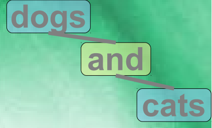


ACL 2013 paper

Coordination Structures in Dependency Treebanks

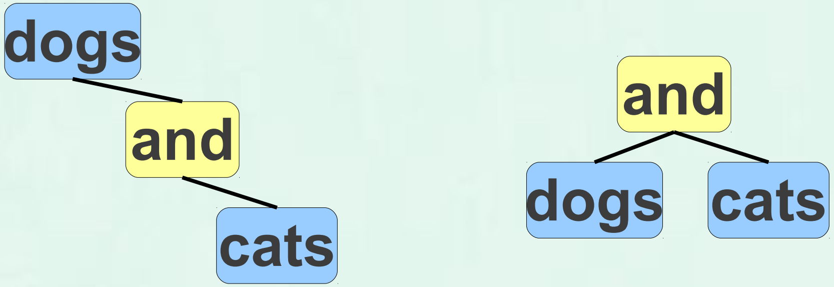
Martin Popel, David Mareček, Jan Štěpánek,
Daniel Zeman, Zdeněk Žabokrtský

Charles University in Prague,
Faculty of Mathematics and Physics,
ÚFAL (Institute of Formal and Applied Linguistics)

September 19th 2013, Příchovice

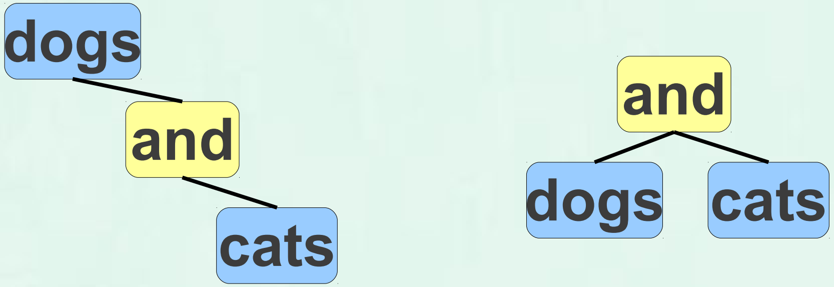
Motivation

- Coordination and Dependency are fundamentally different relations
- Coordinations are difficult to represent in dependency treebanks
- Large inter-treebank differences

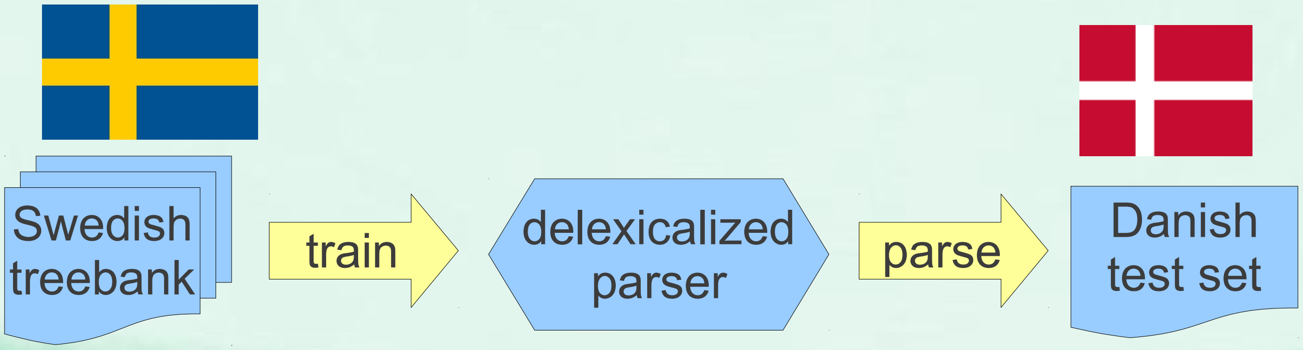


Motivation

- Coordination and Dependency are fundamentally different relations
- Coordinations are difficult to represent in dependency treebanks
- Large inter-treebank differences



- Obstacle for cross-lingual parsing (evaluation)



Outline

- Styles of annotating coordinations
    - Topological styles
    - Labeling styles
- Transformation of styles
- Data: HamleDT (26 languages)

Participants of coordination

- conjunct
- delimiter (separates two conjuncts)
  - Coordinating conjunction
  - Comma or other punctuation (semicolon)
- shared modifier (modifies two or more conjuncts)

Examples:

- lazy dogs , cats and rats — more than two conjuncts ("multi-conjunct c.")
- Mary came home and cried — *home* is a "private modifier"
- John and Mary or Peter — nested (embedded) coordinations
- big and cheap apples and oranges — coordinated shared modifier

Special cases

- Asyndetic coordination = no conjunction

  Don't worry , be happy , keep smiling

Special cases

- Asyndetic coordination = no conjunction

  Don't worry , be happy , keep smiling

- Multi-word conjunction **as well as**

Special cases

- Asyndetic coordination = no conjunction
  Don't worry **,** **be happy** **,** **keep smiling**
- Multi-word conjunction **as well as**
- Single-conjunct coordination **And** **I love her**

Special cases

- Asyndetic coordination = no conjunction

  Don't worry **,** **be happy** **,** **keep smiling**
- Multi-word conjunction **as well as**
- Single-conjunct coordination **And** **I love her**
- One token with more roles **etc.**

  Senatus **Populusque** **Romanus** *que* = coord. enclitic

  (The Senate and the People of Rome)

Special cases

- Asyndetic coordination = no conjunction

  Don't worry , be happy , keep smiling

- Multi-word conjunction **as well as**
- Single-conjunct coordination **And I love her**
- One token with more roles **etc.**

  Senatus Populusque Romanus *que* = coord. enclitic

  (The Senate and the People of Rome)

- Paratactic vs. hypotactic means *(John with Mary)*

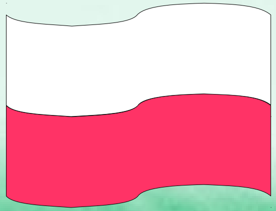
Special cases

- Asyndetic coordination = no conjunction

  Don't worry , be happy , keep smiling
- Multi-word conjunction **as well as**
- Single-conjunct coordination **And I love her**
- One token with more roles **etc.**

  Senatus Populusque Romanus *que* = coord. enclitic

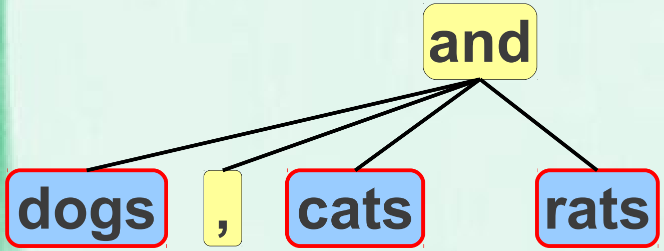
  (The Senate and the People of Rome)
- Paratactic vs. hypotactic means *(John with Mary)*
- *red and white wine = red wine and white wine*
  red and white flag of Poland



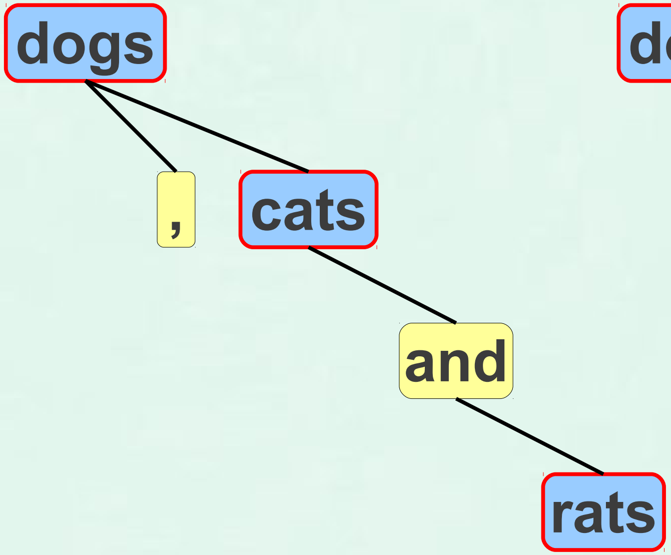
Topological styles (family)

