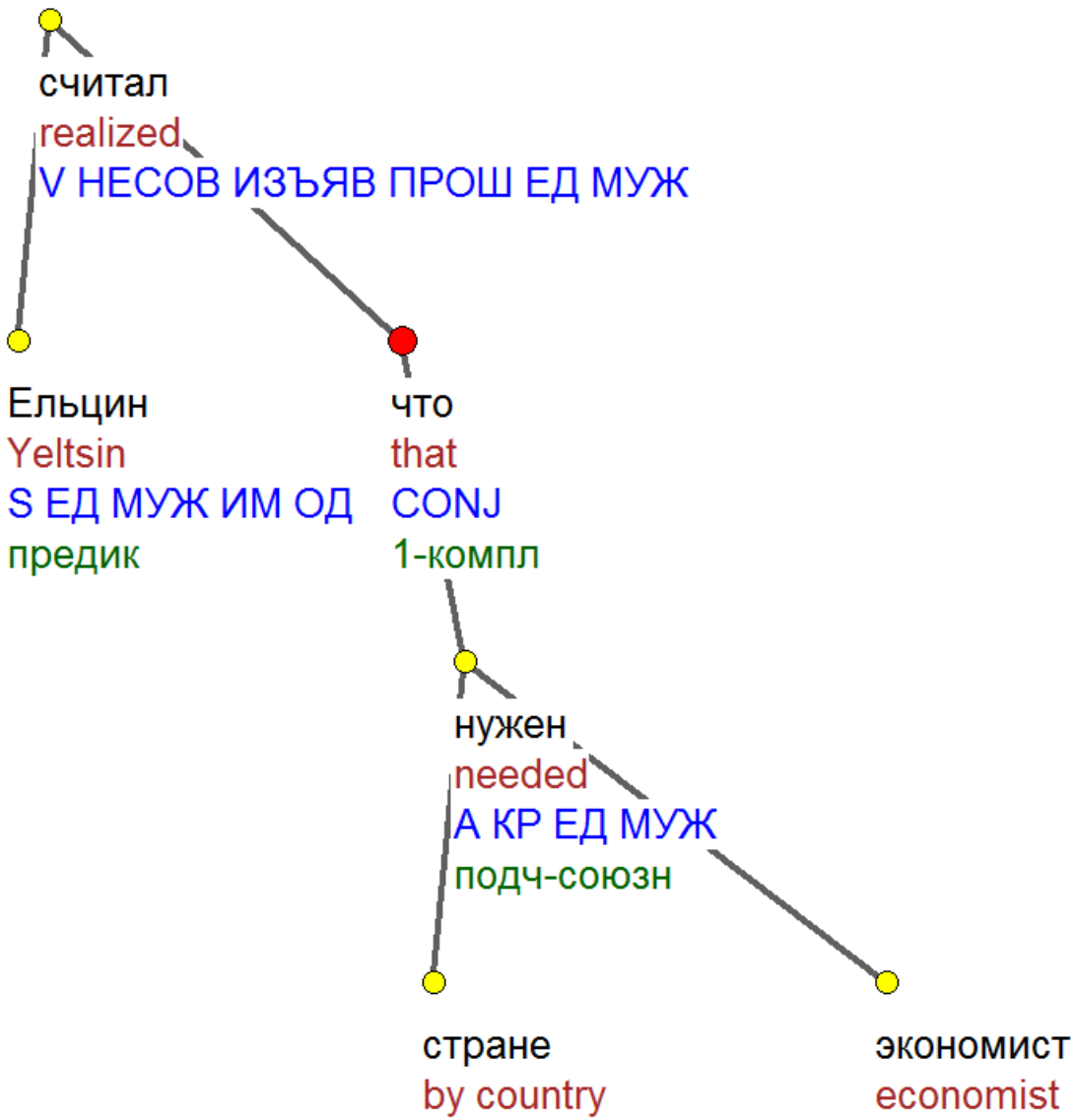
common style.

