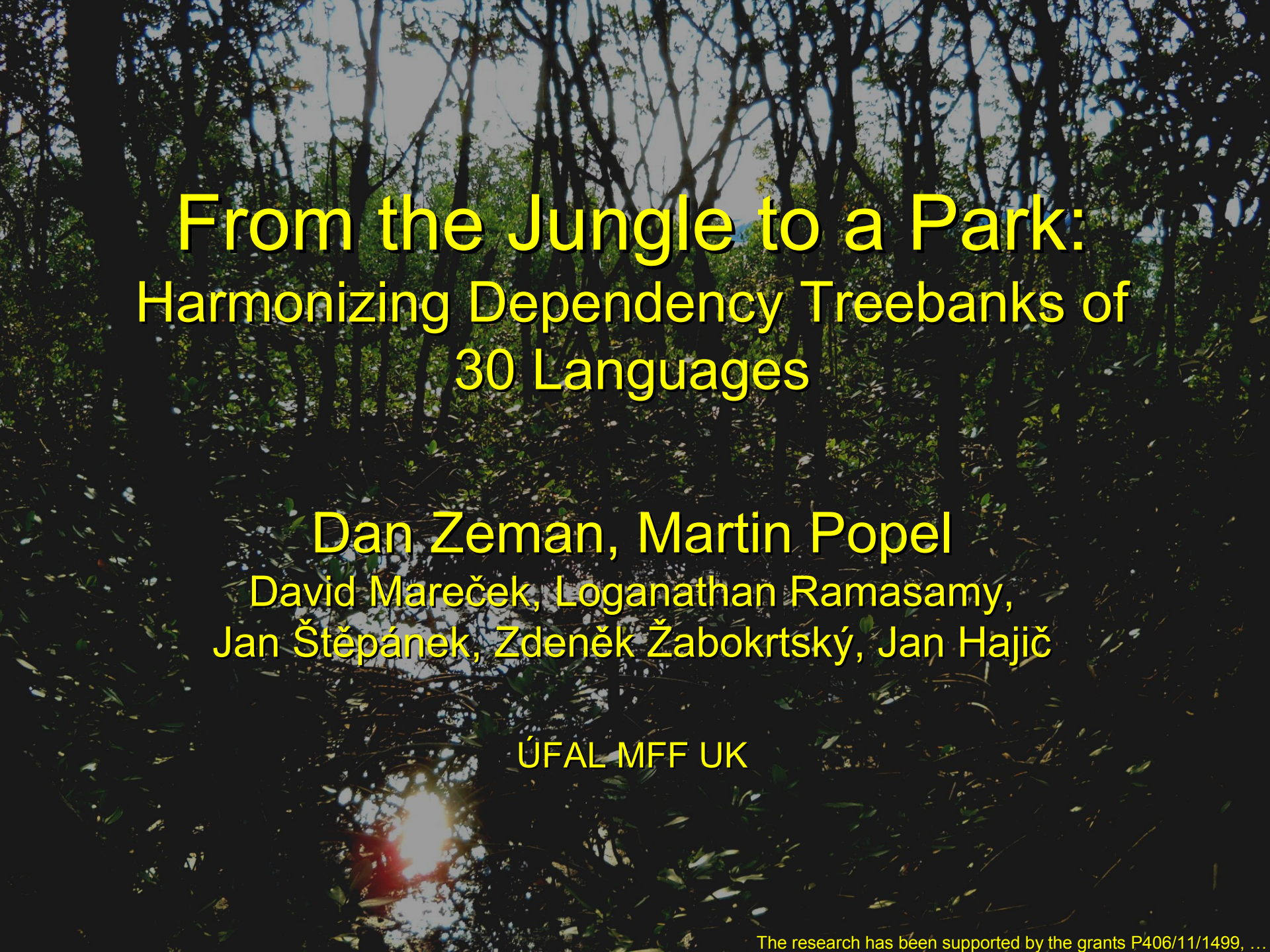


From the Jungle to a Park:
Harmonizing Dependency Treebanks of 30 Languages

Dan Zeman, Martin Popel
David Mareček, Loganathan Ramasamy,
Jan Štěpánek, Zdeněk Žabokrtský, Jan Hajič

ÚFAL MFF UK

The research has been supported by the grants P406/11/1499, …

Results

	McD	Niv	O'N	Rie	Sag	Che	Cor	Cha	Joh
ar	66.9	66.7	66.7	62.7	65.2	65.2	63.5	60.9	64.3
zh	85.9	86.9	86.7	90.0	84.7	84.3	79.9	85.1	72.5
cs	80.2	78.4	76.6	67.4	75.2	76.2	74.5	72.9	71.5
da	84.8	84.8	82.8	83.6	81.6	81.7	81.7	80.6	81.5
nl	79.2	78.6	77.5	78.6	76.6	71.8	71.4	72.9	72.7
de	87.3	85.8	85.4	86.2	84.9	84.1	83.5	84.2	80.4
ja	90.7	91.7	90.6	90.5	90.4	89.9	90.0	89.1	85.6
pt	86.8	87.6	84.7	84.4	86.0	85.1	84.6	84.0	84.6
sl	73.4	70.3	71.1	71.2	69.1	71.4	72.4	69.5	66.4
es	82.3	81.3	79.8	77.4	77.7	80.5	80.4	79.7	78.2
sv	82.6	84.6	81.8	80.7	82.0	81.1	79.7	82.3	78.1
tr	63.2	65.7	57.5	58.6	63.2	61.2	61.7	60.5	63.4

That Was CoNLL-X

- Does it mean that
  - Turkish (63.2) and Arabic (66.9) are difficult *languages;*
  - German (87.3) and Japanese (90.1) are easy *languages?*
- Not necessarily…
  - Data size (<1000 to >50000 training sentences)
  - Chance (small test set, big deviation)
  - Language differences
  - Domain differences (=> sentence length)
  - Annotation style differences

The TASK

- As many languages as possible
- Find style differences
- Unify styles

  => *normalized treebank*
- Try alternatives to the unified style

  => *transformations*

- What is the best for your parser?

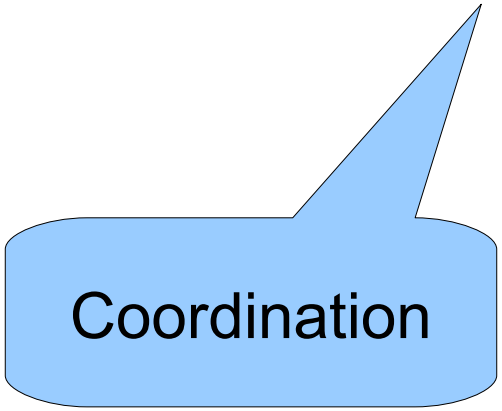
The TALK

Dan:

- Data overview
- Normalization

Martin:

- Transformation overview
- Experiments



Languages

- Ancient Greek (grc)
- Arabic (ar)
- Basque (eu)
- Bengali (bn)
- Bulgarian (bg)
- Catalan (ca)
- Chinese (zh)
- Czech (cs)
- Danish (da)
- Dutch (nl)
- English (en)
- Estonian (et)
- Finnish (fi)
- German (de)
- Greek (el)
- Hebrew (he)
- Hindi (hi)
- Hungarian (hu)
- Icelandic (is)
- Italian (it)
- Japanese (ja)
- Latin (la)
- Portuguese (pt)
- Romanian (ro)
- Russian (ru)
- Slovene (sl)
- Spanish (es)
- Swedish (sv)
- Tamil (ta)
- Telugu (te)
- Turkish (tr)

CoNLL-X: 13

- Ancient Greek (grc)
- Arabic (ar)
- Basque (eu)
- Bengali (bn)
- Bulgarian (bg)
- Catalan (ca)
- Chinese (zh)
- Czech (cs)
- Danish (da)
- Dutch (nl)
- English (en)
- Estonian (et)
- Finnish (fi)
- German (de)
- Greek (el)
- Hebrew (he)
- Hindi (hi)
- Hungarian (hu)
- Icelandic (is)
- Italian (it)
- Japanese (ja)
- Latin (la)
- Portuguese (pt)
- Romanian (ro)
- Russian (ru)
- Slovene (sl)
- Spanish (es)
- Swedish (sv)
- Tamil (ta)
- Telugu (te)
- Turkish (tr)

CoNLL 2007: 10

- Ancient Greek (grc)
- Arabic (ar)
- Basque (eu)
- Bengali (bn)
- Bulgarian (bg)
- Catalan (ca)
- Chinese (zh)
- Czech (cs)
- Danish (da)
- Dutch (nl)
- English (en)
- Estonian (et)
- Finnish (fi)
- German (de)
- Greek (el)
- Hebrew (he)
- Hindi (hi)
- Hungarian (hu)
- Icelandic (is)
- Italian (it)
- Japanese (ja)
- Latin (la)
- Portuguese (pt)
- Romanian (ro)
- Russian (ru)
- Slovene (sl)
- Spanish (es)
- Swedish (sv)
- Tamil (ta)
- Telugu (te)
- Turkish (tr)

CoNLL 2009: 7

- Ancient Greek (grc)
- Arabic (ar)
- Basque (eu)
- Bengali (bn)
- Bulgarian (bg)
- Catalan (ca)
- Chinese (zh)
- Czech (cs)
- Danish (da)
- Dutch (nl)
- English (en)
- Estonian (et)
- Finnish (fi)
- German (de)
- Greek (el)
- Hebrew (he)
- Hindi (hi)
- Hungarian (hu)
- Icelandic (is)
- Italian (it)
- ~~Japanese (ja)~~
- Latin (la)
- Portuguese (pt)
- Romanian (ro)
- Russian (ru)
- Slovene (sl)
- Spanish (es)
- Swedish (sv)
- Tamil (ta)
- Telugu (te)
- Turkish (tr)

ICON 2009-2010: 3

- Ancient Greek (grc)
- Arabic (ar)
- Basque (eu)
- Bengali (bn)
- Bulgarian (bg)
- Catalan (ca)
- Chinese (zh)
- Czech (cs)
- Danish (da)
- Dutch (nl)
- English (en)
- Estonian (et)
- Finnish (fi)
- German (de)
- Greek (el)
- Hebrew (he)
- Hindi (hi)
- Hungarian (hu)
- Icelandic (is)
- Italian (it)
- Japanese (ja)
- Latin (la)
- Portuguese (pt)
- Romanian (ro)
- Russian (ru)
- Slovene (sl)
- Spanish (es)
- Swedish (sv)
- Tamil (ta)
- Telugu (te)
- Turkish (tr)

