Pravidlová extrakce slovesných kolokátů z parsovaného korpusu

Silvie Cinková

Workshop tří projektů

UFAL, 15.4. 2013

Strukturní i sémantické variabilita užití

- *The wooden chair gave a frightened squeak.*
- *Mom gave me a cookie.*
- *The results gave them quite a shock.*
- *Joanna gave her a disgusted look.*
- *The audience gave him the raspberry*.
- *Finally, they had to give up*.

Motivace

- Rozpoznávání patternů (daných lexikonem)
  - Morfosyntaktická definice patternu
  - Sémantická podobnost kolokátů na stejné syntakt. pozici

6 **[Route|Watercourse|Aperture|Building|Natural Landscape Feature {trail|path|door|window|hallway|...} access [Location|Building]**
[Route|Watercourse|Aperture|Building|Natural Landscape Feature] leads to [Location|Building]

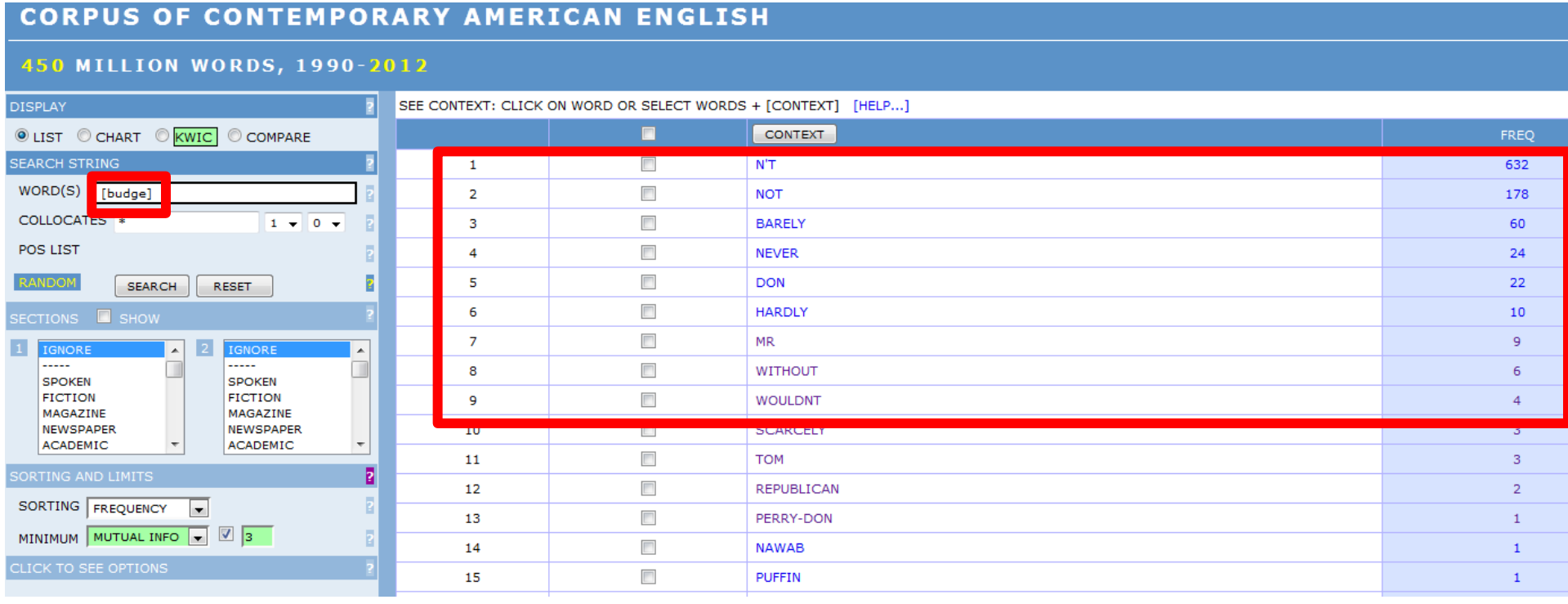
7 **[Human | {student|pupil|...}] access [[Activity | {seminar|programme|course|...}]^[{school|college|university|academy|...}]]**
[Human] joins or enrolls in an [Activity]|{school|college|university|academy|...}