Prepositions

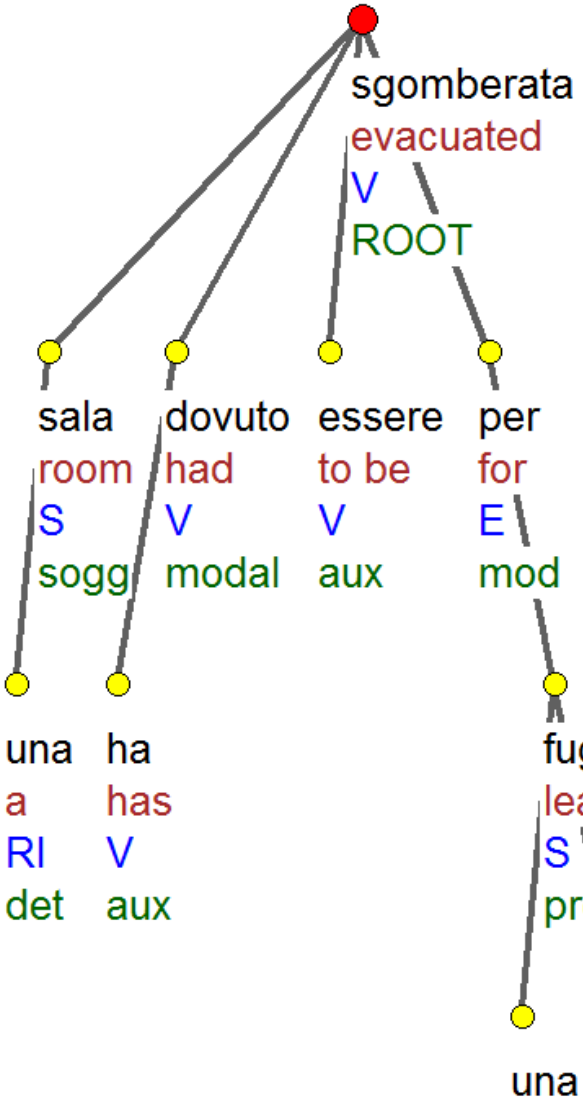


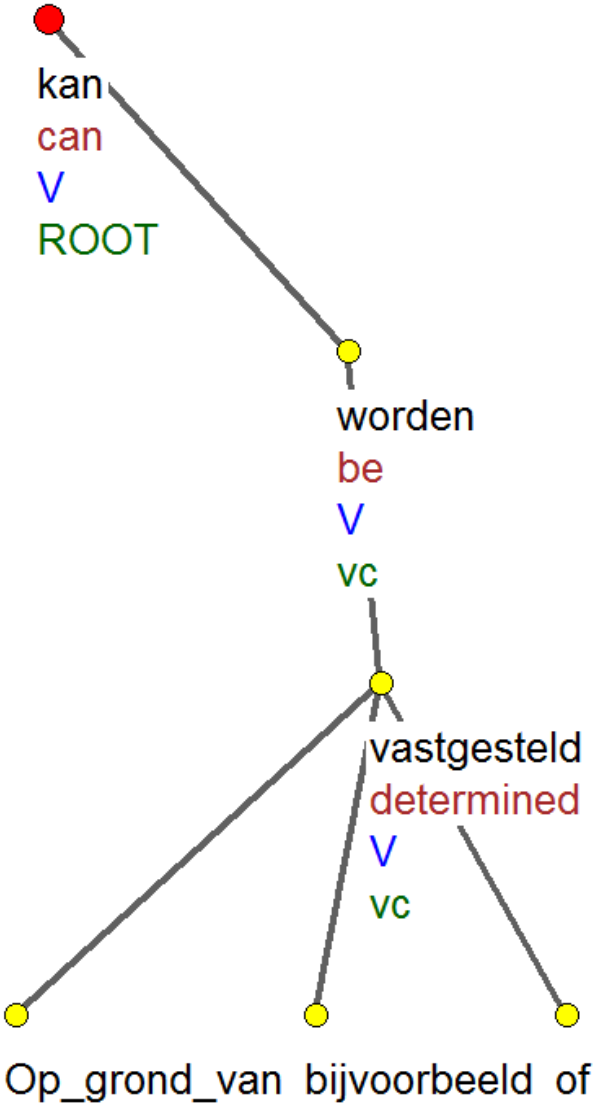
Subordinate Clauses



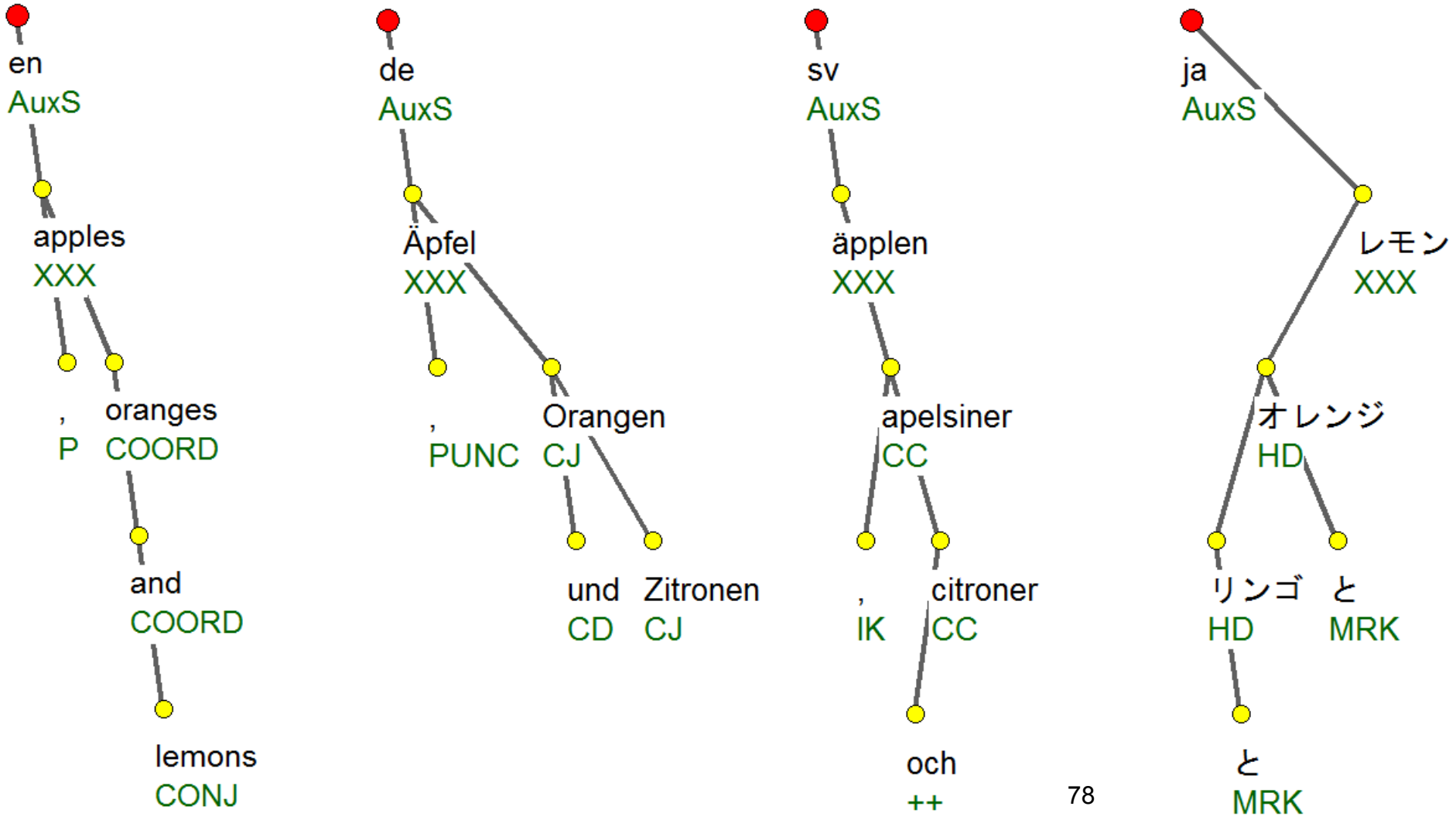


Verb Groups

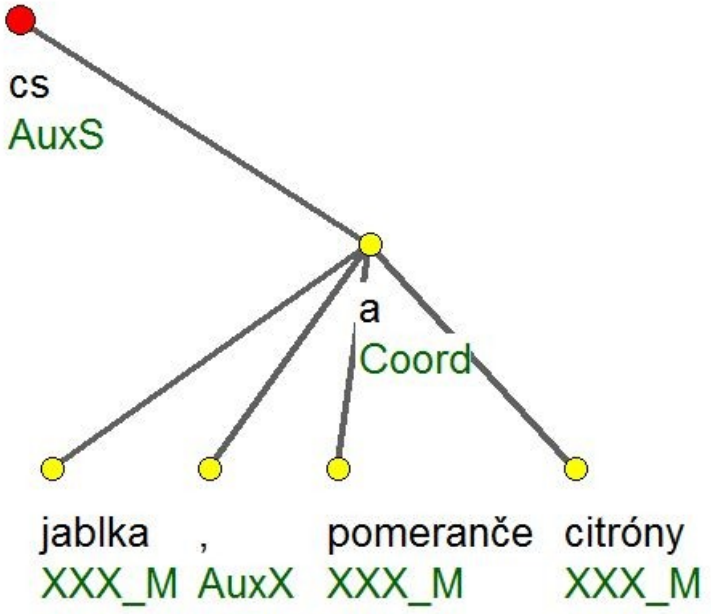

sgomberata
evacuated
V
ROOT
sala
room
S
sogg
dovuto
had
V
modal
essere
to be
V
aux
per
for
E
mod
una
a
RI
det
ha
has
V
aux
fuga
leak
S
prep
una
di