Other: 10

- Ancient Greek (grc)
- Arabic (ar)
- Basque (eu)
- Bengali (bn)
- Bulgarian (bg)
- Catalan (ca)
- Chinese (zh)
- Czech (cs)
- Danish (da)
- Dutch (nl)
- English (en)
- Estonian (et)
- Finnish (fi)
- German (de)
- Greek (el)
- Hebrew (he)
- Hindi (hi)
- Hungarian (hu)
- Icelandic (is)
- Italian (it)
- Japanese (ja)
- Latin (la)
- Portuguese (pt)
- Romanian (ro)
- Russian (ru)
- Slovene (sl)
- Spanish (es)
- Swedish (sv)
- Tamil (ta)
- Telugu (te)
- Turkish (tr)

Dependencies (PDT-like) / Converted / Constituents

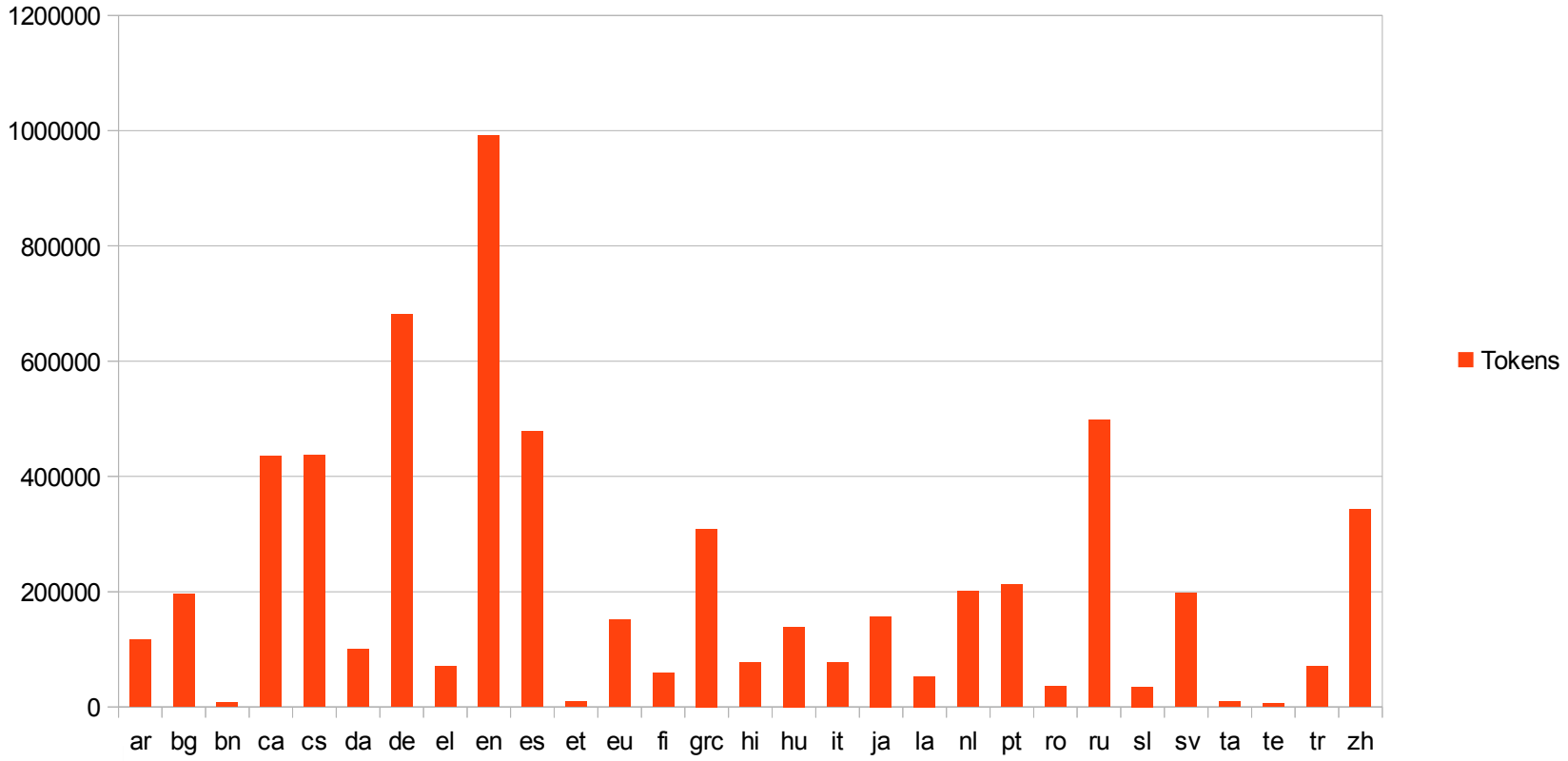
- Ancient Greek (grc)
- Arabic (ar)
- Basque (eu)
- Bengali (bn)
- Bulgarian (bg)
- Catalan (ca)
- Chinese (zh)
- Czech (cs)
- Danish (da)
- Dutch (nl)
- English (en)
- Estonian (et)
- Finnish (fi)
- German (de)
- Greek (el)
- Hebrew (he)
- Hindi (hi)
- Hungarian (hu)
- Icelandic (is)
- Italian (it)
- Japanese (ja)
- Latin (la)
- Portuguese (pt)
- Romanian (ro)
- Russian (ru)
- Slovene (sl)
- Spanish (es)
- Swedish (sv)
- Tamil (ta)
- Telugu (te)
- Turkish (tr)

We Cannot Process (yet)

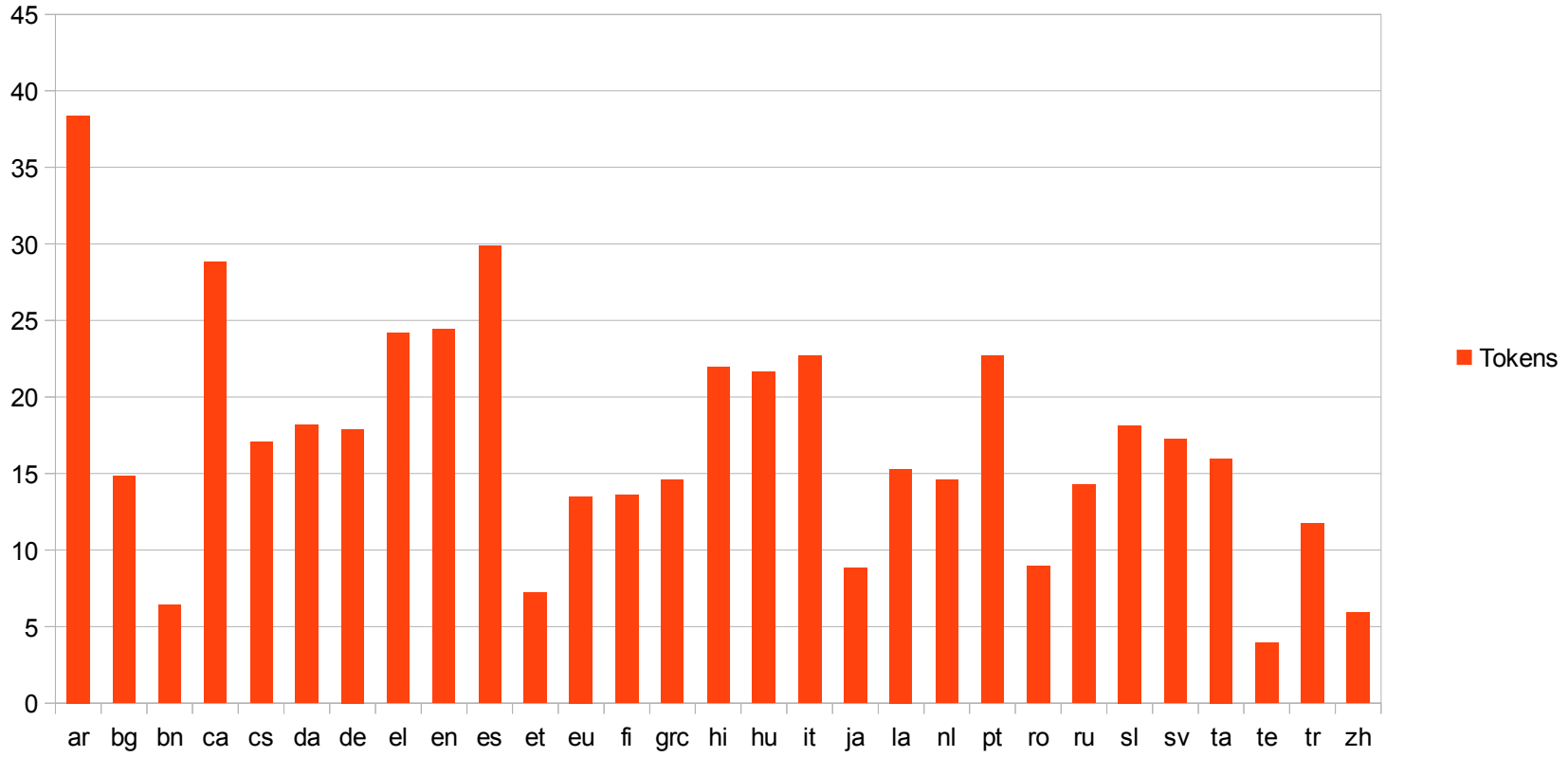
- Chinese (Sinica Treebank)
- Hebrew (constituents)
- Icelandic (constituents)