8 **[Human|Animal] access [Physical Object]**
[Human|Animal] approaches and handles [Physical Object]
It is complicated and care must be taken

#agrs	AL	JT	SC	EK	multi	adj	
0	x	2.a	2	1.a	1.a,2,2.a,x	2.a	Directly **accessing** the experience of service users avoids inadvertently confounding measu owes more to staff activity (in this case , in setting achievable goals) than to real differenc
1	x	1.a	1	1	1,1.a,x	1.a	**Accessing** information would be improved by showing the file location in the index , and ma
1	u	5.a	u	1.a	1.a,5.a,u	u	A huge , thermionic valve-based , test-rig to **access** their performance has now been by-pa
1	x	x	u	3.f	3.f,u,x	u	The ` kids ' get to air their ` views ' to the ` programme ' makers in THE WORD : **ACCESS**
1	4	3	4.a	4	4,4.a	4.a	There are a variety of different ways in which an individual may **access** the hospital system

Které kolokáty nás zajímají

- Argumenty a volná doplnění, negace
- Slovesní sourozenci v koordinacích
- Gramatické kategorie u slovesa i kolokátů
- Větný druh (oznamovací, rozkazovací…)



Jak se pozná sémantická podobnost

- Rozpoznání pojmenovaných entit
- WordNet
- Vlastní ontologie (BSO, Omega, DeepDict entities)
- Distribuční sémantika – distribuční model
  - Kvantitativní informace o distribuční podobnosti mezi každými dvěma slovy
  - Možnost generalizace z většího korpusu

Extrakce stejné syntaktické pozice

- Pattern: oznamovací věta hlavní, většinou aktivní
- Pevný slovosled (neutrální)
- „Canonical Sequence“

→ seřadíme větné členy do „Canonical Sequence“

Nezávisle na významu slovesa

– Strukturální ambiguita větných členů

-> nemůžeme věřit parseru

They called him[OBJacc/OBJdat] *an idiot*[OBJ/PreNom/OBJ-acc]

They gave him[OBJacc/OBJdat] *food.*

They called him[OBJacc/OBJdat] *a doctor*[OBJ/PreNom/OBJ-acc]

.

They gave the flowers to the girls.
They sent him to Prague.

1. Analysis is easy for humans, because they consider the meaning of each verb.
2. The parser does mostly not.
3. **We need the same extraction rules for all verbs!**

Příklady

- *The car (that was) stolen yesterday*
- *A growing problem*
- *It was John who broke the window*
- *The topic I was interested in*
- *They gave me a picture*
- *They gave a picture to me*
- *He is easy to please*
- *I find it interesting that babies have to burp*
- *“Certainly”, he said.*
- *What do you fear?*