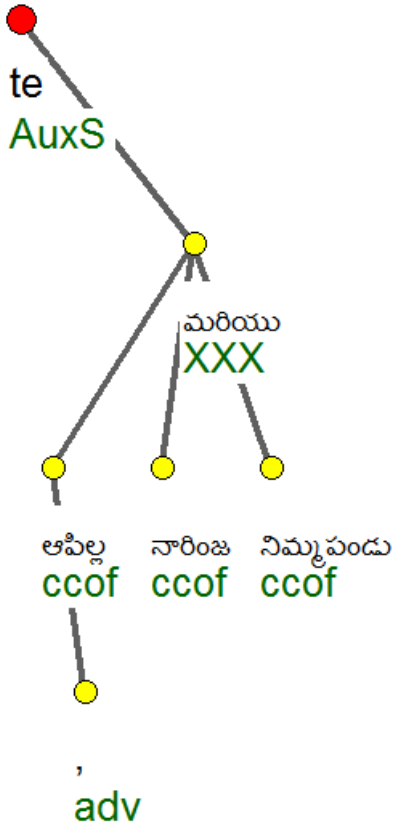

kan
can
V
ROOT
worden
be
V
vc
vastgesteld
determined
V
vc
Op_grond_van
bijvoorbeeld
of

Coordination: Mel'čuk

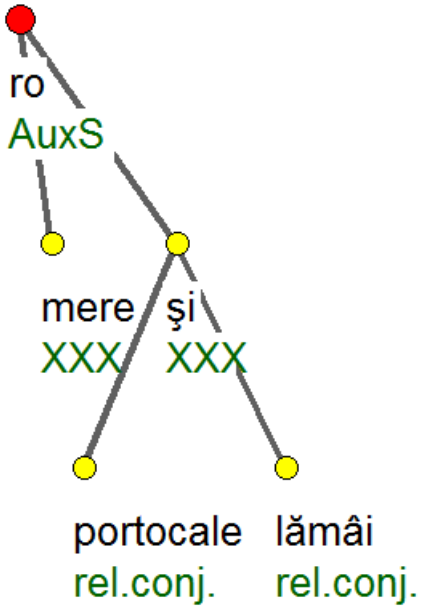


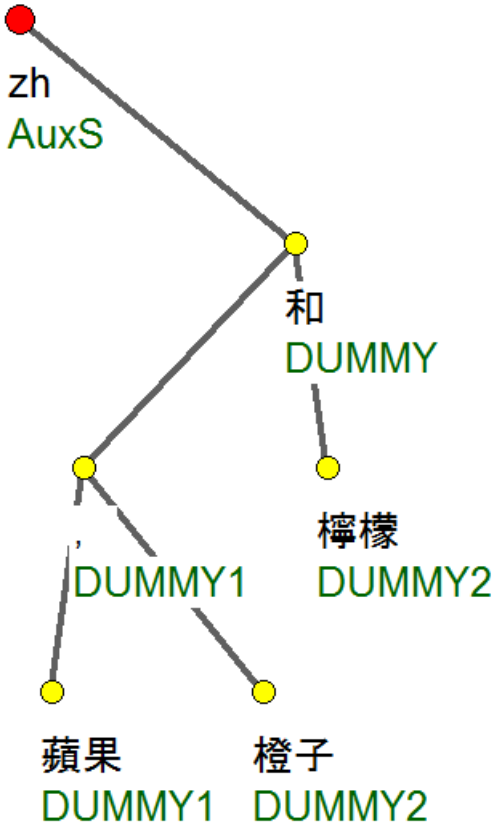
Coordination: Prague



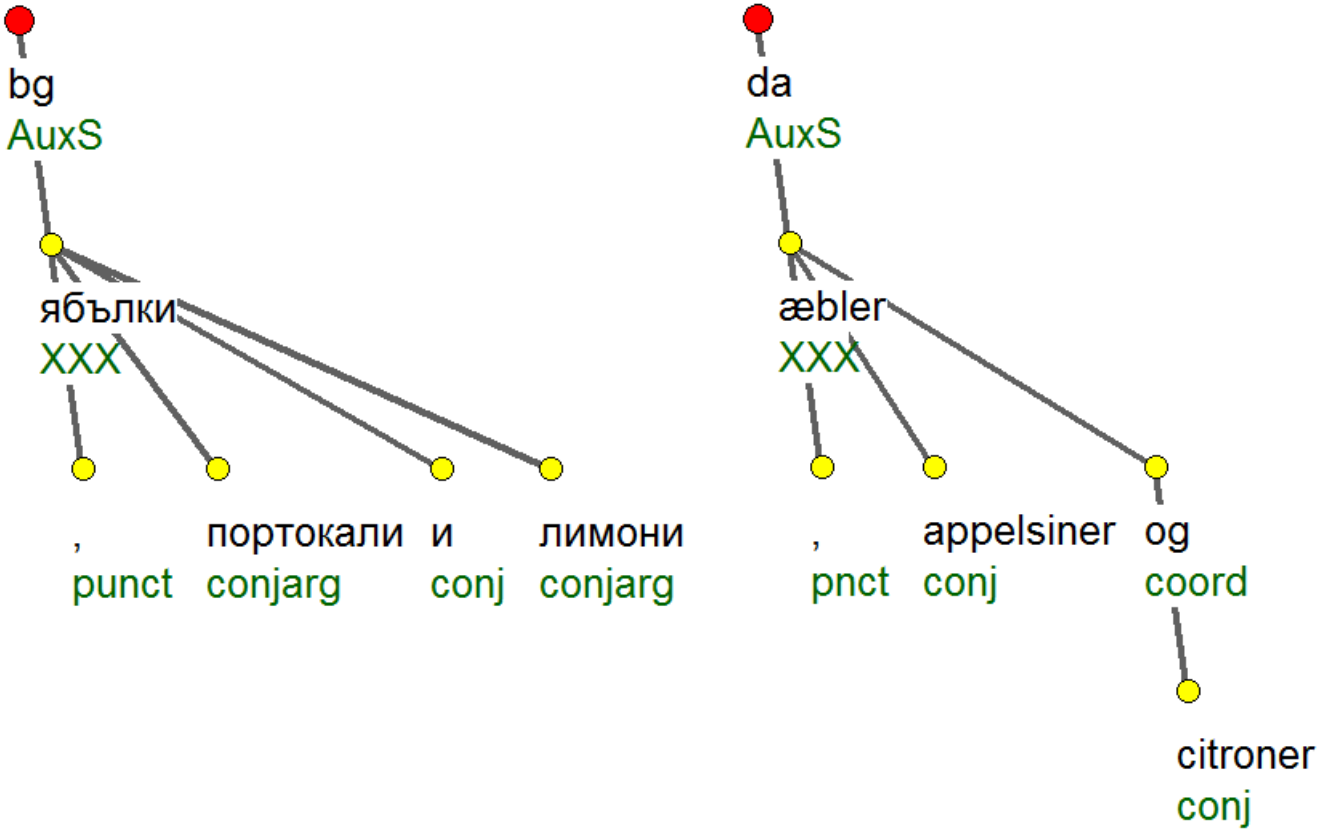


Coordination: [ro, zh]