Main “family” – configuration of conjuncts

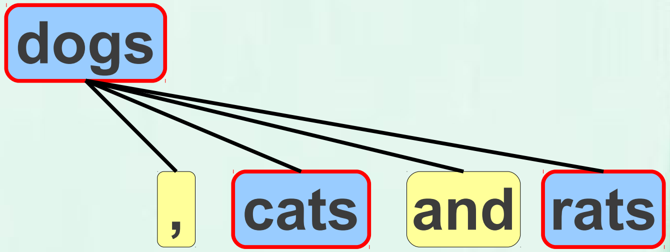
Prague



Moscow



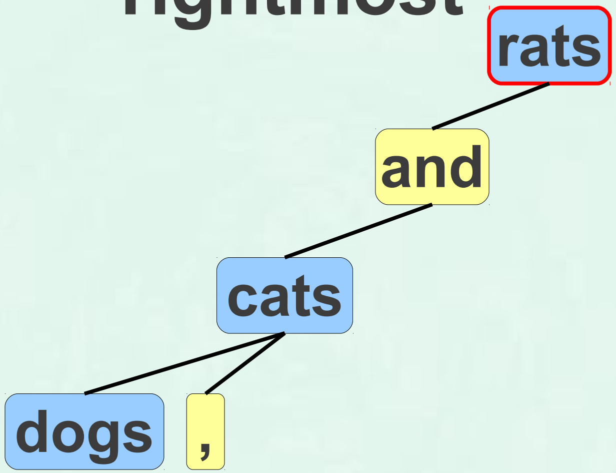
Stanford



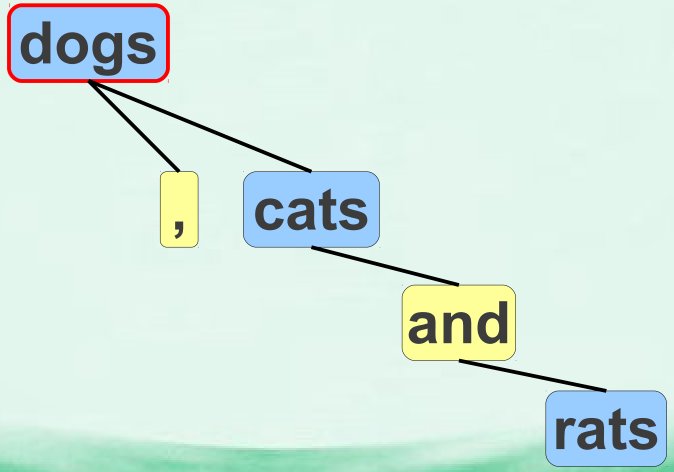
Topological styles (head)

Choice of head (which delimiter/conjunct to choose):

rightmost

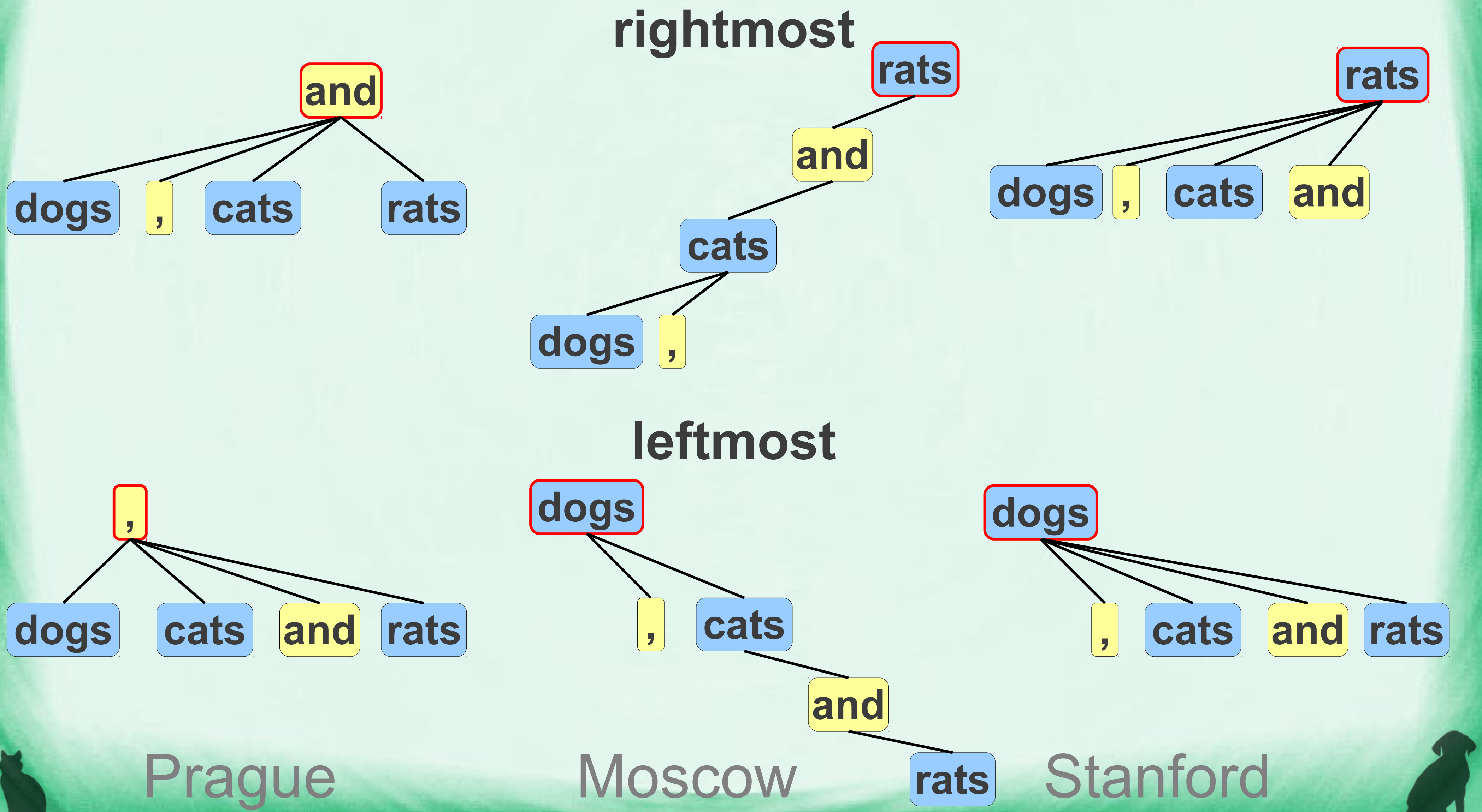


leftmost



Topological styles (head)

Choice of head (which delimiter/conjunct to choose):