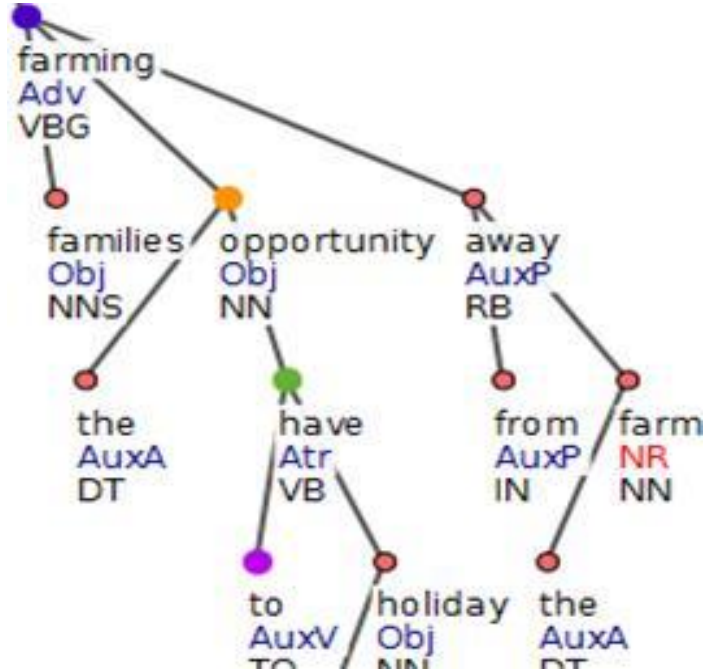
Doplnění

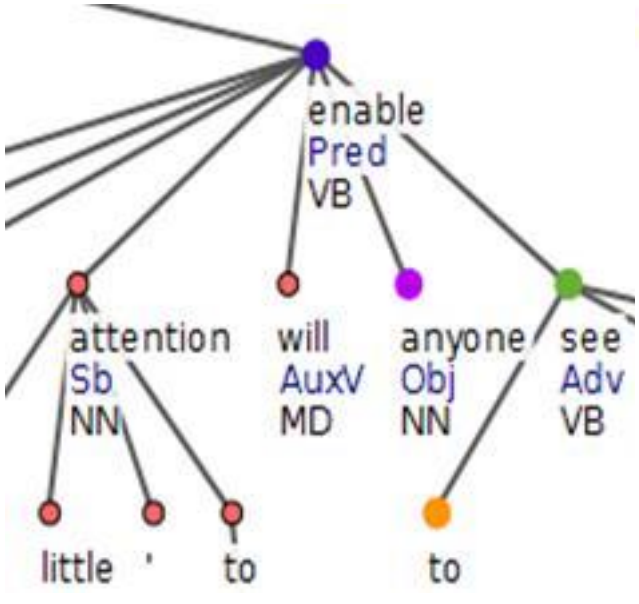
- Může tam být realizováno?
- Je tam?
- Jaké má lemma?

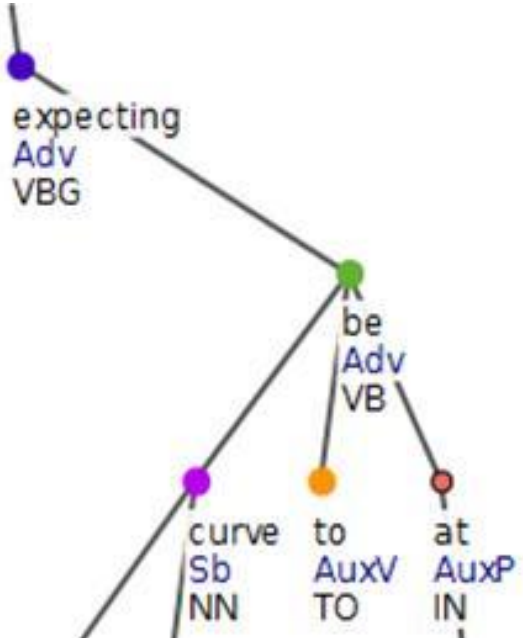
Strukturální mnohoznačnost

Mostly random parse solutions

- *... persuaded the visitor to leave*
- *… worked enough days to deserve one day off*
- *... shut the door to hide*
- *... hated the woman to go*
- *... became the first player to score*