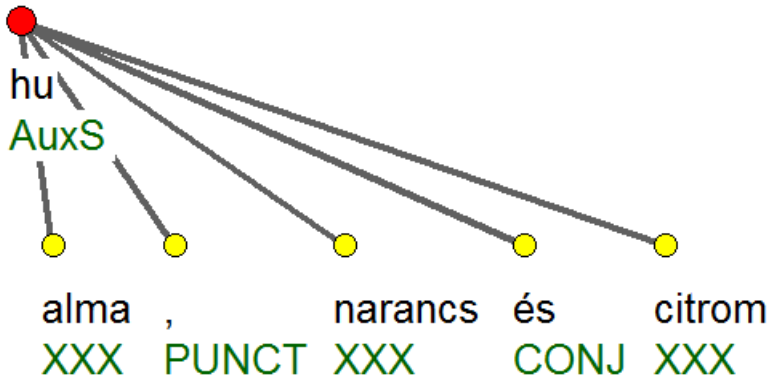




Coordination: Stanford



Coordination: Tesnière



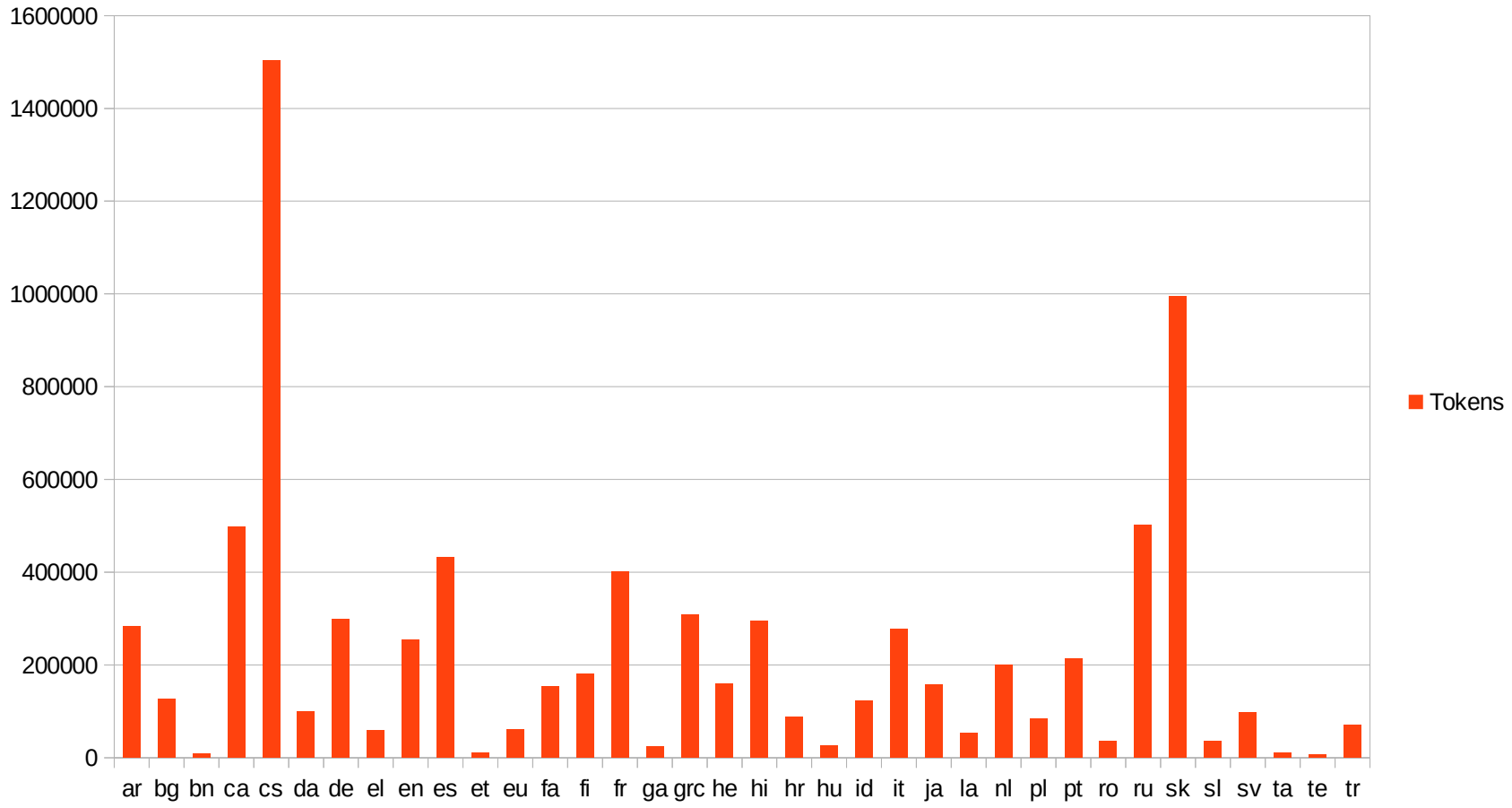
36 Languages (HamleDT 3.0)

- Ancient Greek (grc)
- Arabic (ar)
- Basque (eu)
- Bengali (bn)
- Bulgarian (bg)
- Catalan (ca)
- Croatian (hr)
- Czech (cs)
- Danish (da)
- Dutch (nl)
- English (en)
- Estonian (et)
- Finnish (fi)
- French (fr)
- German (de)
- Greek (el)
- Hebrew (he)
- Hindi (hi)
- Hungarian (hu)
- Indonesian (id)
- Irish (ga)
- Italian (it)
- Japanese (ja)
- Latin (la)
- Persian (fa)
- Polish (pl)
- Portuguese (pt)
- Romanian (ro)
- Russian (ru)
- Slovak (sk)
- Slovene (sl)
- Spanish (es)
- Swedish (sv)
- Tamil (ta)
- Telugu (te)
- Turkish (tr)