- The others:
  - Various levels of success

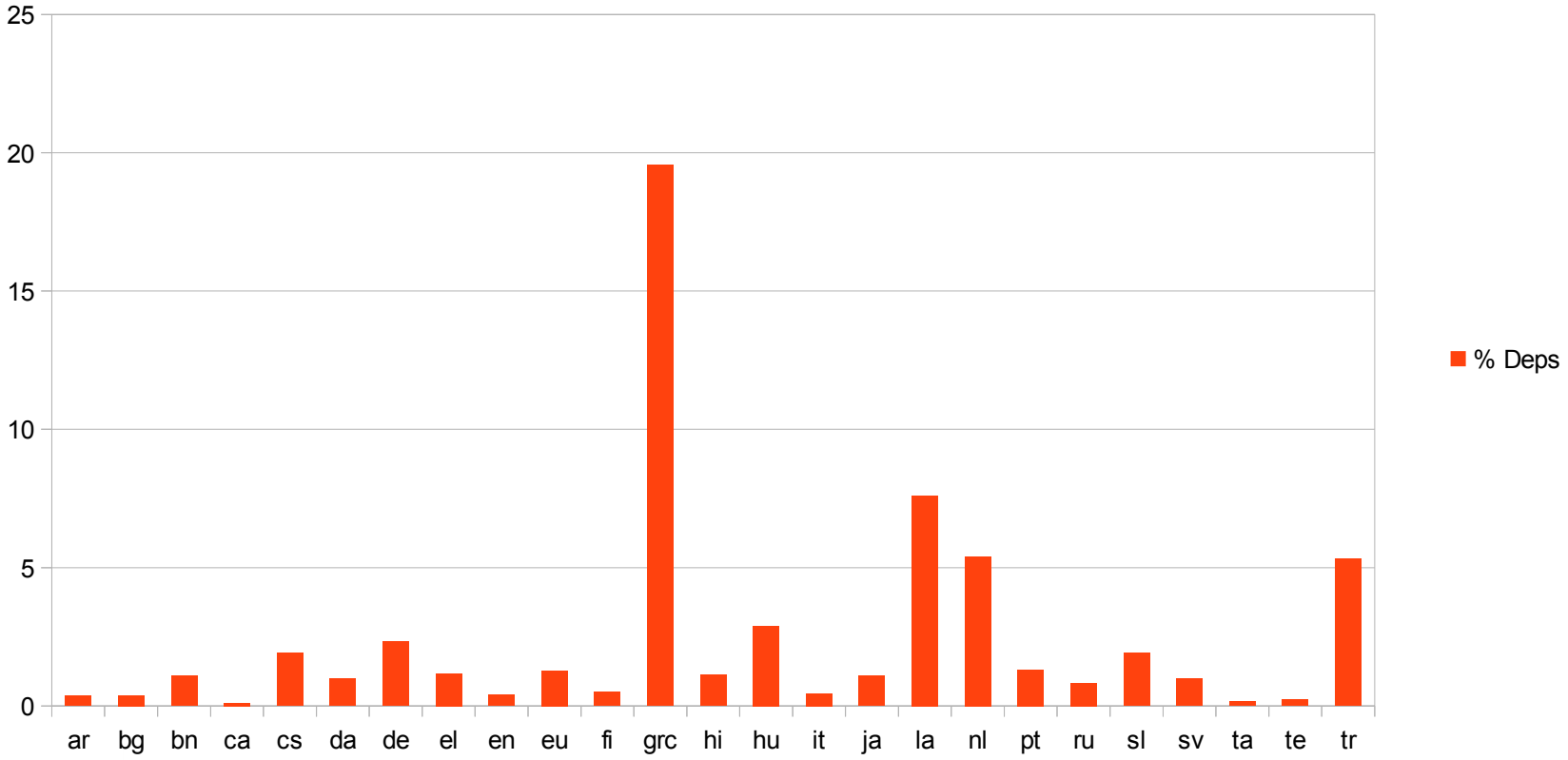
Data Size


1200000
1000000
800000
600000
400000
200000
0
ar bg bn ca cs da de el en es et eu fi grc hi hu it ja la nl pt ro ru sl sv ta te tr zh
Tokens

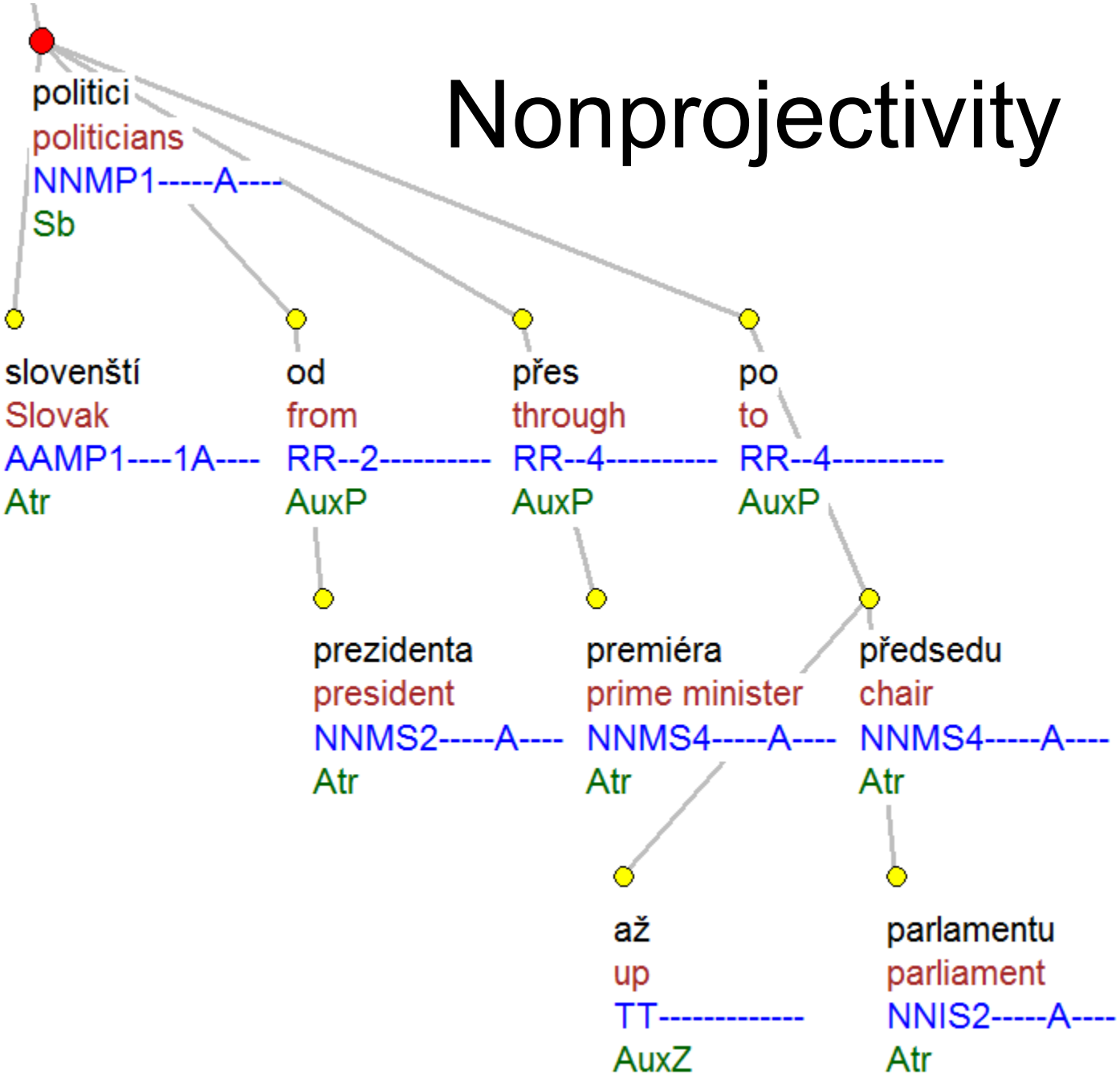
Sentence Length



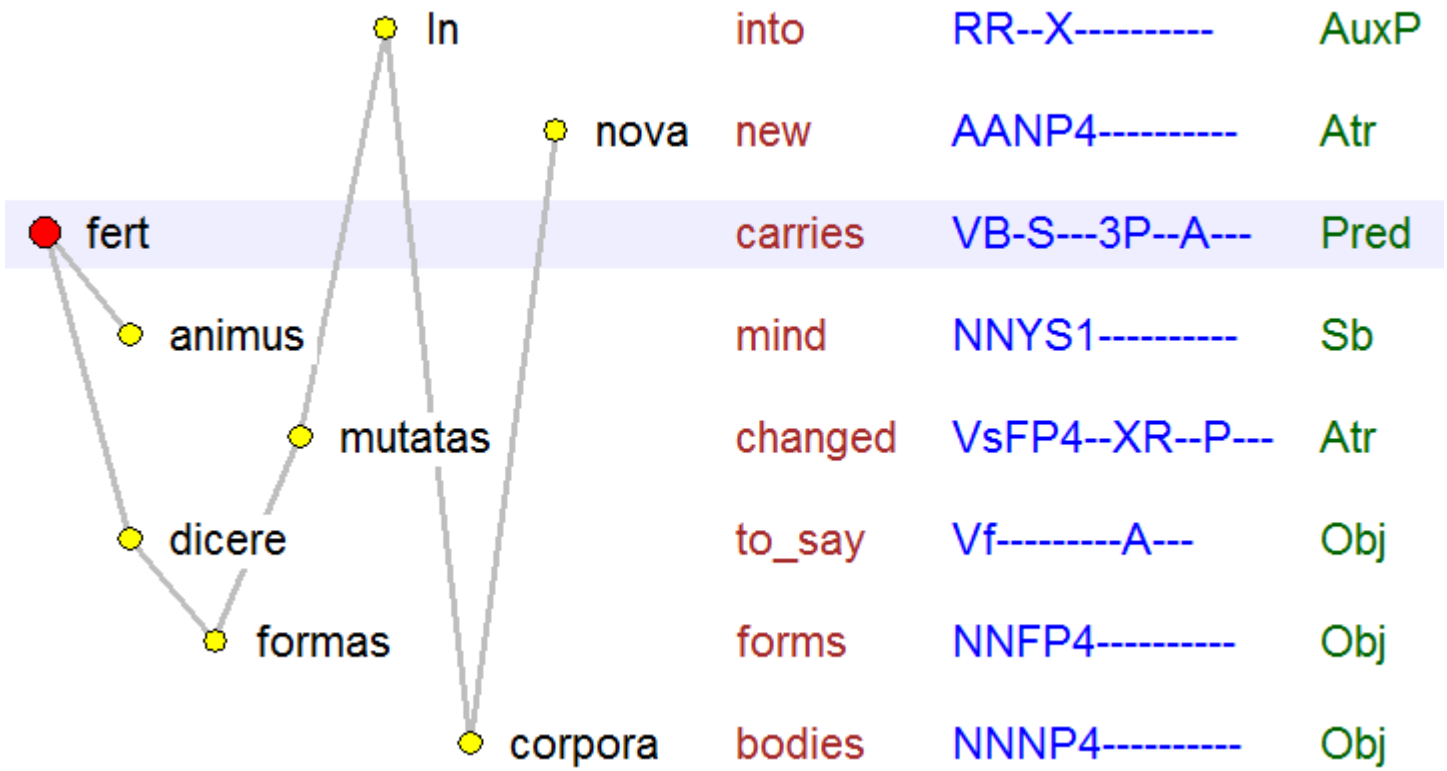
Nonprojective Dependencies



Nonprojectivity



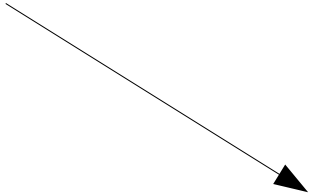
Nonprojective Latin



? the mind carries to say forms changed into new bodies ?