rightmost

and
dogs , cats rats

rats
and
cats
dogs ,

rats
dogs , cats and

leftmost

,
dogs cats and rats

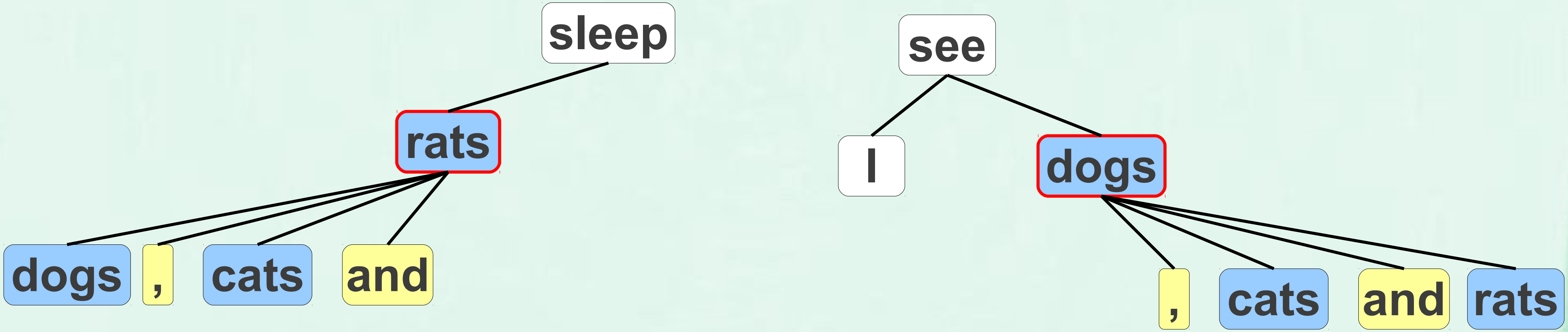
dogs
, cats
and
rats

dogs
, cats and rats

Prague Moscow Stanford

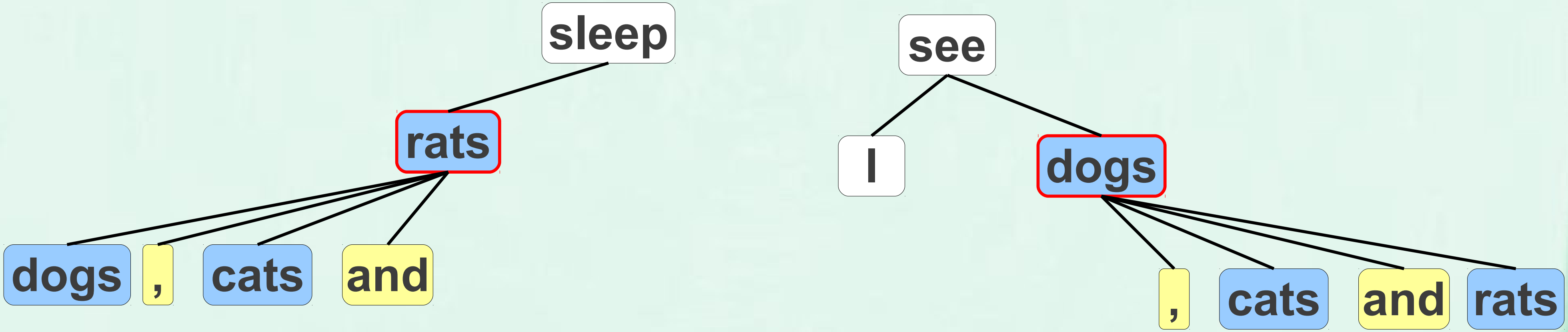
Topological styles (head)

Choice of head: leftmost, rightmost or **mixed**



Topological styles (head)

Choice of head: leftmost, rightmost or **mixed**

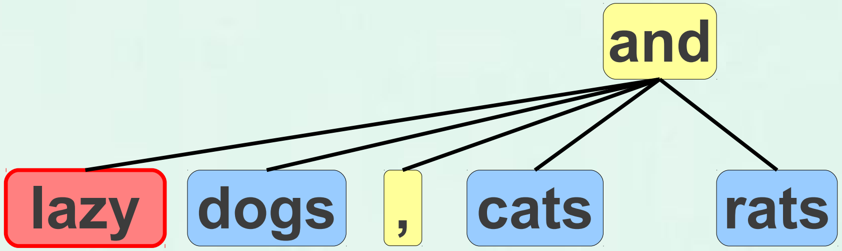


Persian treebank: rightmost for coordination of verbs leftmost otherwise

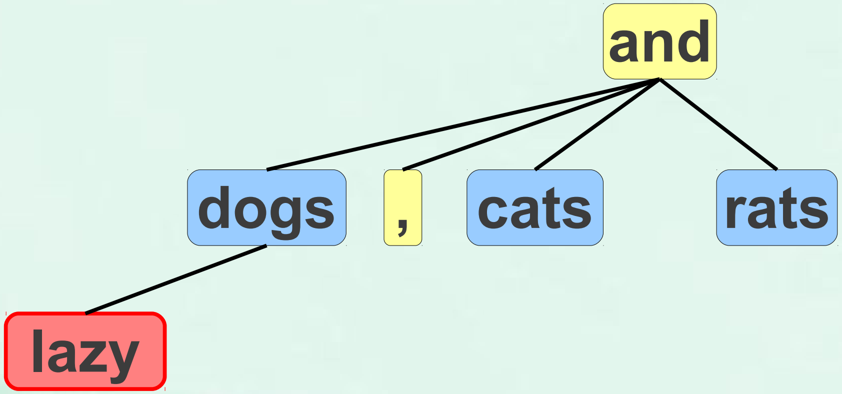
Topological styles (shared modifiers)

Attachment of shared modifiers:

below **the head**



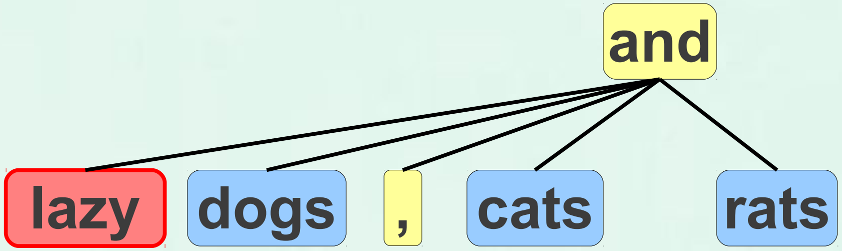
below **the nearest conjunct**

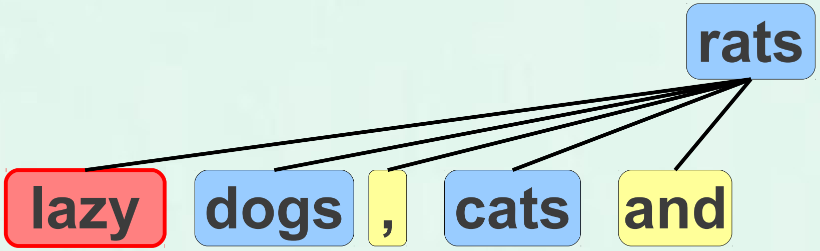


Topological styles (shared modifiers)

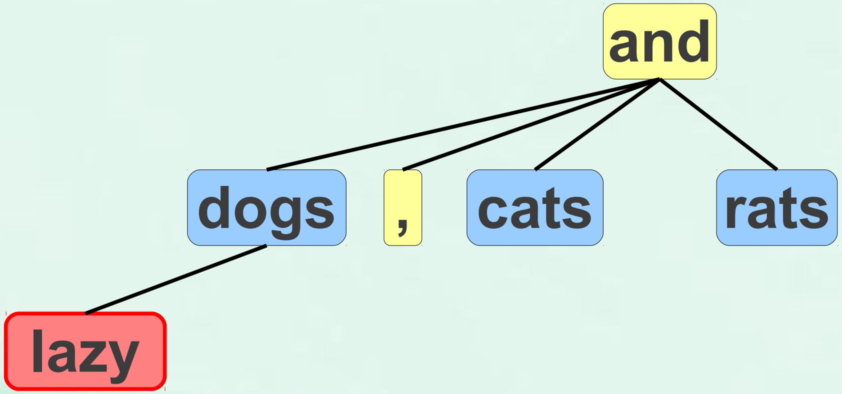
Attachment of shared modifiers:

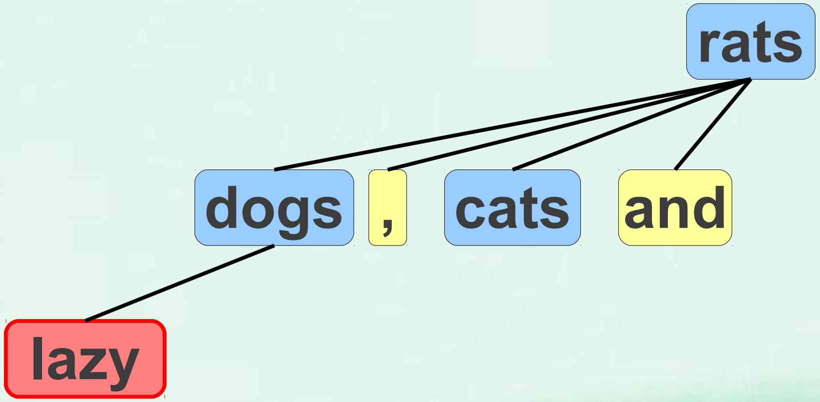
below **the head**





below **the nearest conjunct**





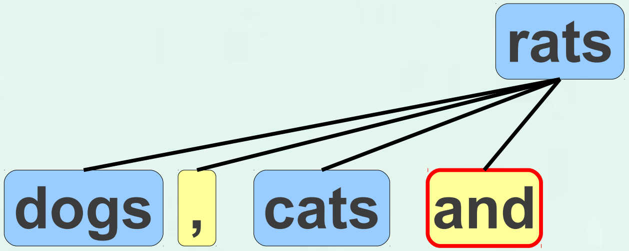
Prague

Stanford

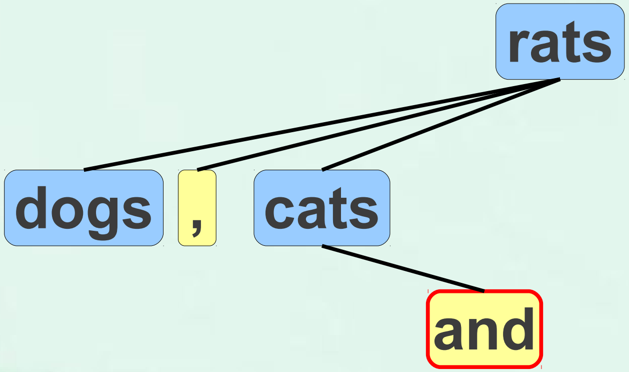
Topological styles (conjunction)