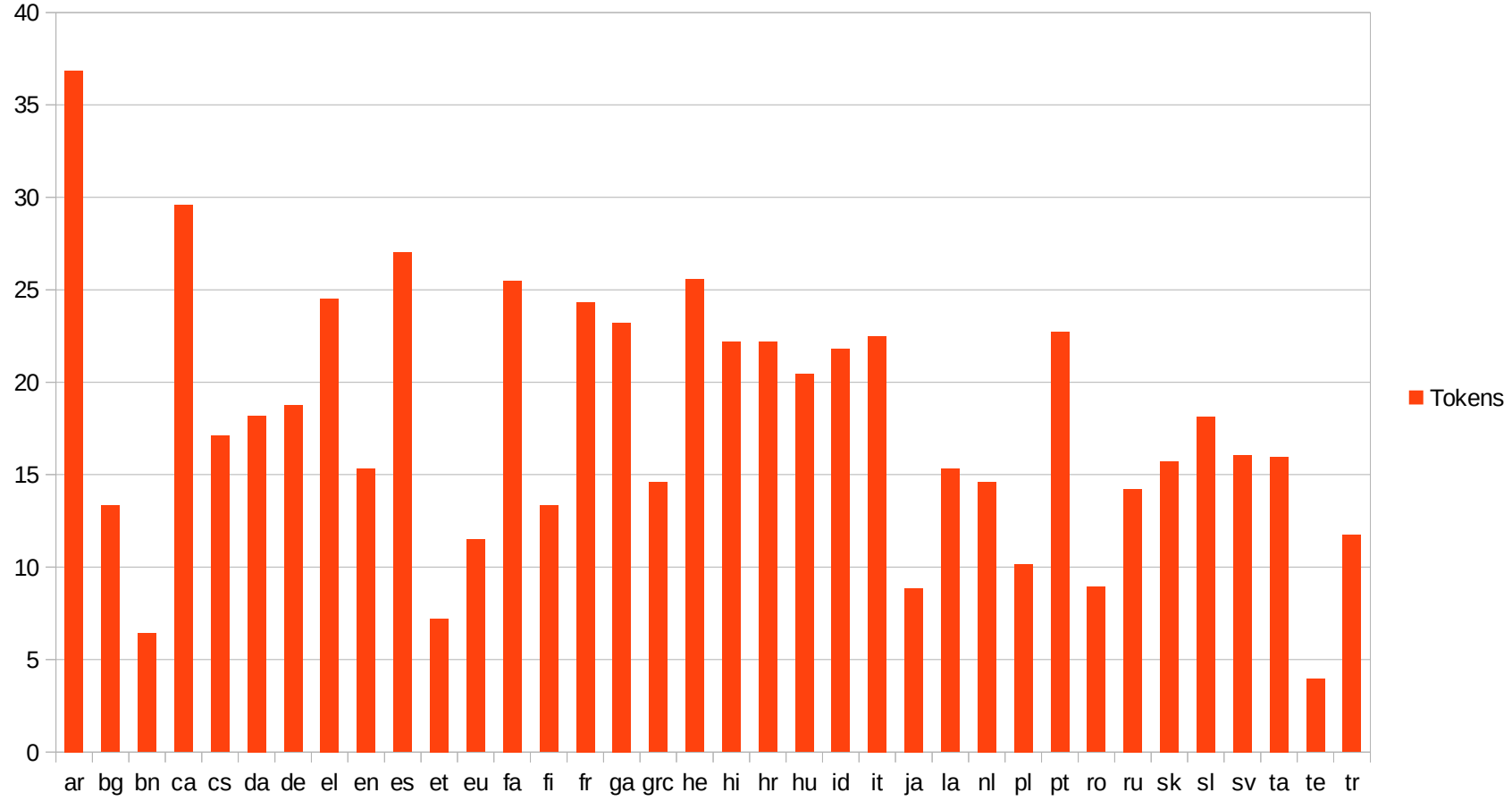
UD 1.1 (May 2015): 18

- Ancient Greek (grc)
- Arabic (ar)
- Basque (eu)
- Bengali (bn)
- Bulgarian (bg)
- Catalan (ca)
- Croatian (hr)
- Czech (cs)
- Danish (da)
- Dutch (nl)
- English (en)
- Estonian (et)
- Finnish (fi)
- French (fr)
- German (de)
- Greek (el)
- Hebrew (he)
- Hindi (hi)
- Hungarian (hu)
- Indonesian (id)
- Irish (ga)
- Italian (it)
- Japanese (ja)
- Latin (la)
- Persian (fa)
- Polish (pl)
- Portuguese (pt)
- Romanian (ro)
- Russian (ru)
- Slovak (sk)
- Slovene (sl)
- Spanish (es)
- Swedish (sv)
- Tamil (ta)
- Telugu (te)
- Turkish (tr)

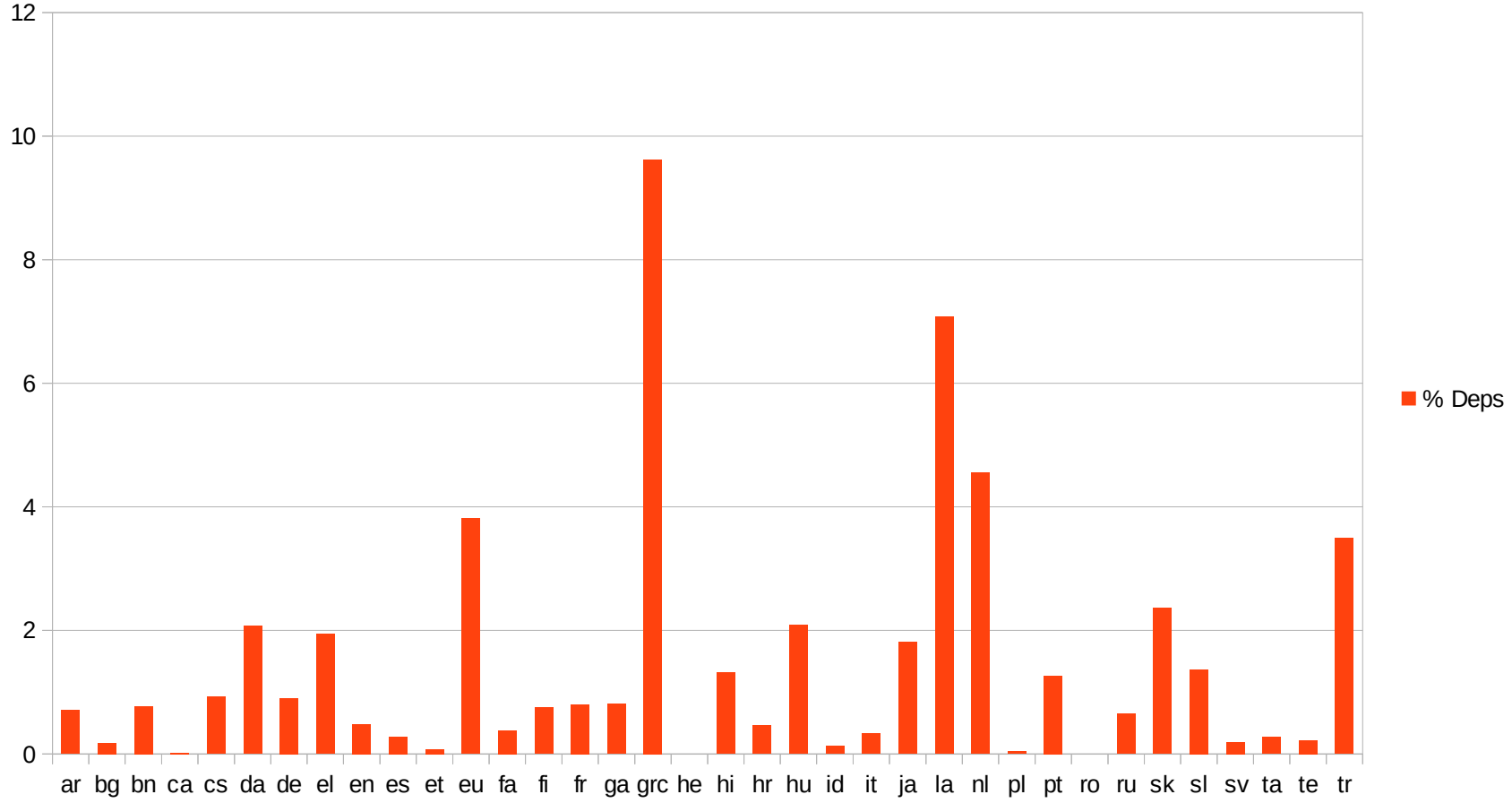
Data Size


1600000
1400000
1200000
1000000
800000
600000
400000
200000
0
ar bg bn ca cs da de el en es et eu fa fi fr ga grc he hi hr hu id it ja la nl pl pt ro ru sk sl sv ta te tr
Tokens