DZ Interset: Unify Morpho-Tags

- t-prppfa-



- Pos: verb
- Verbform: part
- Gender: fem
- Number: plu
- Case: acc
- Tense: past
- Voice: pass
- Aspect: perf

DZ Interset: Unify Morpho-Tags

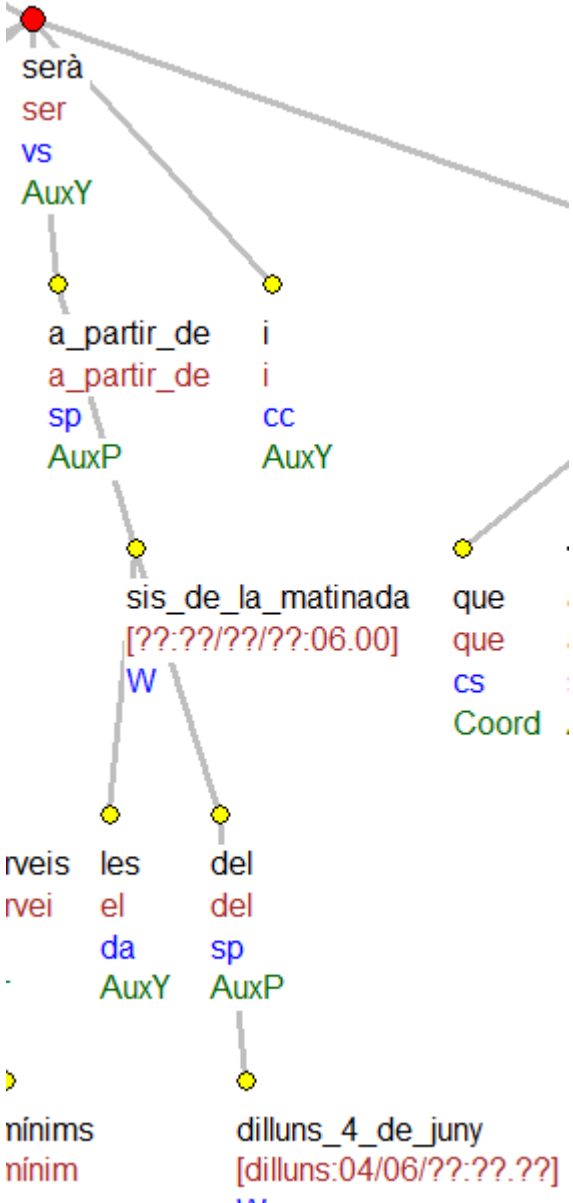
- la: t-prppfa-
- en: IN
- de: ADJA Pos|Dat|Sg|Neut
- ru: S ЕД СРЕД ВИН
- fi: ALL|SG|DV-JA|N
- pt: pron pron-indp <dem>|M|S
- ja: PSE

- **VsFP4--XR--P---**
- **RR--X----------**
- **AANS3----1-----**
- **NNNS4----------**
- **NNXSX----------**
- **PDYSX----------**
- **TT------------**

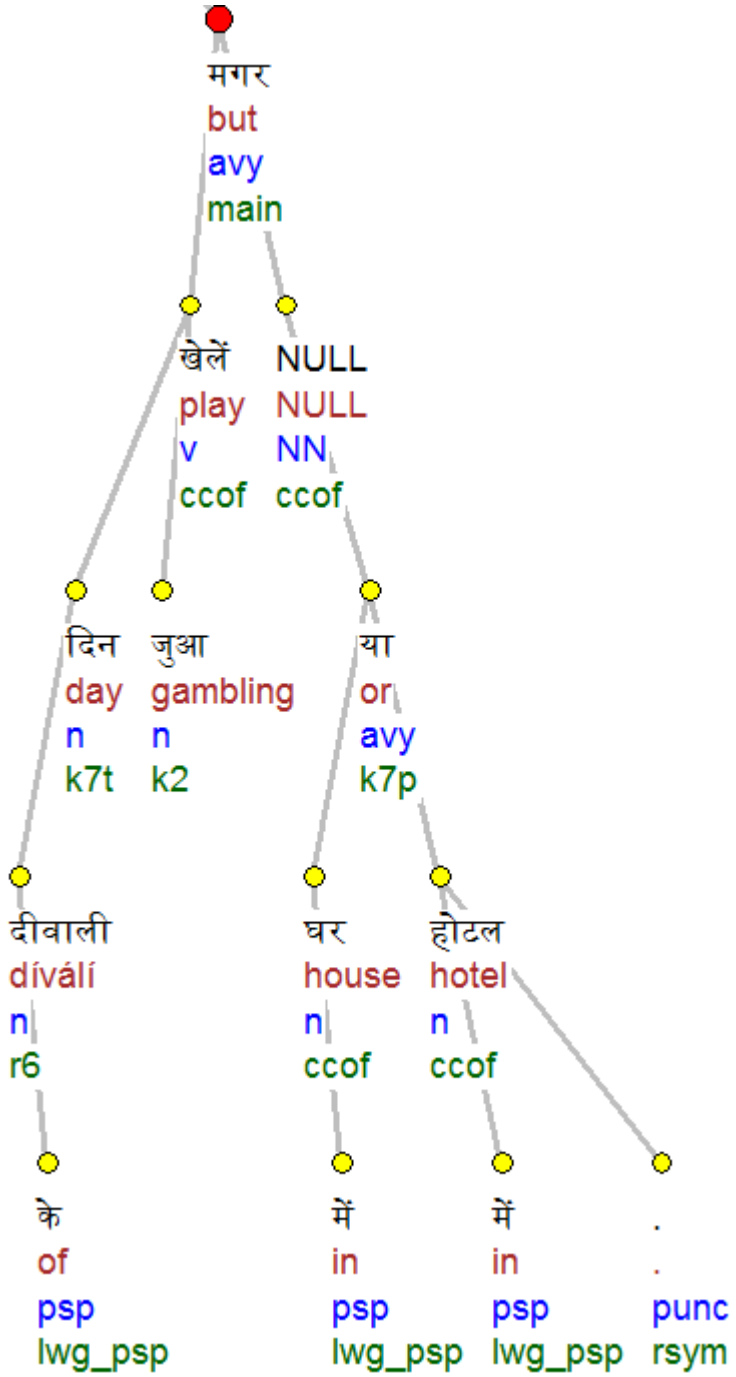
Original Morphology
Manual / Auto / Both

- Ancient Greek (grc)
- Arabic (ar)
- Basque (eu)
- Bengali (bn)
- Bulgarian (bg)
- Catalan (ca)
- Chinese (zh)
- Czech (cs)
- Danish (da)
- Dutch (nl)
- English (en)
- Estonian (et)
- Finnish (fi)
- German (de)
- Greek (el)
- Hebrew (he)
- Hindi (hi)
- Hungarian (hu)
- Icelandic (is)
- Italian (it)
- Japanese (ja)
- Latin (la)
- Portuguese (pt)
- Romanian (ro)
- Russian (ru)
- Slovene (sl)
- Spanish (es)
- Swedish (sv)
- Tamil (ta)
- Telugu (te)
- Turkish (tr)

We Don't Touch Tokenization



- Multi-word expressions = single tokens / nodes in some treebanks
- Separated elsewhere
- We don't normalize it



Tokenization

- NULL nodes

- दीवाली के दिन जुआ खेलें मगर NULL घर में या होटल में .
- dīvālī ke dina juā khelėṁ magara NULL ghara meṁ yā hoṭala meṁ .
- *On Diwali they gamble but* [they do so] *at home or hotel.*

Dependency Relation Labels

- DEPREL column in CoNLL data
- Afun ("analytical function") in Prague Treebanks
- Examples:
  - Sb, Pred, Obj, Adv, Atr, AuxP [cs]
  - nobj [da] … noun object (e.g. of preposition)
  - PG [de] … phrasal genitive (von-PP instead of gen)
  - AMS [de] … measure argument of adj *(zwei Jahre alt)*
  - k1 (karta / agent), k2 (karma / patient), k3 (karana / instrument), k4 (sampradaana / recipient) … [hi]