Attachment of coordinating conjunctions:

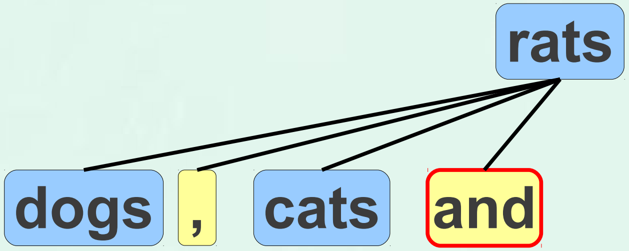
“between” conjuncts



below the **previous conjunct**



following conjunct

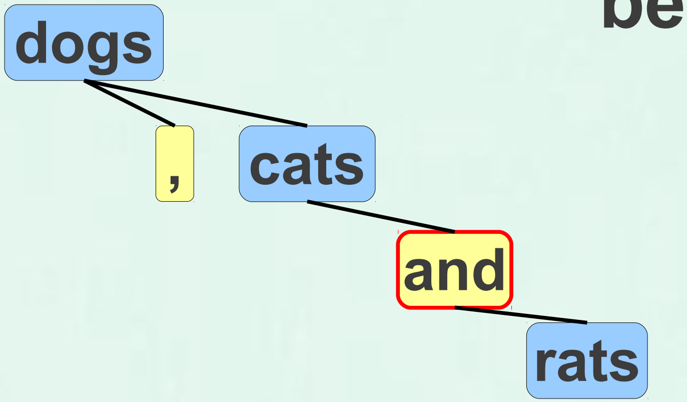


Stanford, head=rightmost

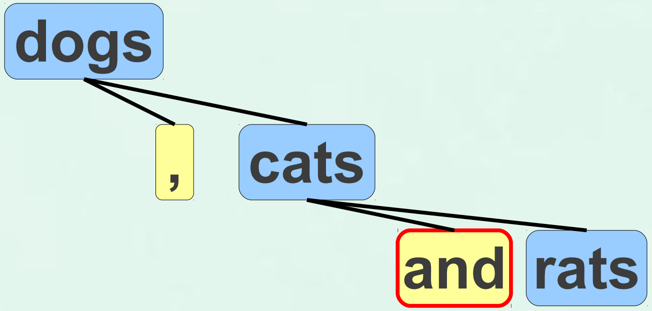
Topological styles (conjunction)

Attachment of coordinating conjunctions:

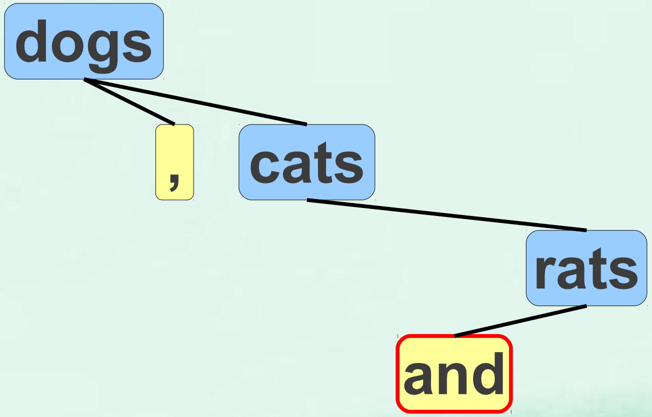
“between” conjuncts



below the **previous conjunct**



following conjunct

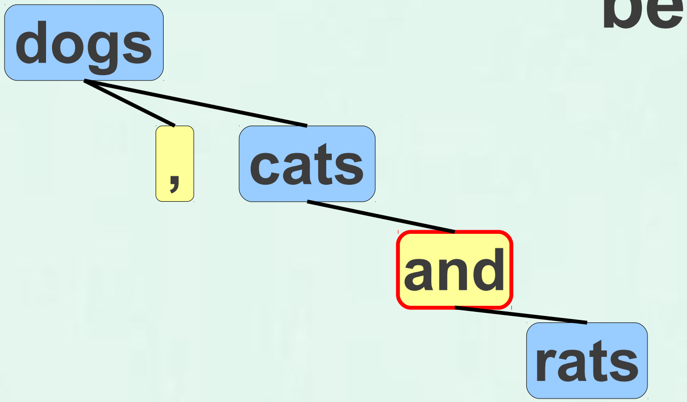


Moscow, head=leftmost

Topological styles (conjunction)

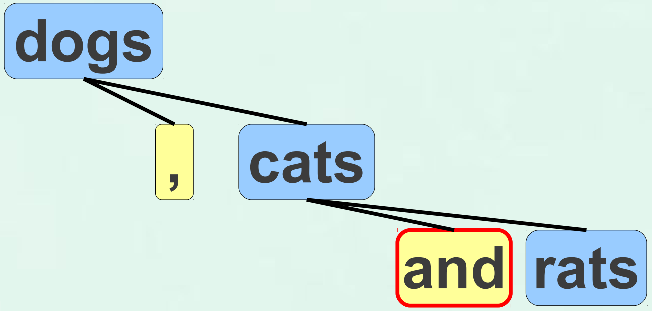
Attachment of coordinating conjunctions:

“between” conjuncts

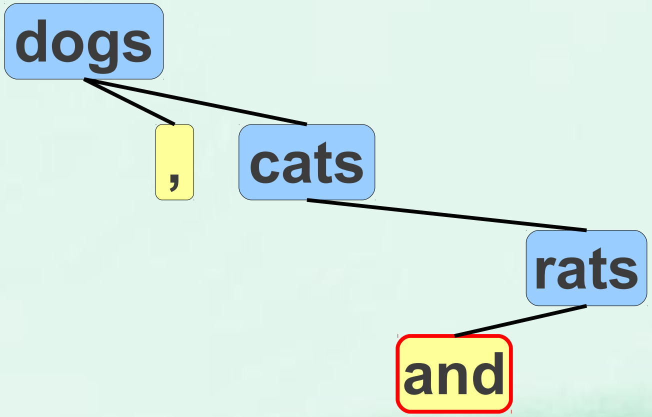


“as the head”
for Prague (the only applicable)

below the **previous conjunct**



following conjunct

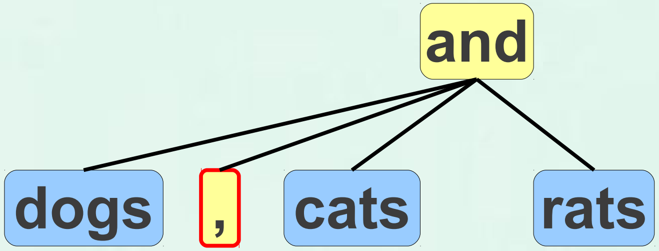


Moscow, head=leftmost

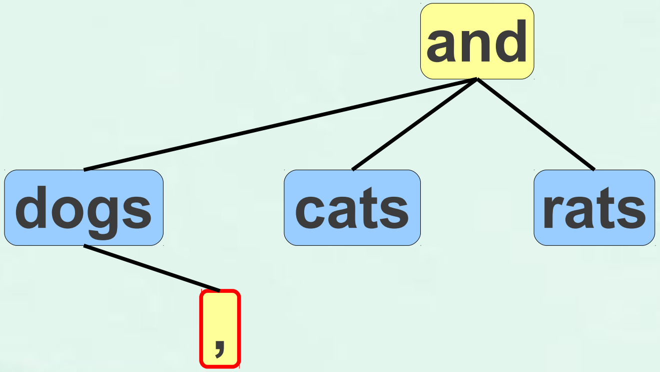
Topological styles (punctuation)

Attachment of punctuation delimiters:

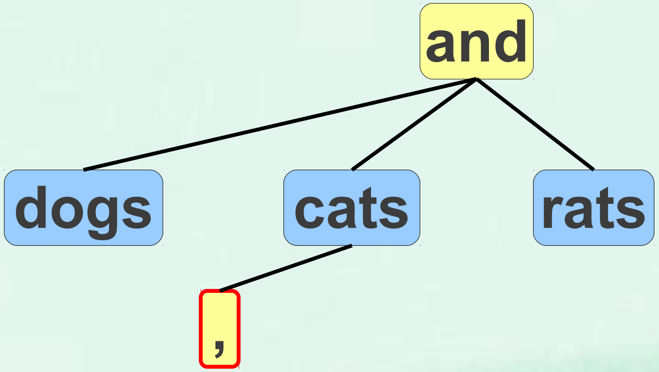
“between” conjuncts



below the **previous conjunct**



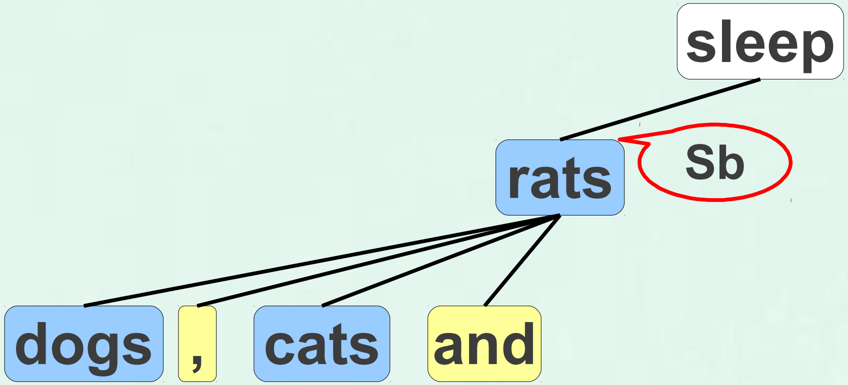
following conjunct

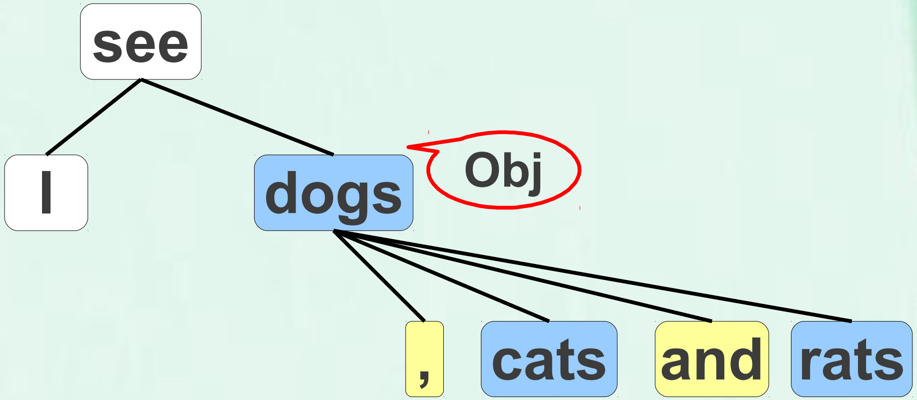


Prague

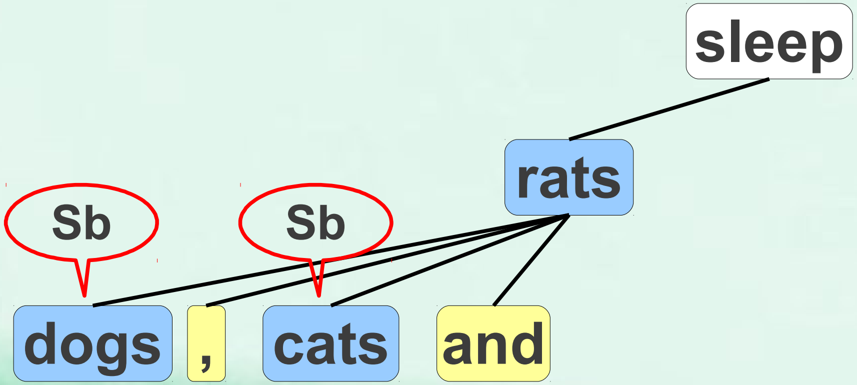
Labeling styles (dependency rel.)

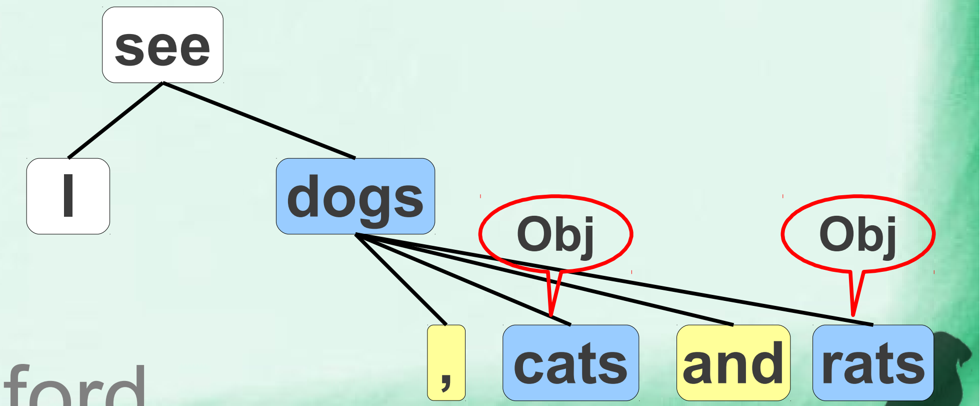
Dependency relation at “**upper level**” = with the head node





Dependency relation at “**lower level**” = with the conjuncts

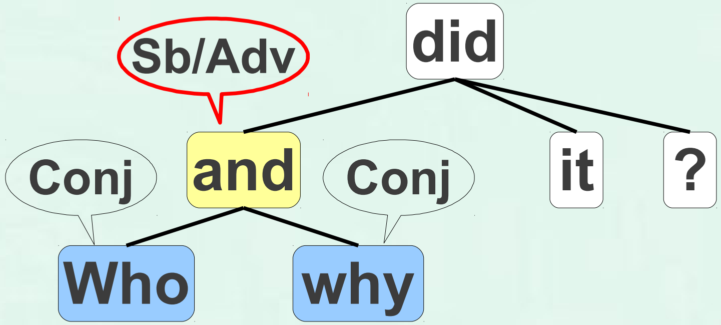




Stanford

Labeling styles (dependency rel.)

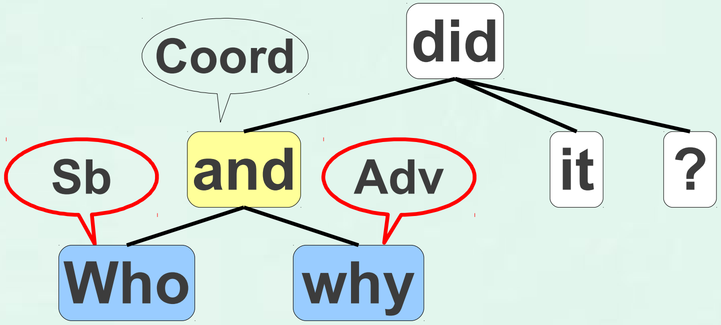
Dependency relation at “**upper level**” = with the head node



Dependency relation at “**lower level**” = with the conjuncts

Allows different labels

of conjuncts.



Prague

Labeling styles (other)

- Are **conjuncts** annotated?
  - additional attribute (`is_member`) or
  - encoded into the dependency label: Sb_M, Obj_M, Atr_M,...
- Are **shared modifiers** annotated?
  - In PDT not explicitly, but it can be deduced.
- Proposed, but unseen in treebanks:
  co-indexation attributes or bubbles
  for nested coordinations and shared modifiers

Annotation styles – overview

How many treebanks (out of 26 in HamleDT 1.0) use a given style?

- **Family** (Prague=14, Moscow=5, Stanford=6)
- **Head** (Leftmost=10, Rightmost=14, Mixed=1)
- **Shared modifiers** (below Head=11, Nearest conjunct=15)
- **Conjunctions** (Previous=2, Following=1, Between=8, as Head=14)
- **Punctuation** (Previous=7, Following=1, Between=15, Missing=2)

- **Dependency relation** (Upper=17, Lower=9)
- **Annotated conjuncts** (yes=21, no=5)
- **Annotated shared modifiers** (yes=8, no=18)

Annotation styles – overview

How many possible styles?

2*3*2*3*3+1*3*2*1*3 = 126 topological

* 8 labeling variants = 1008

How many styles really found?