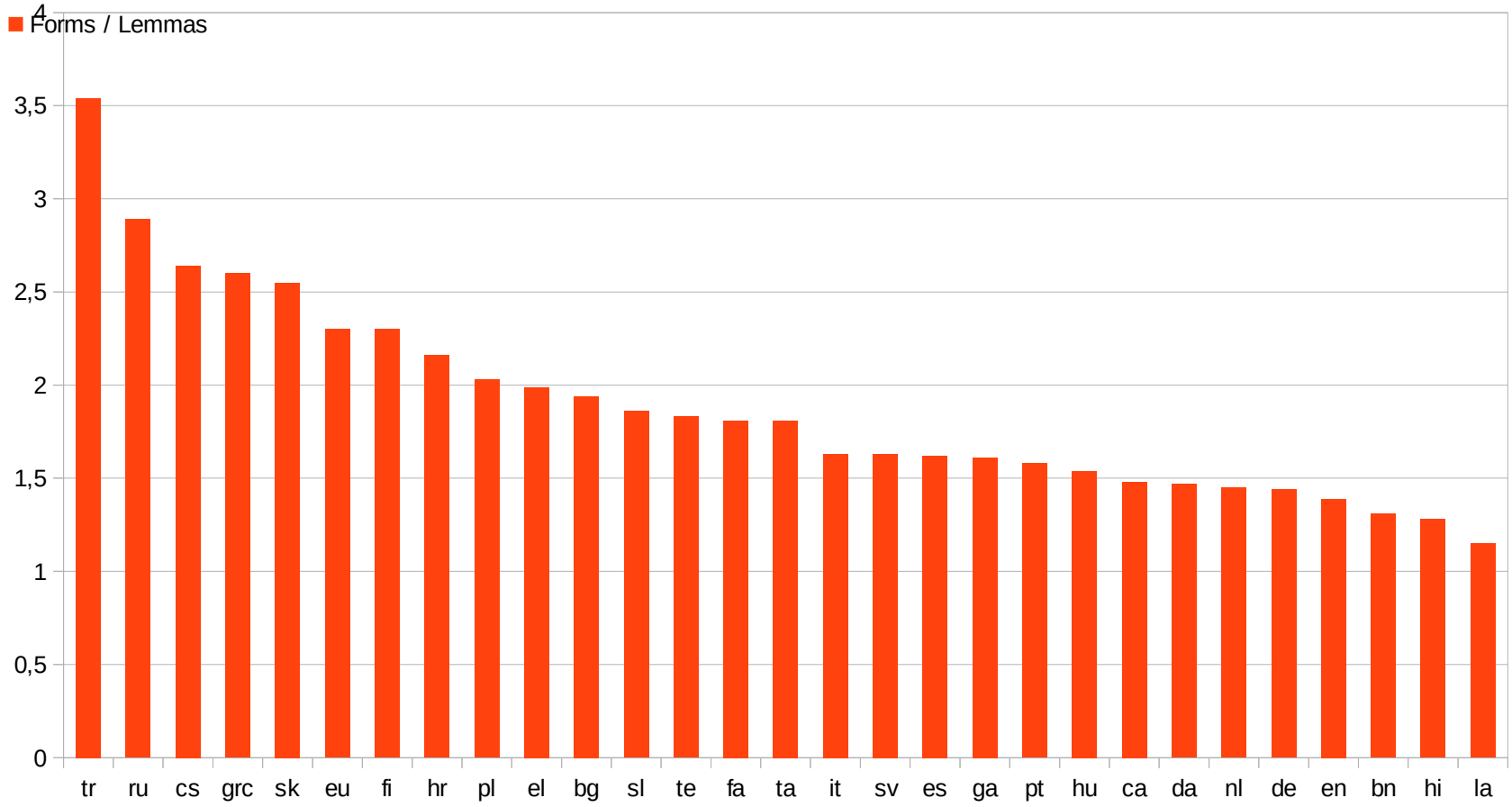
Sentence Length



Nonprojective Dependencies



Morphological Richness



How Can You Get It?

- 36 languages in the UD style
- 28 directly downloadable
- patches and/or free software for the rest (if you have the data)

- Stay tuned to:

http://ufal.mff.cuni.cz/hamledt/

(HamleDT 3.0 should be available before the end of summer.)