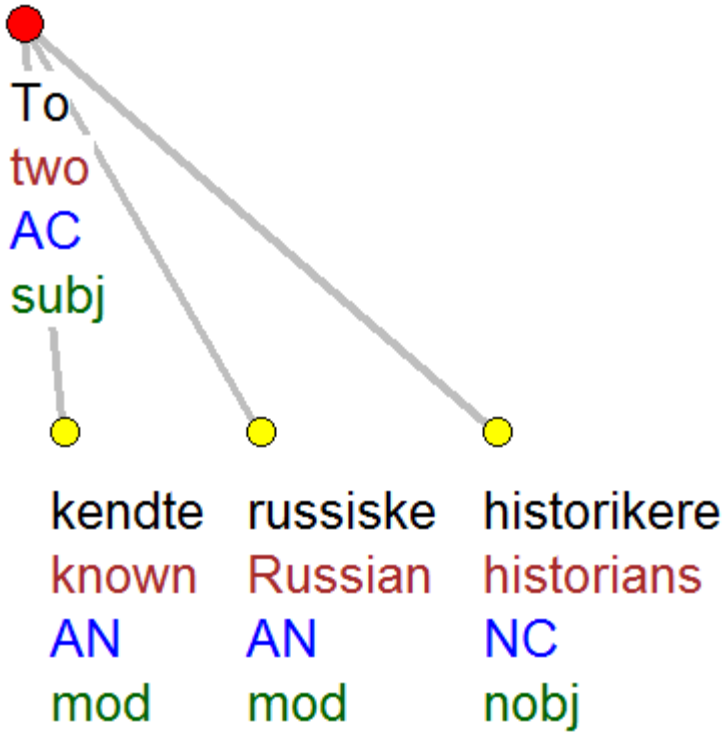
Structural Variations

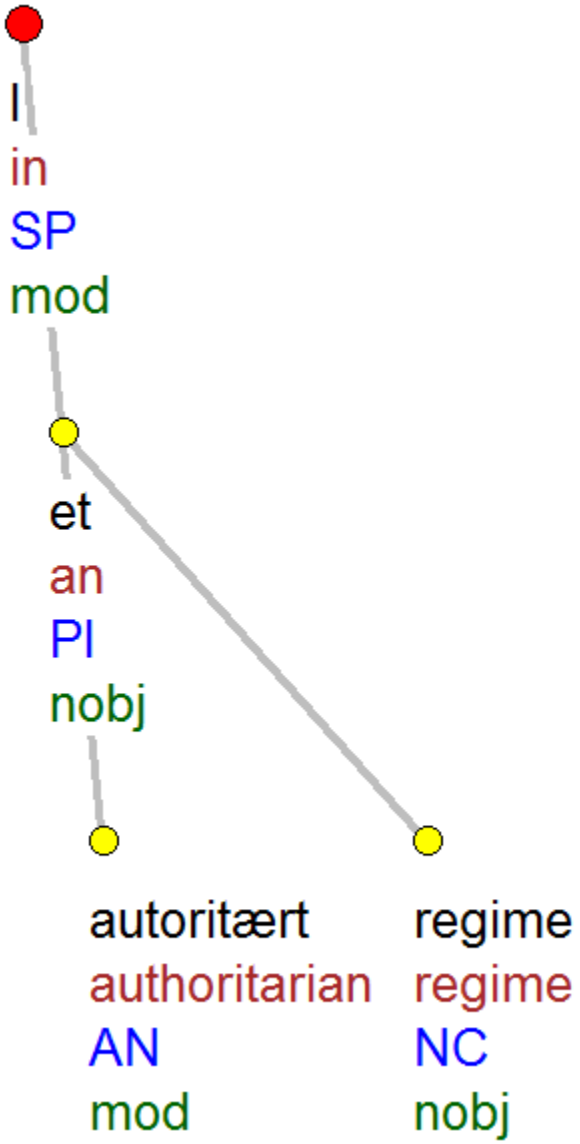
- DDT: exotic animal

- Prepositions and/or postpositions
- Subordinated clauses
- Verb groups
- Punctuation
- Apposition
- Coordination

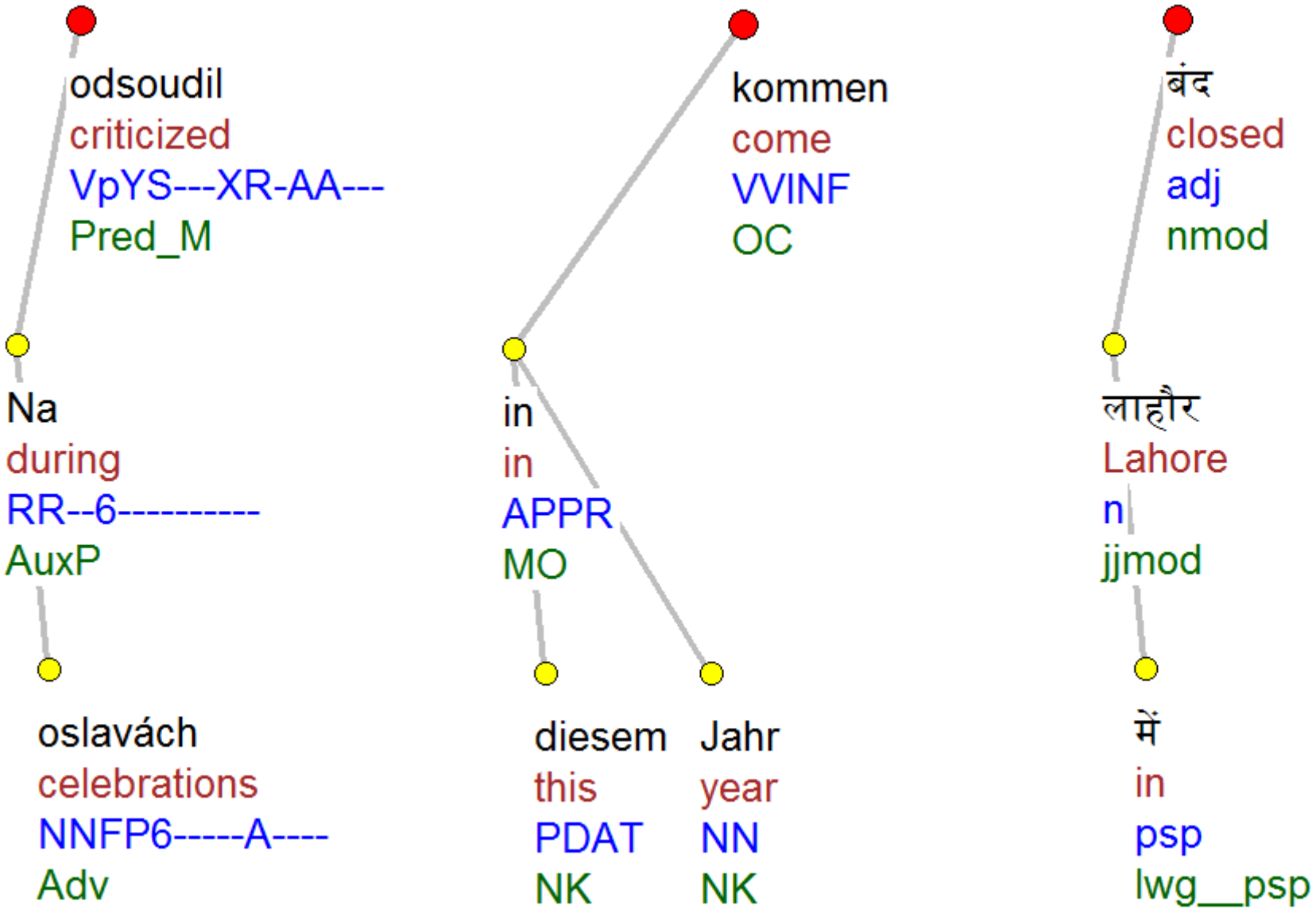
We try to automatically identify these constructions and restructure them as in PDT.