Kanonická posloupnost

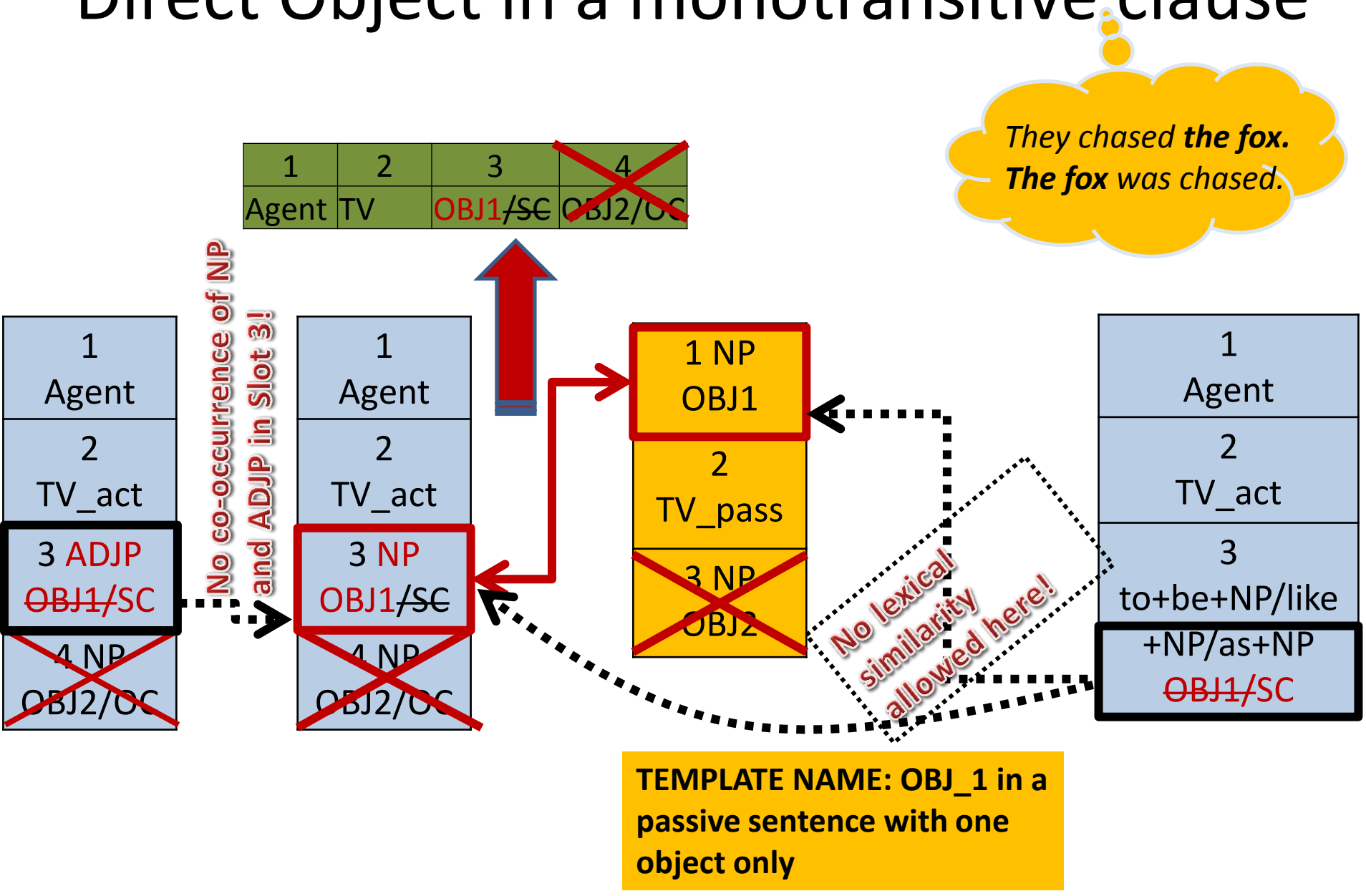
1	2	3	4	5
Agent	TV	OBJ1/SC	OBJ2/OC	Prep+NP/ADV/RP/NPquant

Bez sémantické informace často neumíme přiřadit syntaktickou nálepku,
Ale aspoň víme, jaké je pořadí členů.

Clause templates (CLT)
Klauzální šablony

- Pravidelné gramatické transformace Kanonické posloupnosti
  - Pasívum, přívlastkové věty, cleft sentences, participia…
- PMLTQ dotaz hledá cílové sloveso a jeho doplnění na pozicích daných příslušnou klauzální šablonou

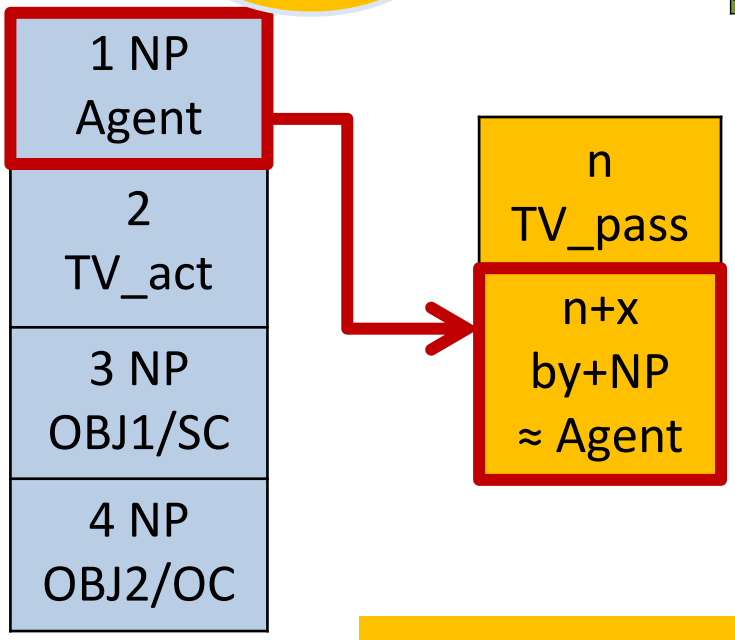
Direct Object in a monotransitive clause