Outline

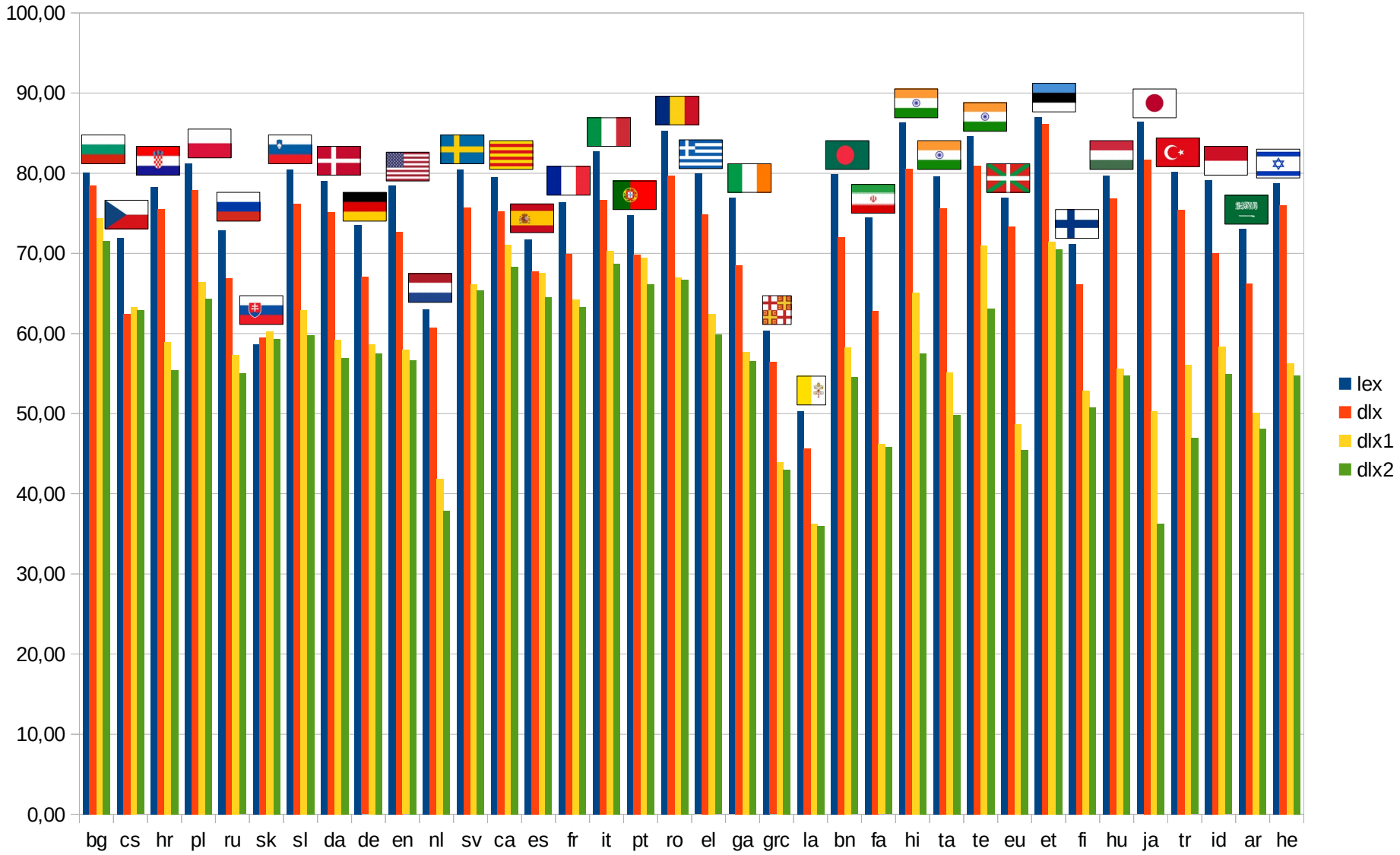
- Cross-language learning (historical motivation)
- Normalization: morphology
- Normalization: dependencies
- Cross-language learning (current work)

Default Setup

- Malt Parser, stack-lazy algorithm
  - same configuration for all, no optimization
  - same selection of training features for all treebanks

- Trained on the first 1000 sentences only
- Tested on the whole test set
- Default score: UAS
- Only harmonized data used

Malt Trained on 1000 Sents.



Who Helps Whom?

- Czech (62.44) ⇦ Croatian (63.27), Slovene (62.87)
- Slovak (59.47) ⇦ Croatian (60.28), Slovene (59.32)
- Polish (77.92) ⇦ Croatian (66.42), Slovene (64.31)
- Russian (66.86) ⇦ Croatian (57.35), Slovak (55.01)
- Croatian (75.52) ⇦ Slovene (58.96), Polish (55.42)
- Slovene (76.17) ⇦ Croatian (62.92), Finnish (59.79)
- Bulgarian (78.44) ⇦ Croatian (74.39), Slovene (71.52)

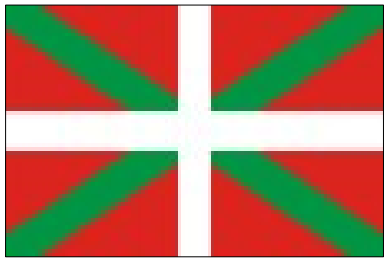
Who Helps Whom?

- Catalan (75.28) ⇦ Italian (71.07), French (68.30)
- Italian (76.66) ⇦ French (70.37), Catalan (68.66)
- French (69.93) ⇦ Spanish (64.28), Italian (63.33)
- Spanish (67.76) ⇦ French (67.61), Catalan (64.54)
- Portuguese (69.89) ⇦ Italian (69.48), French (66.12)
- Romanian (79.74) ⇦ Croatian (67.01), Latin (66.75)

Who Helps Whom?

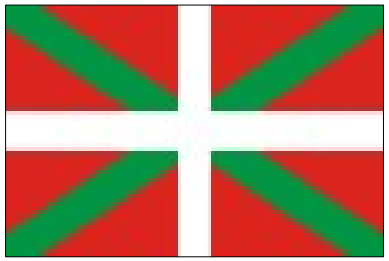
- Swedish (75.73) ⇦ Danish (66.17), English (65.41)
- Danish (75.19) ⇦ Swedish (59.23), Croatian (56.89)
- English (72.68) ⇦ German (57.95), French (56.70)
- German (67.04) ⇦ Croatian (58.68), Swedish (57.48)
- Dutch (60.76) ⇦ Hungarian (41.90), Finnish (37.89)

Who Helps Basque?



Who Helps Basque?

- Basque (73.36) ⇦
  - Hungarian (48.72)
  - Estonian (45.49)
  - Croatian (44.37)



- Basque ⇨
  - Hindi (54.89); best is Tamil (65.07)
  - Tamil (49.58); best is Hindi (55.11)

Morphological Features: Do They Help?

- The SVM learner discriminates between useful and useless features.

BUT!

- What if the target data lack the “useful” features?
  - (What if they lack all features, e.g. UD1.1 German, French, Spanish, Indonesian?)

Feature Ranking

1. All features
2. Lex features (PronType, NumType, Poss, Reflex)
3. No features (only part of speech tag)
4. Lex + Person
5. Only Person
6. Only Case
7. Lex + Case
8. …