Danish Dependency Treebank

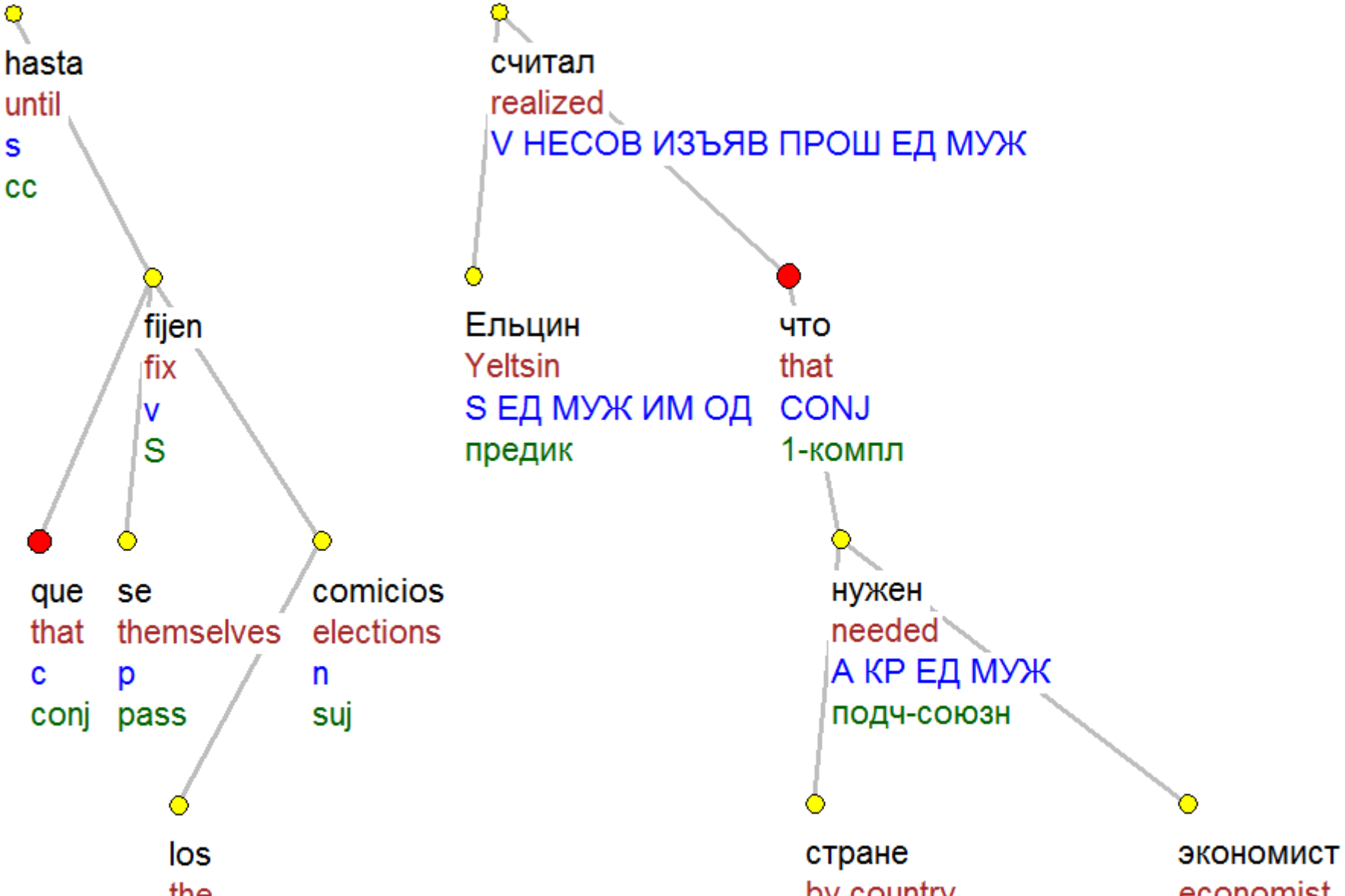




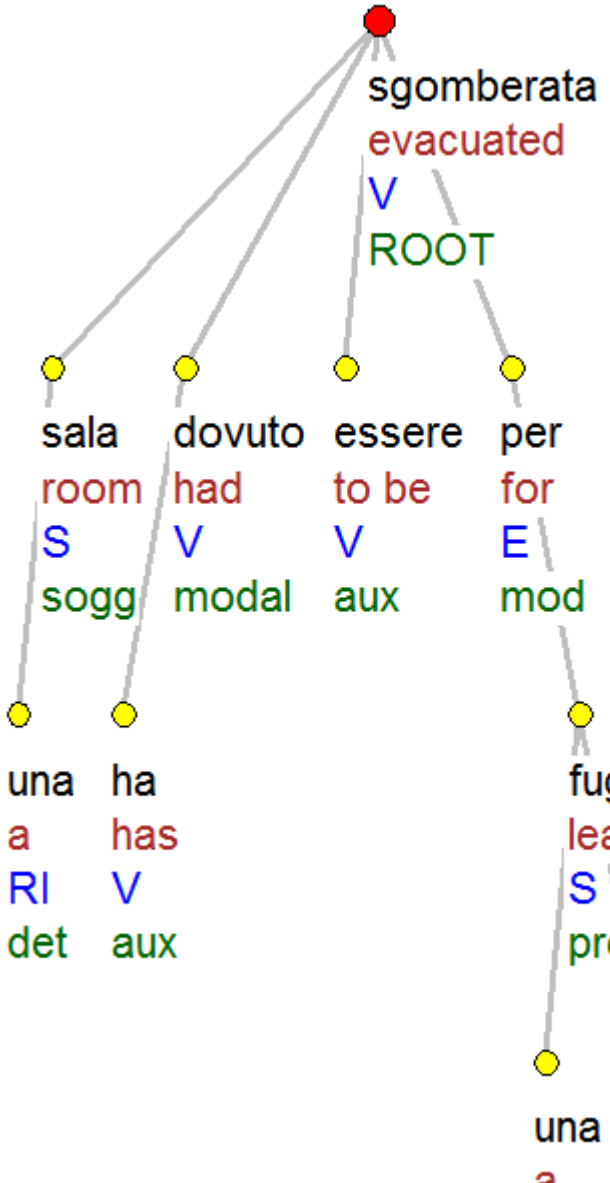
Prepositions

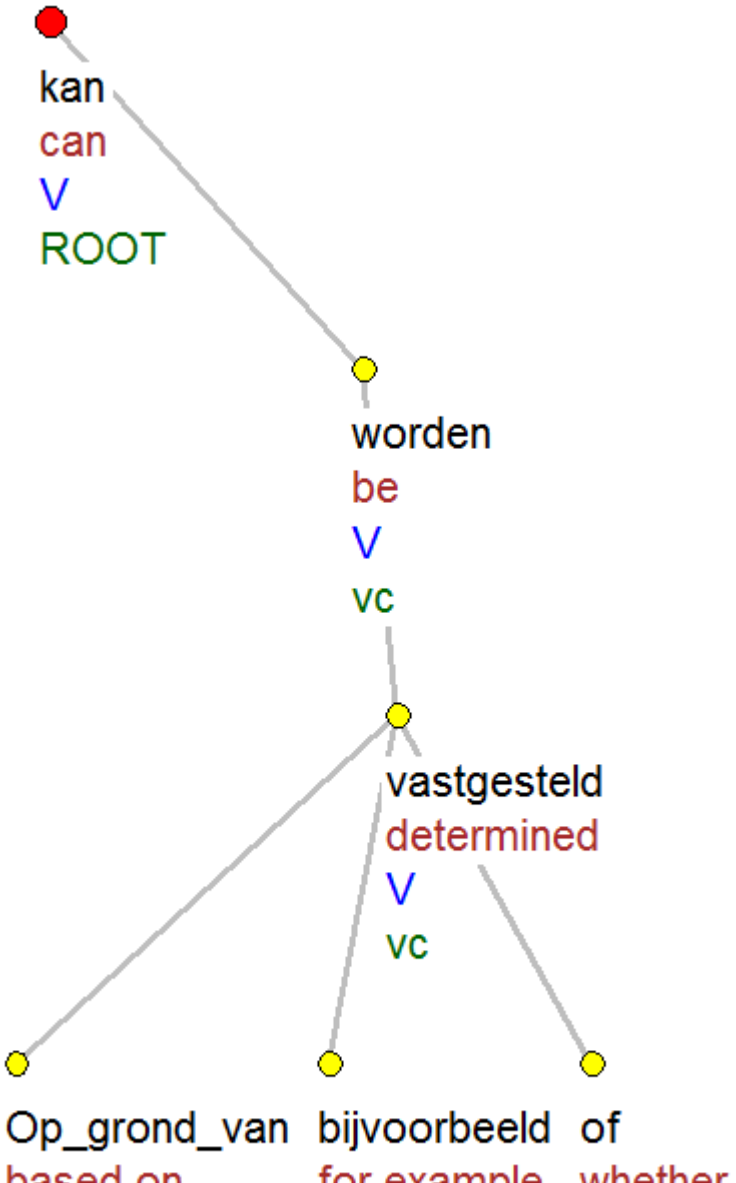


Subordinated Clauses



Verb Groups

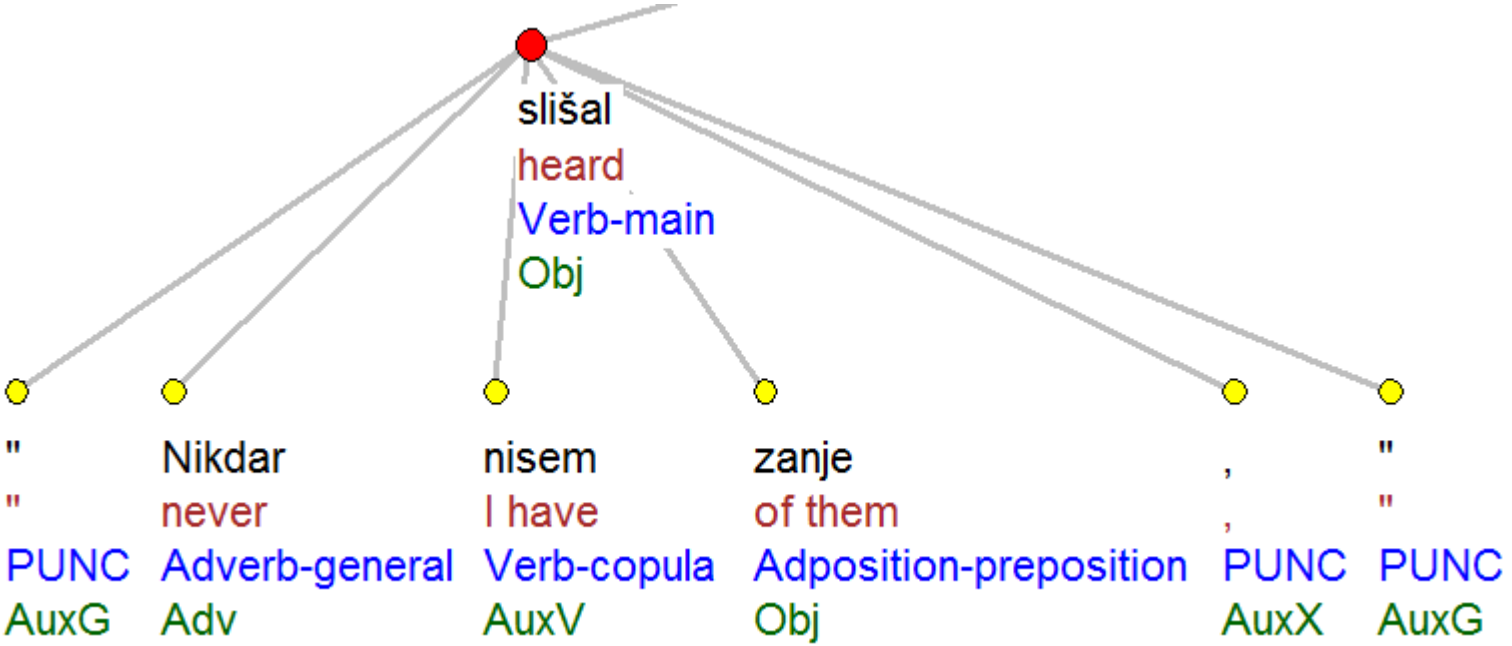

sgomberata
evacuated
V
ROOT
sala
room
S
sogg
dovuto
had
V
modal
essere
to be
V
aux
per
for
E
mod
una
a
RI
det
ha
has
V
aux
fuga
leak
S
prep
una
di