Passive by-phrase as Agent

A cockatoo *bit John.*
John was bitten ***by a cockatoo****.*
~~*The deadline was missed* ***by a week***~~*.*

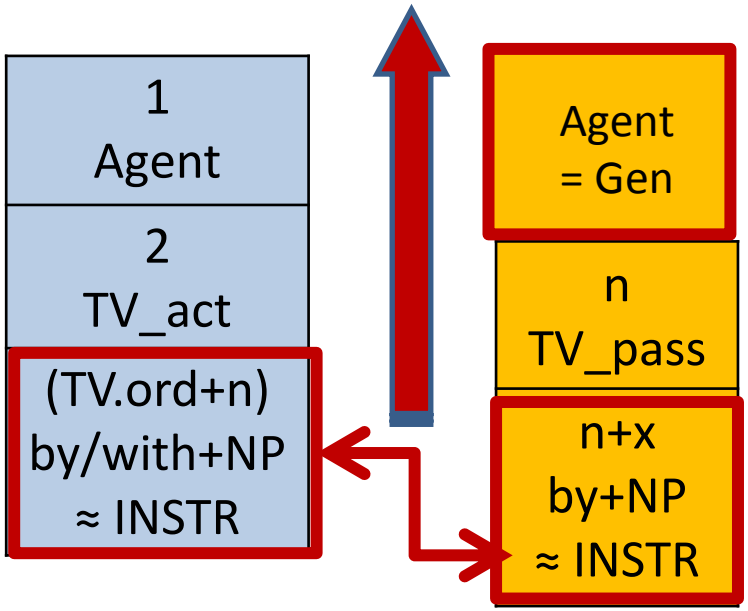
1	2	3	4	5
Agent	TV	OBJ1/SC	OBJ2/OC	(TV.ord+n) by/with+NP ≈ INSTR