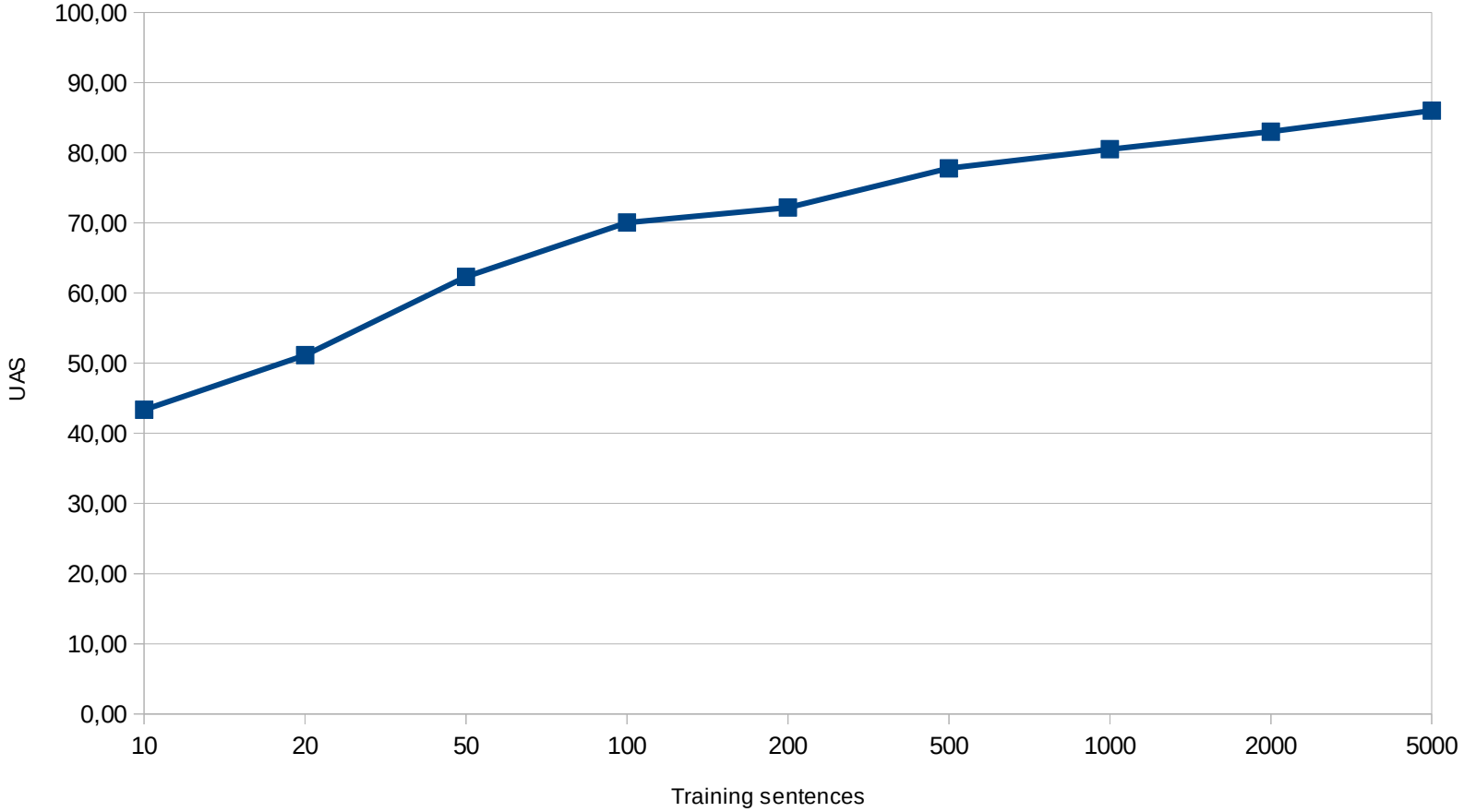
Some Exceptions

- None: fr ← it, fi ← hu
- Lexical: hr ← bg, he ← it, et ← hu, da ← sv
- Lexical + Tense + Aspect + Mood + Voice: eu ← hu
- VerbForm + Person: ga ← he

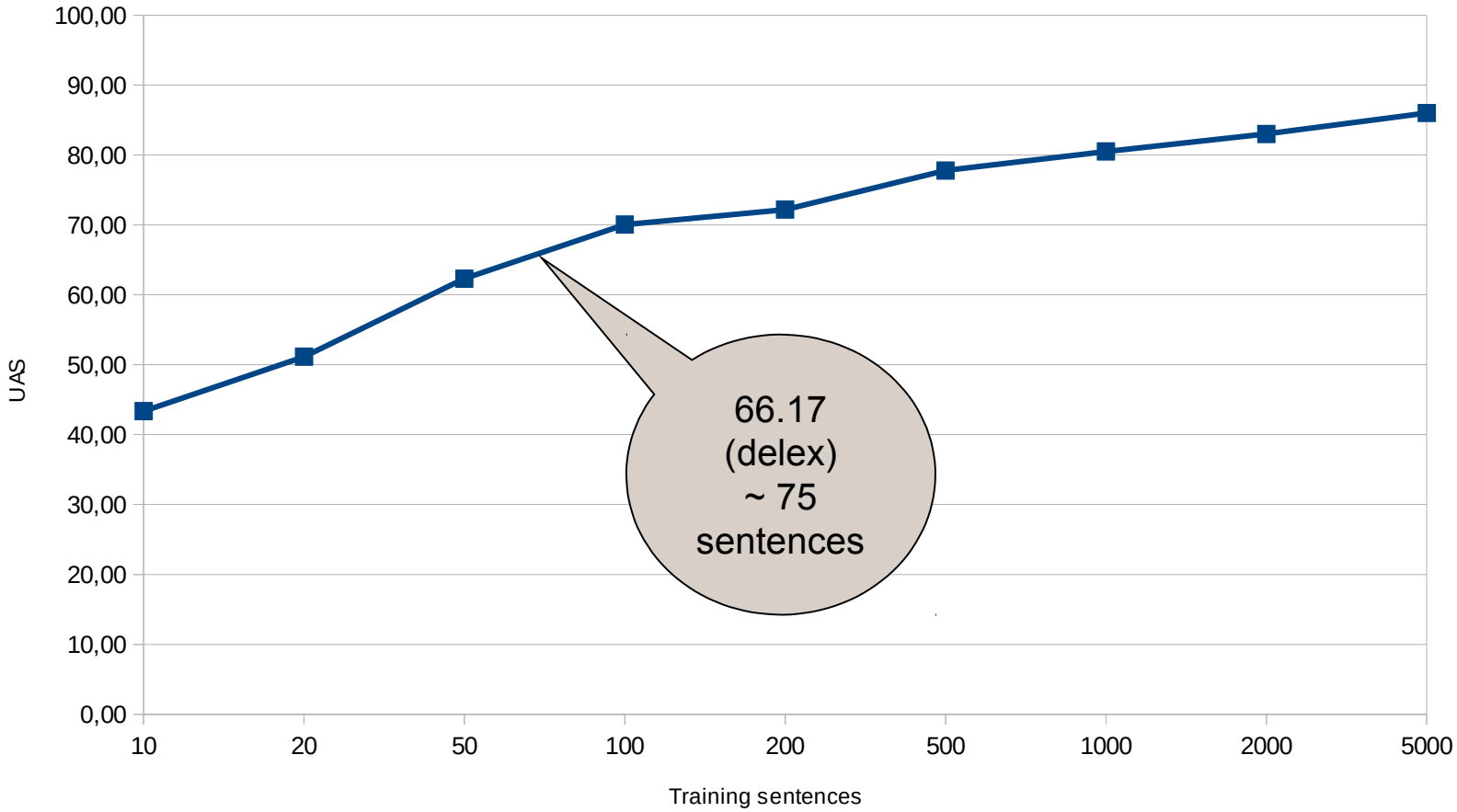
So What's Next?

- Ongoing work — preliminary results
- Unsupervised target POS + morphology
- Other settings of Malt parser, other parsers
- Combination of source languages; prediction of the best source language(s)
  - cf. Rudolf Rosa's talk from yesterday

Back at the Start



Back at the Start




mulțumesc
gratias
tak
শুক্রিয়া
danke
teşekkür ederim
謝謝
спасибо
gràcies
děkujeme
ధన్యవాదాలు
благодаря
grazie
dank
thank you
köszönöm
شكرا
hvala
তোমাকে
ধন্যবাদ
kiitos
gracias
obrigado
நன்றி
ευχαριστώ
ありがとう
tack
aitäh
bakka bér
eskerrik asko