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

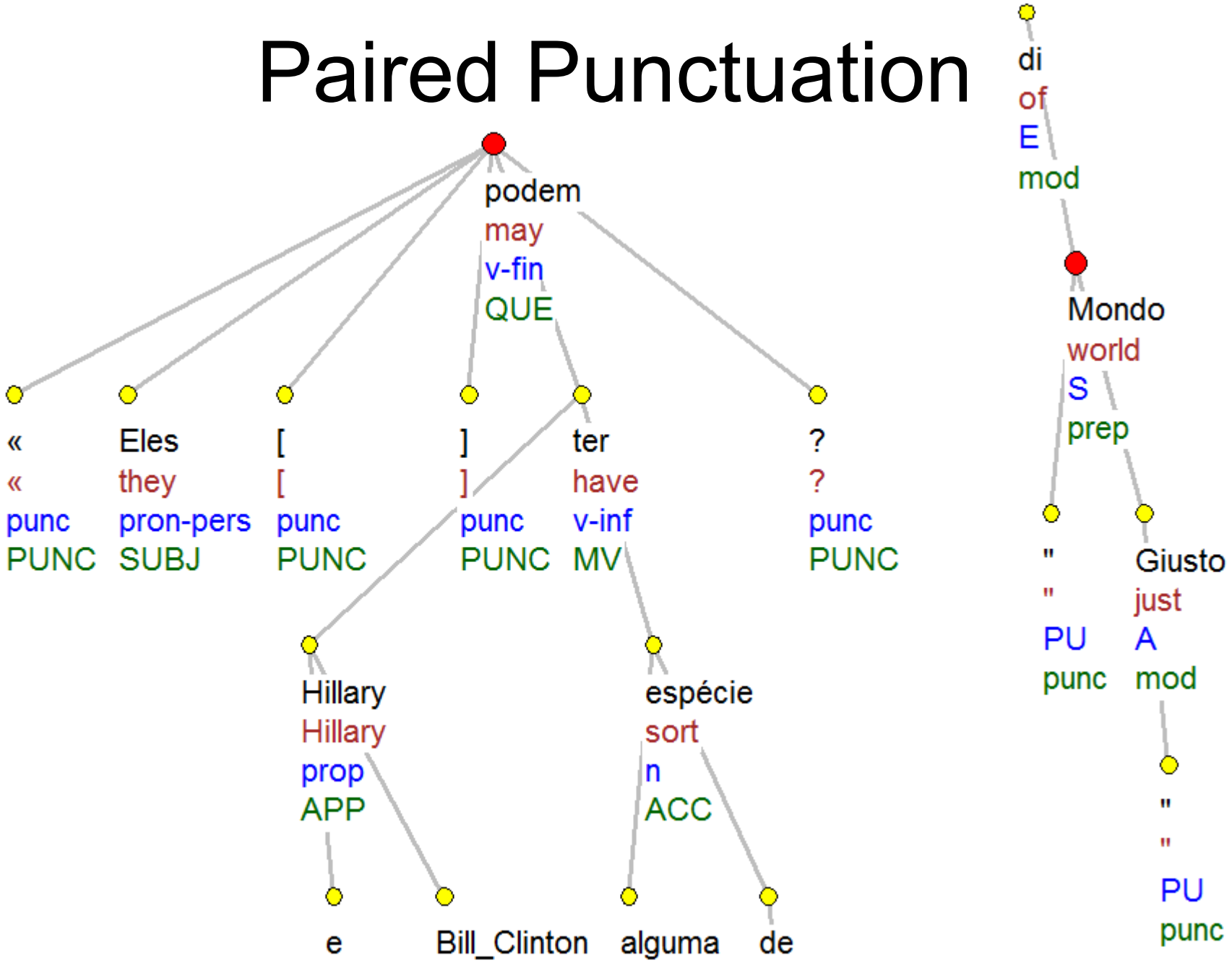
Final Punctuation

- On artificial root [cs, ar, sl, grc, ta]
- Between artificial root and main predicate [tr]
- On main predicate [bg, ca, da, de, en, es, et, fi, hu, …]
- On the predicate of the last clause [hi]
- On previous token [eu, it, ja, nl]
- No punctuation [ru, ro]