TEMPLATE NAME: BY-PHRASE_PASSIVE IN PASSIVE CLAUSES

Passive by-phrase as Instrumental Adverbial

1	2	3	4	5
Agent	TV	OBJ1/SC	OBJ2/OC	(TV.ord+n) by/with+NP ≈ INSTR

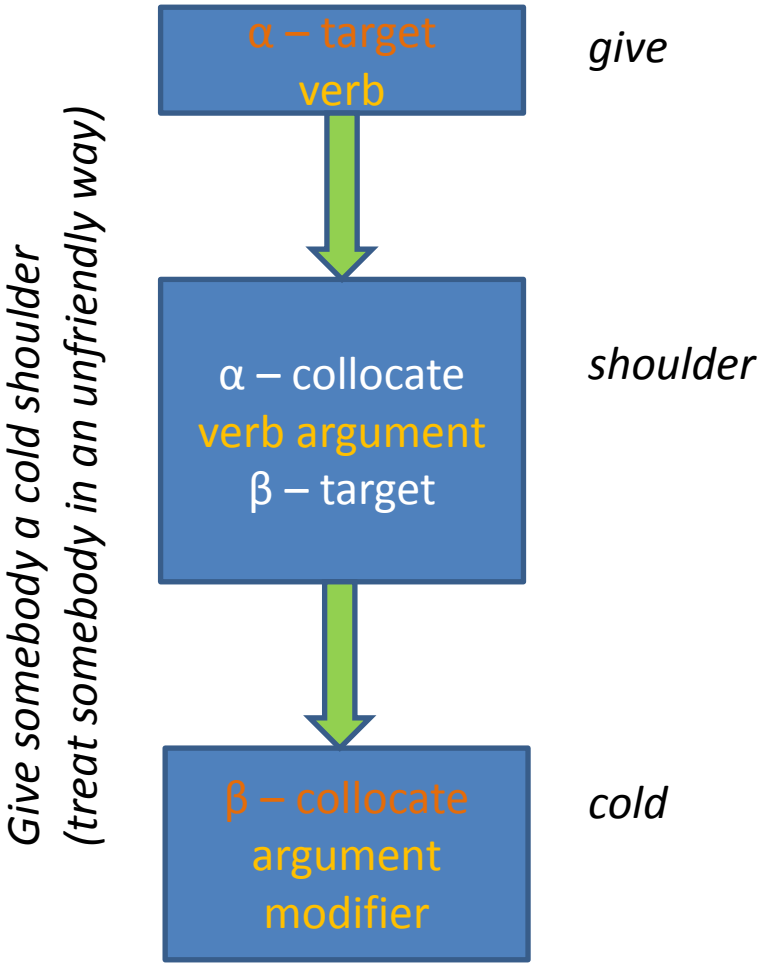


They broke the window ***by a hammer.***
The window was broken ***by/with a hammer.***

TEMPLATE NAME: BY-PHRASE
IN ACTIVE CLAUSES

TEMPLATE NAME: BY-PHRASE_PASSIVE
IN PASSIVE CLAUSES

Relations between verb, arguments and their modifiers



Targets & collocates ≈ SYNTAGMS

- No matter the number of tokens
- They can be defined by:
  - Morphological features (e.g. 'genitive')
  - Lexical features (lemma/form)

QUERY TEMPLATES

- Verb-Form Templates (VFT): α - target
- Verb-Argument Templates (VAT): α – collocates
- Noun-Argument Templates (NAT): β - collocates
- Adjective-Argument Templates (AAT): γ - collocates
- Supra-Clause Templates (SCT):
  - combines VFT, VAT, NAT, AAT
  - syntactic labeling (⇨ see Sentence Position Model)

Verb Form Templates (VFT)

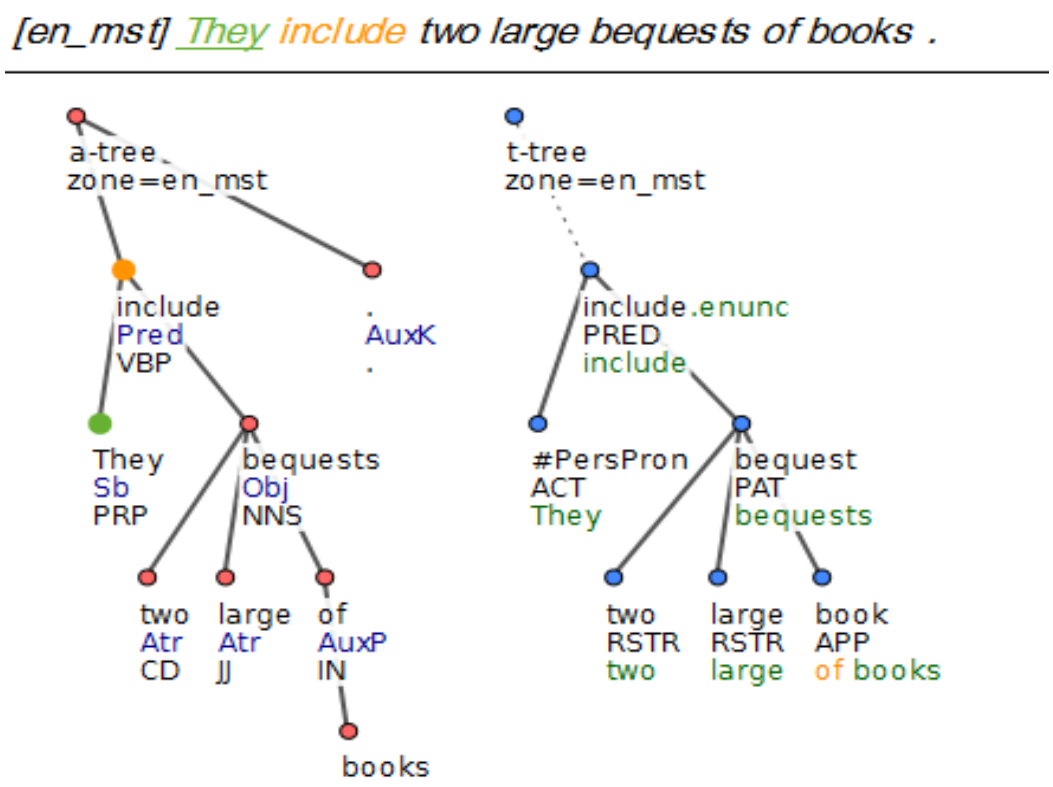
- **“instead of grammatemes”**
- different combinations of grammatical verb categories get their respective labels
  - finite x infinitive x participle
  - present x past
  - prespast x perfect
  - simple x progressive
  - auxdo
  - member of coordination (left, right, none)