16 (in 26 treebanks)

Transformations of styles

Subtasks

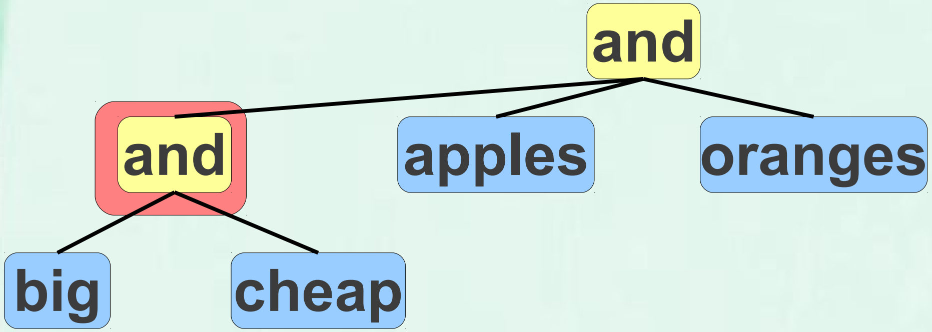
1. Detect coordinations in a sentence
   (esp. boundaries of nested coordinations)

2. Classify participants of coordinations
   (conjunct, commas, conjunctions, shared m.)

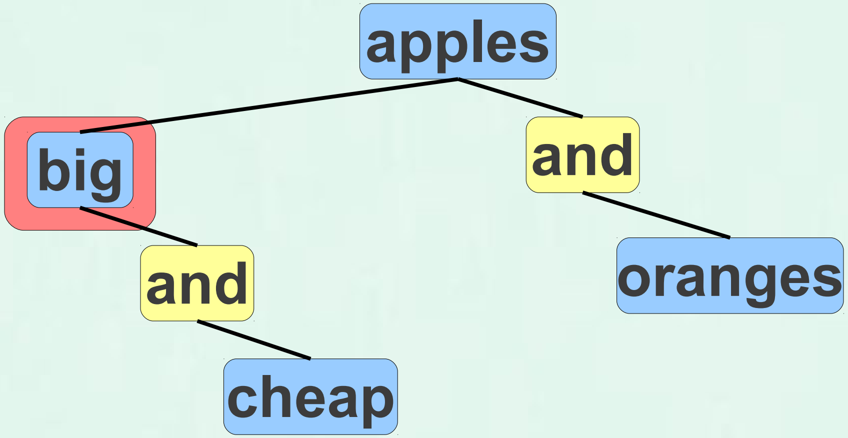
3. Transform each coordination to the target style
   (depth-first recursion, start with inner coord.)

Problematic cases

big and cheap apples and oranges

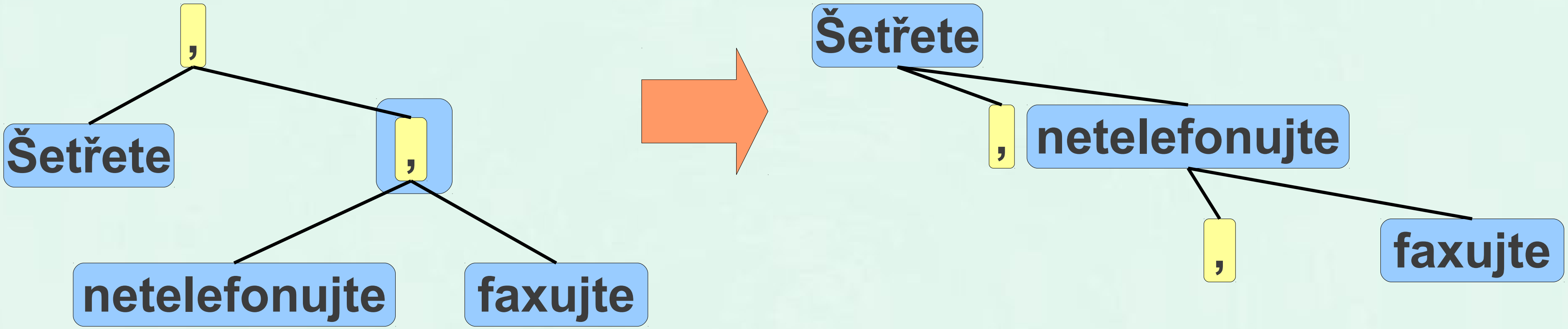


Prague



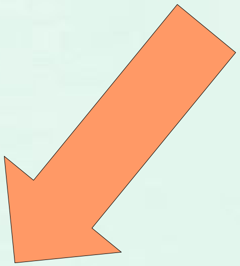
Moscow

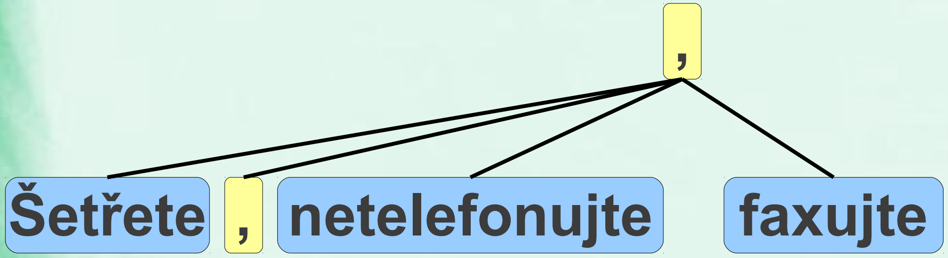
Problematic cases



Prague

Moscow

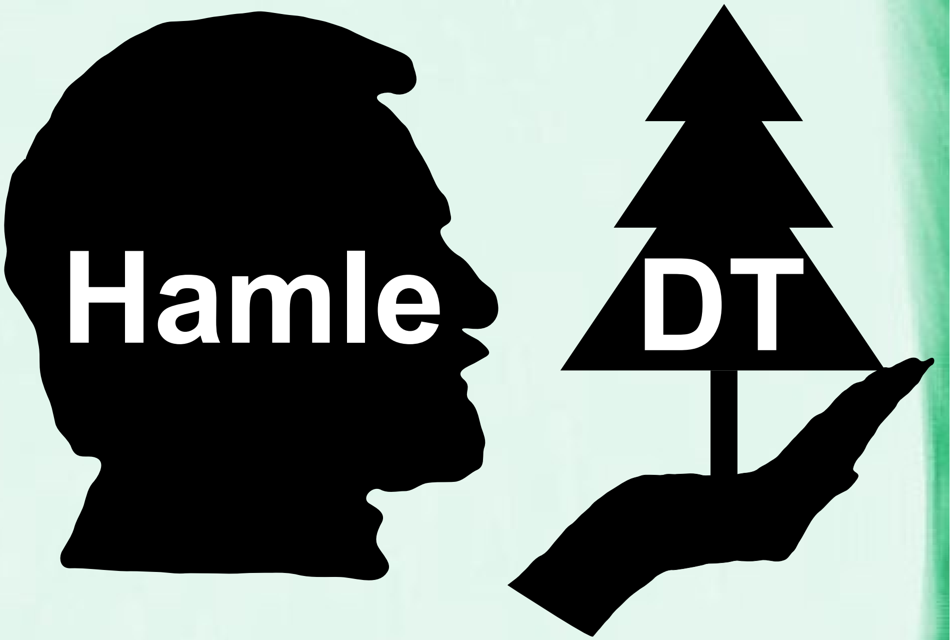




“Save money, don't phone, use fax.”