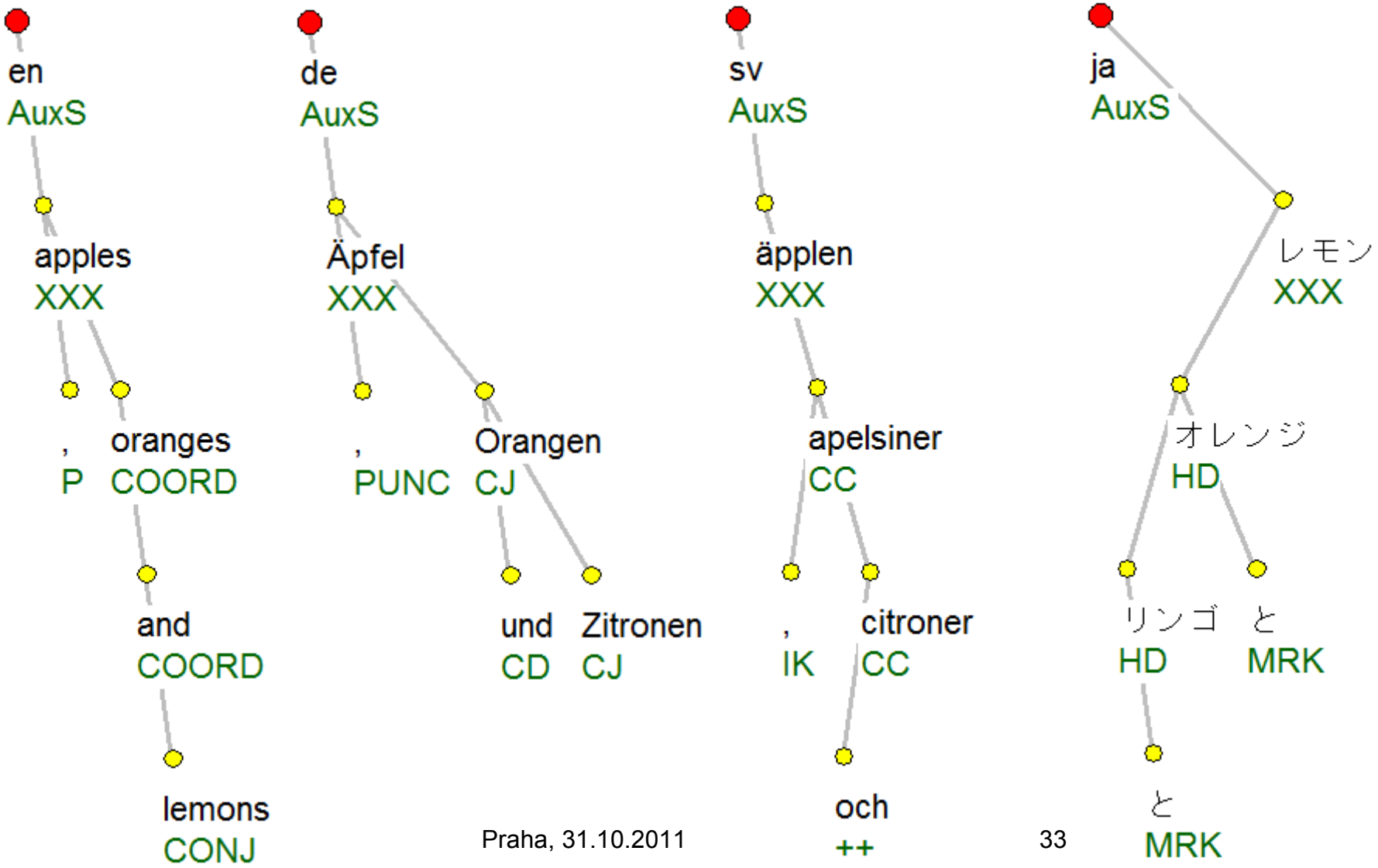
Paired Punctuation



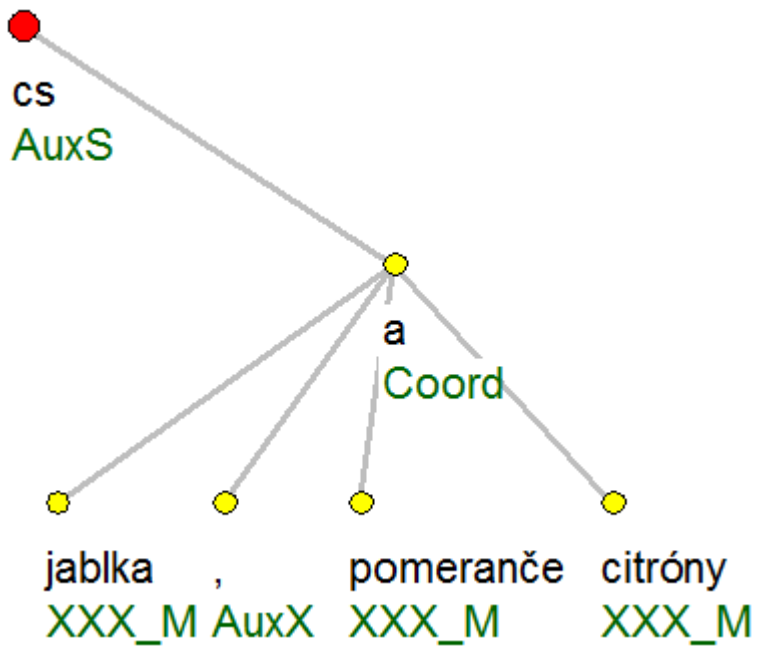
Paired Punctuation

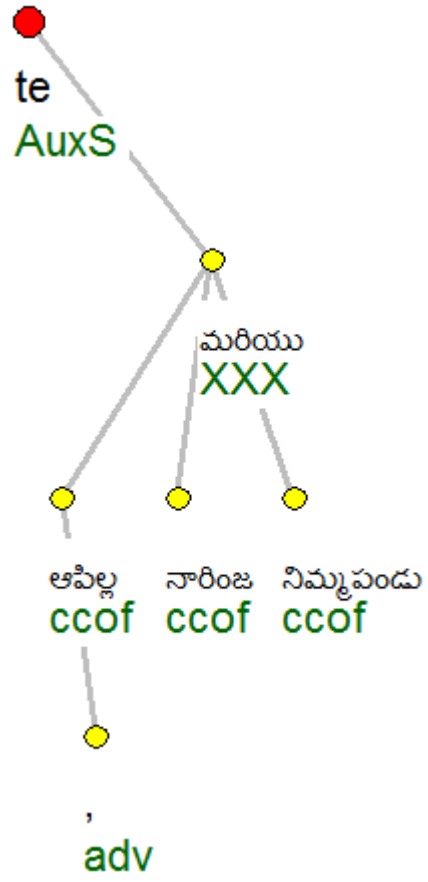


Coordination: Mel'čuk

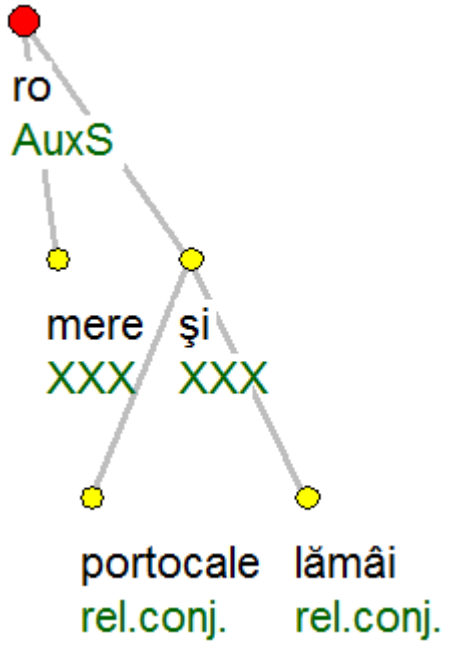


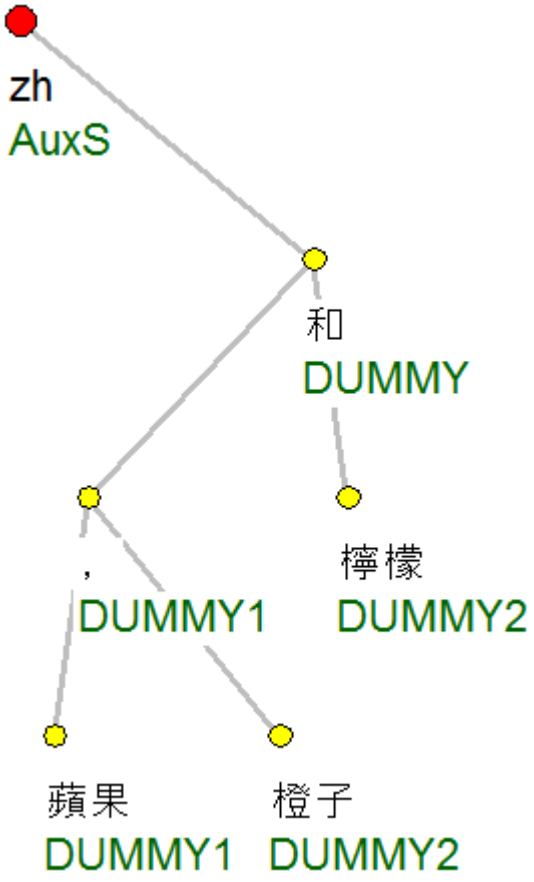
Coordination: Prague



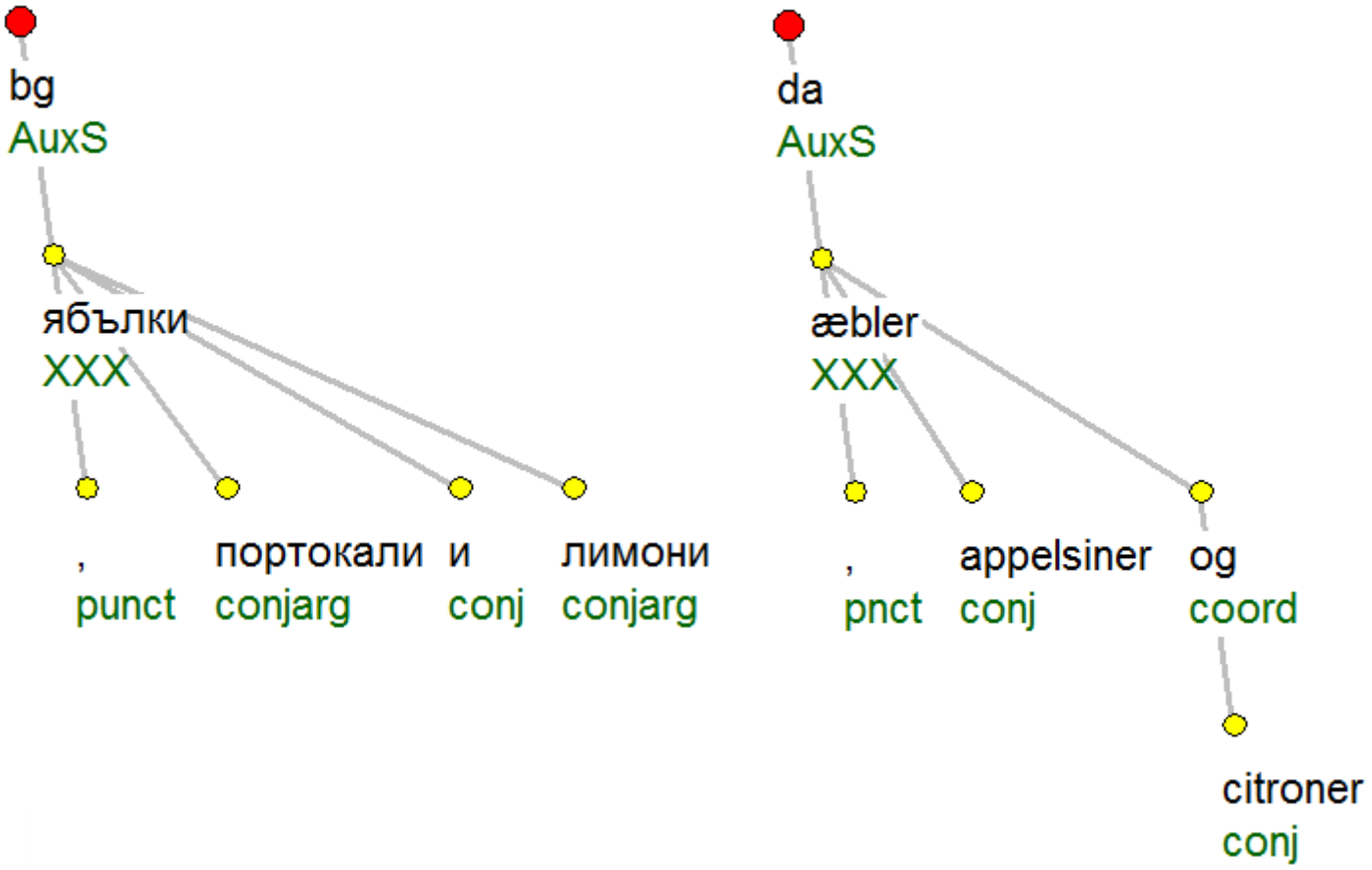


Coordination: [ro, zh]

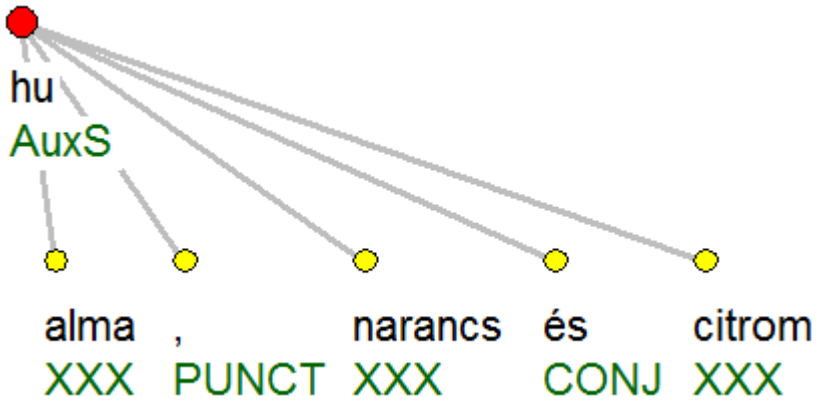




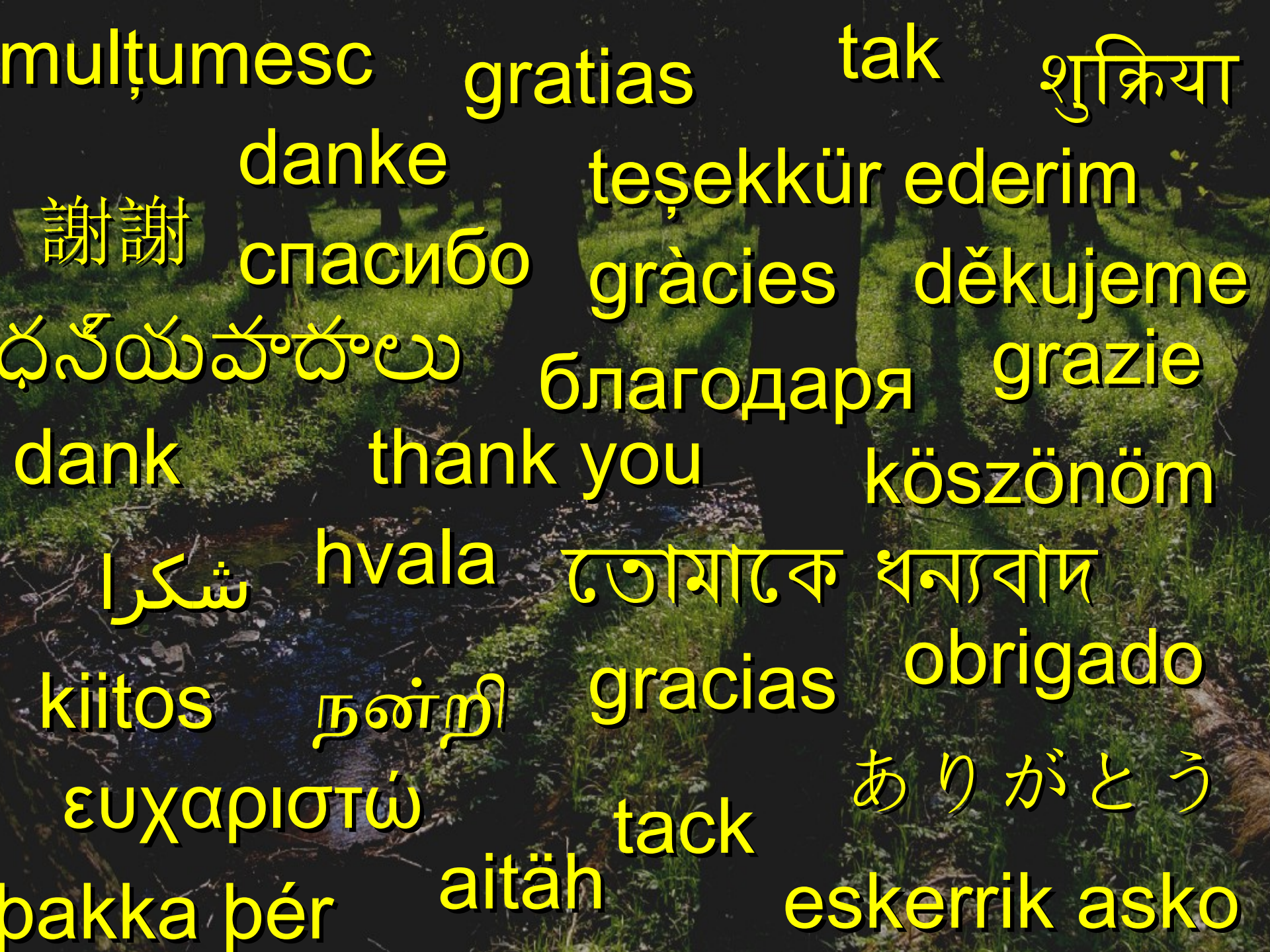
Coordination: Stanford



Coordination: Tesnière




END OF PART ONE


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