PDT 2.0

HamleDT v1.0 collection of treebanks



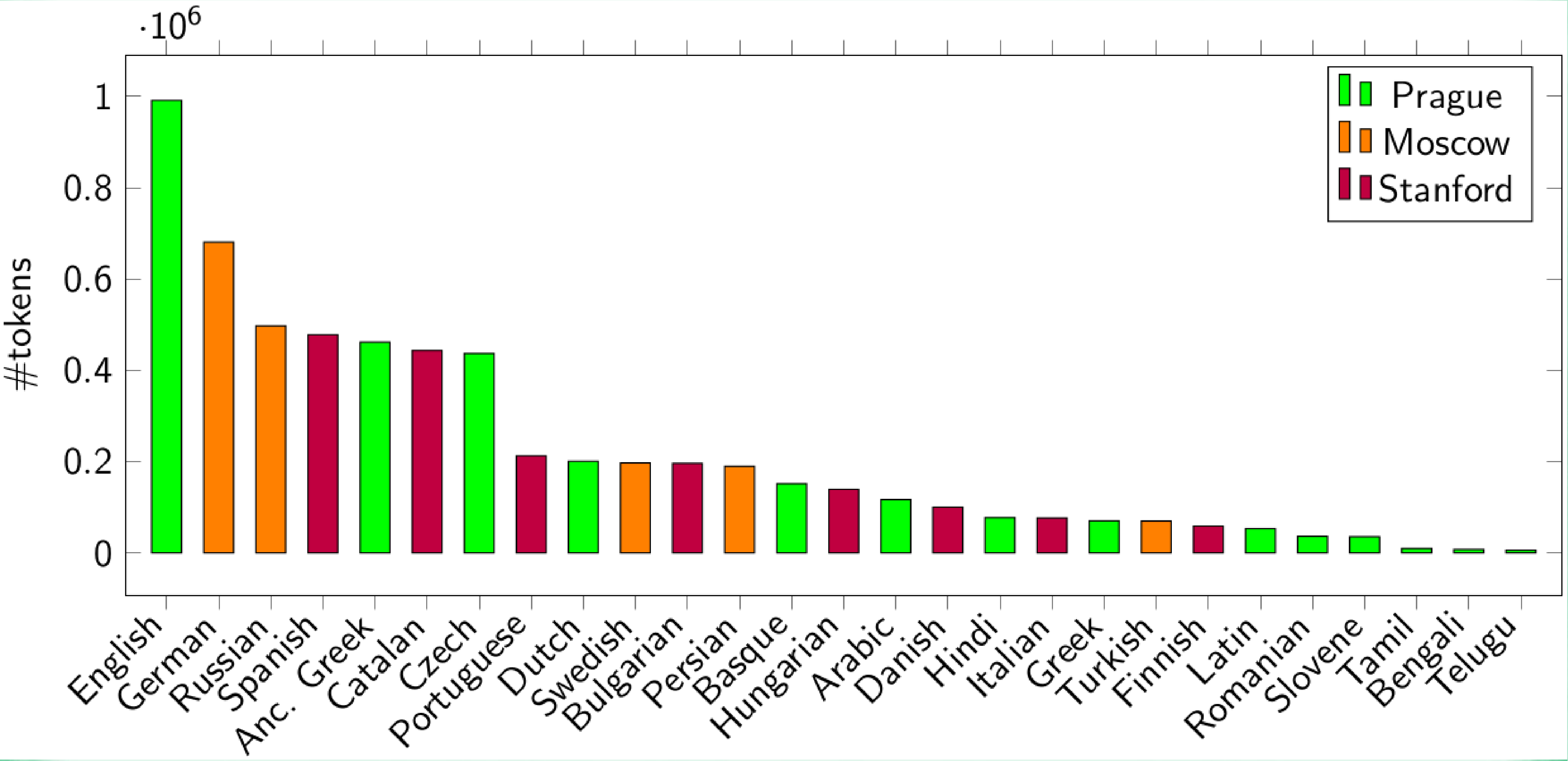
- HArmonized Multi-LanguagE Dependency Treebank
  http://ufal.mff.cuni.cz/hamledt/

- Sources: CoNLL, ICON, other
- We tried to harmonize also:
  prepositions, determiners,
  subordinated clauses, punctuation
- We plan to harmonize:
  verb groups, tokenization, …
- Recent “competitor”: Google Universal Treebanks

HamleDT v1.0 statistics

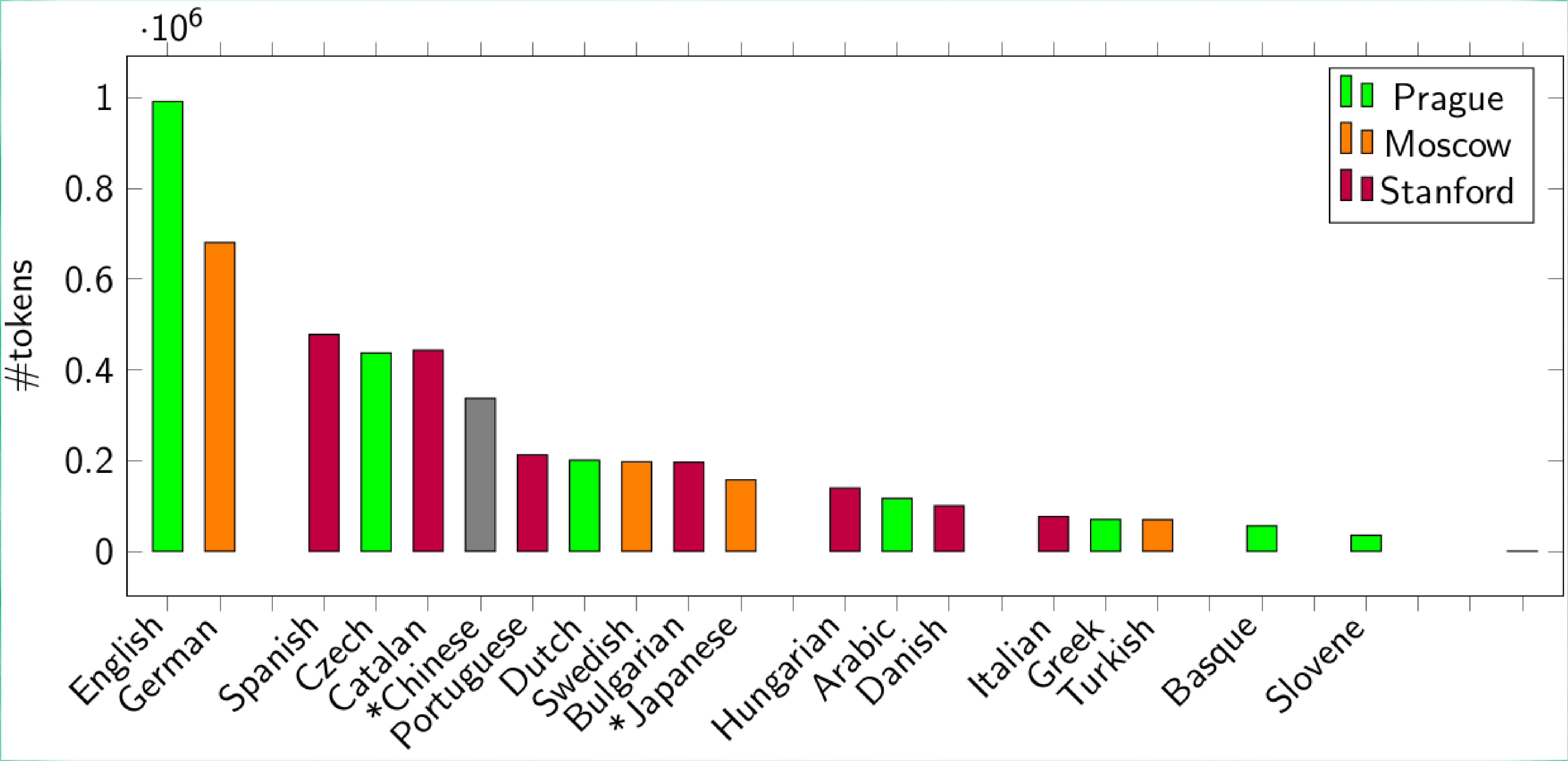
Language	Orig. type	Data set	Sents.	Tokens	Original CS style code	CSs / 100 tok.	CJs / CS	SMs / CS	Nested CS[%]	RT UAS
Ancient Greek	dep	prim.	31 316	461 782	fP hR sH cH pB dL m11	6.54	2.17	0.16	10.3	97.86
Arabic	dep	C07	3 043	116 793	fP hL sH cH pB dL m00	3.76	2.42	0.13	10.6	96.69
Basque	dep	prim.	11 225	151 593	fP hR sN cH pP dU m00	3.37	2.09	0.03	5.1	99.32
Bengali	dep	I10	1 129	7 252	fP hR sH cH pP dU m11	4.87	1.71	0.05	24.1	99.97
Bulgarian	phr	C06	13 221	196 151	fS hL sN cB pB dU m10	2.99	2.19	0.00	0.0	99.74
Czech	dep	C07	25 650	437 020	fP hR sH cH pB dL m11	4.09	2.16	0.20	14.6	99.42
Danish	dep	C06	5 512	100 238	fS* hL sN cP pB dU m10	3.68	1.93	0.13	7.5	99.76
Dutch	phr	C06	13 735	200 654	fP hR sN cH pP dU m10	2.06	2.17	0.05	3.3	99.47
English	phr	C07	40 613	991 535	fP hR sH cH pB dU m10	2.07	2.33	0.05	6.3	99.84
Finnish	dep	prim.	4 307	58 576	fS hL sN cB pB dU m10	4.06	2.41	0.00	6.4	99.70
German	phr	C09	38 020	680 710	fM hL sN cP pP dU m10	2.79	2.09	0.01	0.0	99.73
Greek	dep	C07	2 902	70 223	fP hR sH cH pB dL m11	3.25	2.48	0.18	7.2	99.43
Hindi	dep	I10	3 515	77 068	fP hR sH cH pP dU m11	2.45	1.97	0.04	10.3	98.35
Hungarian	phr	C07	6 424	139 143	fT hX sN cX pX dL m00	2.37	1.90	0.01	2.2	99.84
Italian	dep	C07	3 359	76 295	fS hL sN cB pB dU m10	3.32	2.02	0.03	3.8	99.51
Latin	dep	prim.	3 473	53 143	fP hR sH cH pB dL m11	6.74	2.24	0.41	12.3	97.45
Persian	dep	prim.	12 455	189 572	fM*hM sN cB pP dU m00	4.18	2.10	0.18	3.7	99.82
Portuguese	phr	C06	9 359	212 545	fS hL sN cB pB dU m10	2.51	1.95	0.26	11.1	99.16
Romanian	dep	prim.	4 042	36 150	fP* hR sN cH p0 dU m10	1.80	2.00	0.00	0.0	100.00
Russian	dep	prim.	34 895	497 465	fM hL sN cB p0 dU m10	4.02	2.02	0.07	3.9	99.86
Slovene	dep	C06	1 936	35 140	fP hR sH cH pB dL m00	4.31	2.49	0.00	10.8	98.87
Spanish	phr	C09	15 984	477 810	fS hL sN cB pB dU m10	2.79	1.98	0.14	12.7	99.24
Swedish	phr	C06	11 431	197 123	fM hL sN cF pF dU m10	3.94	2.19	0.13	0.7	99.66
Tamil	dep	prim.	600	9 581	fP hR sH cH pB dL m11	1.66	2.46	0.22	3.8	99.67
Telugu	dep	I10	1 450	5 722	fP hR sH cH pP dU m11	3.48	1.59	0.06	5.0	100.00
Turkish	dep	C07	5 935	69 695	fM hR sN cB pB dL m10	3.81	2.04	0.00	34.3	99.23

HamleDT v1.0

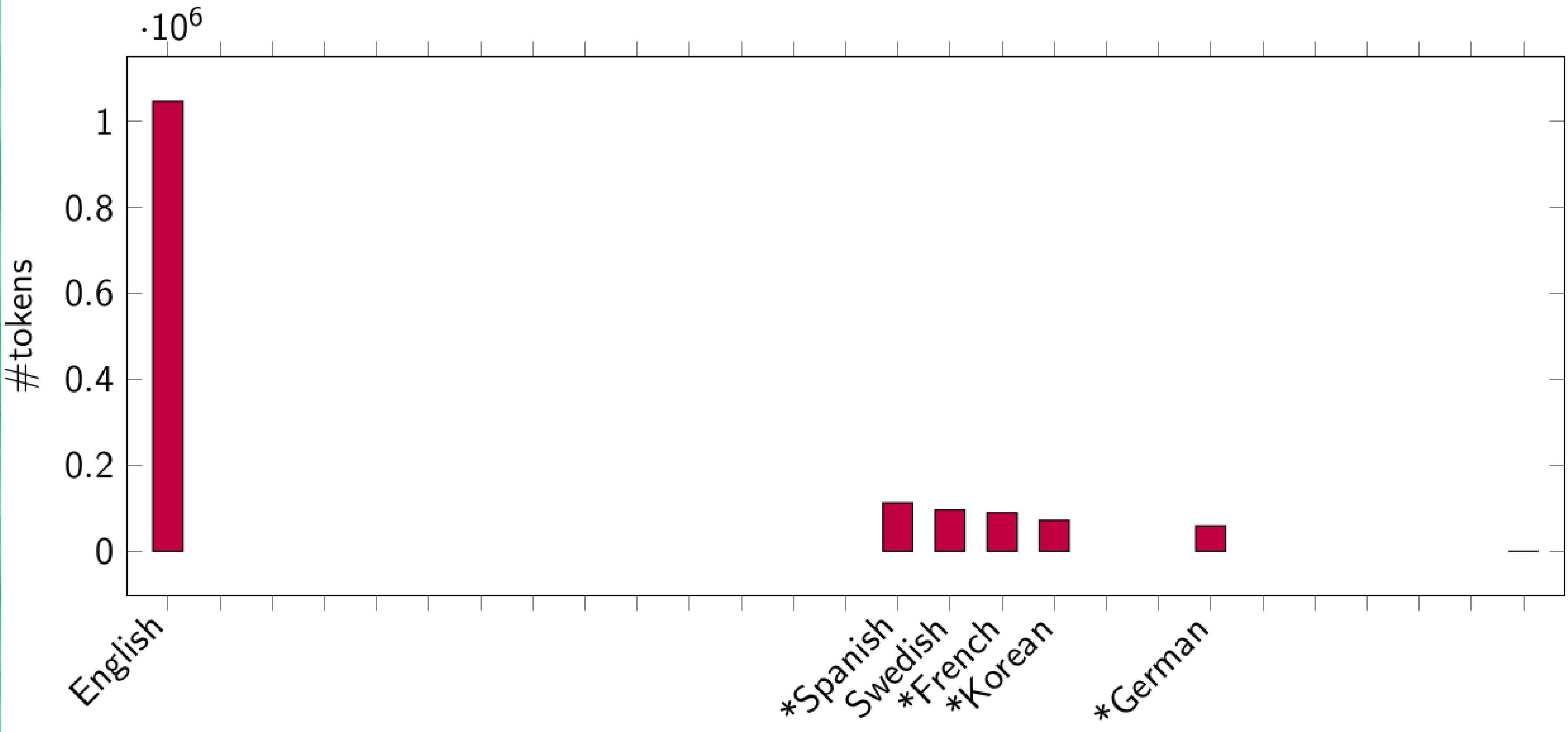


·10^6
#tokens
1
0.8
0.6
0.4
0.2
0
🟩 Prague
🟧 Moscow
🟥 Stanford
English
German
Russian
Spanish
Anc. Greek
Catalan
Czech
Portuguese
Dutch
Swedish
Bulgarian
Persian
Basque
Hungarian
Arabic
Danish
Hindi
Italian
Greek
Turkish
Finnish
Latin
Romanian
Slovene
Tamil
Bengali
Telugu

CoNLL (2006-2010)



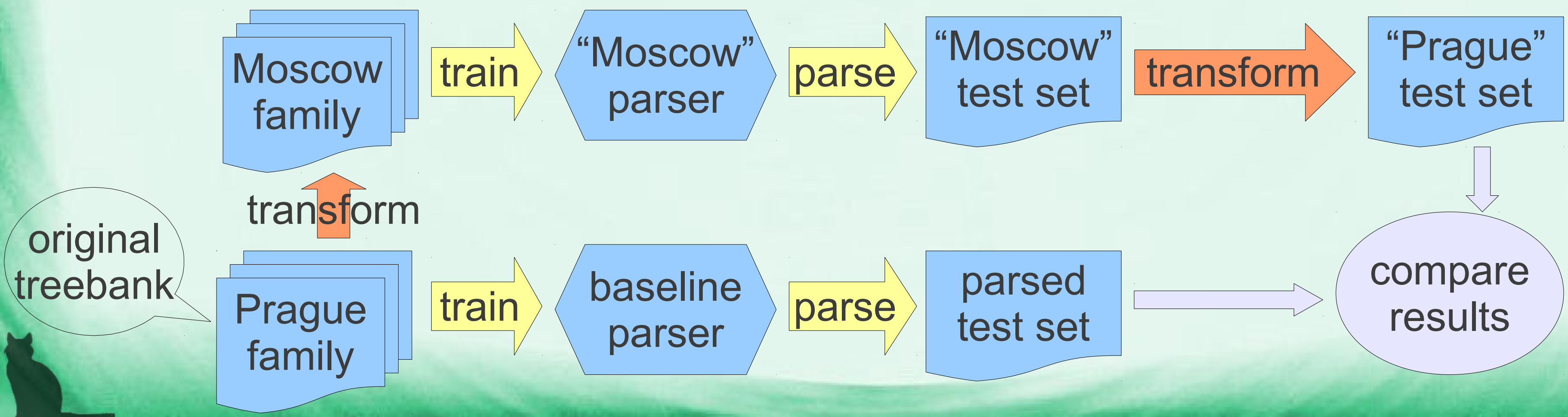
Google Universal Treebank v1.0



Current / Future work

- HamleDT 1.5 (29 languages, done)
- HamleDT 2.0 (Rudolf Rosa, Jan Mašek)
  - More consistent, bigger, more languages (Hebrew, Polish, Korean, French, Northern Sami,...)
  - Stanford dependencies instead Afun
  - English translations and alignments (Google Translate)
- Experiments with parsers and learnability

  Different styles may be better for different parsers.



Thank you

Questions?