Verb Argument Templates (VATs)

- NP
- NP POS
- NP DT
- NP_quant
- NP WH coref REL
- NP WH coref pseudocleft
- Expletive *it* is subject of active *that*-clause
- Expletive *it* is subject of active *wh*-clause

COL_TV_NP

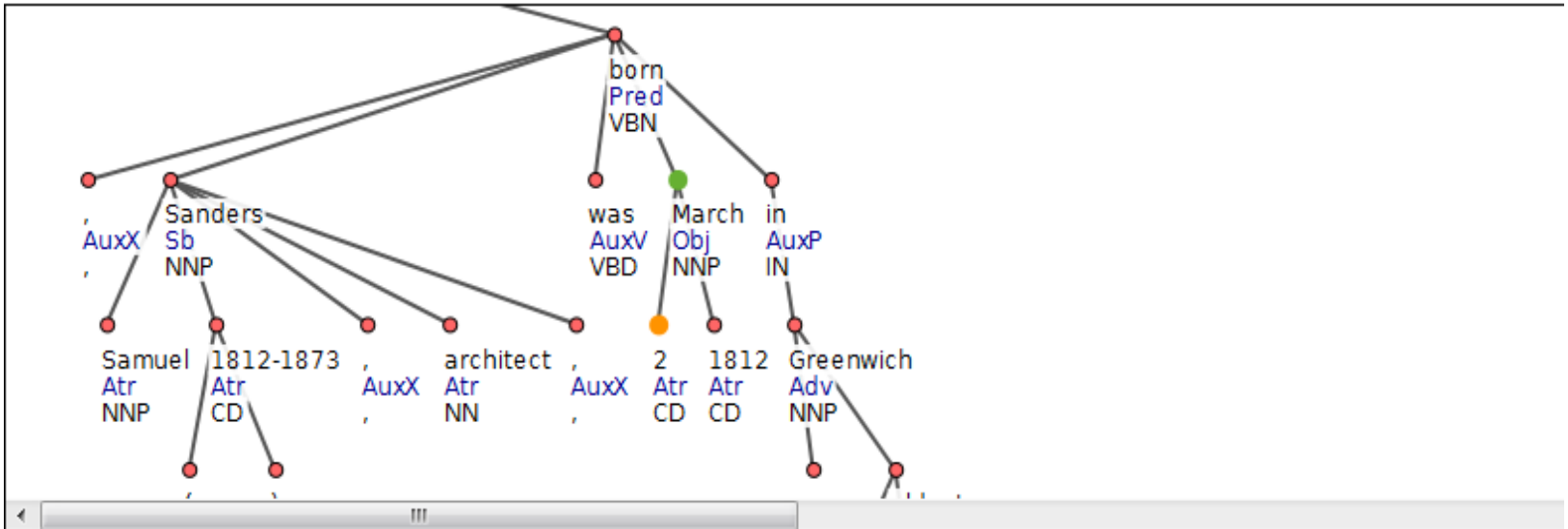
- Nouny word (Noun, non-relative nouny wh-word, personal pronoun, numeral) as an α-collocate



COL_TV_NP_quant

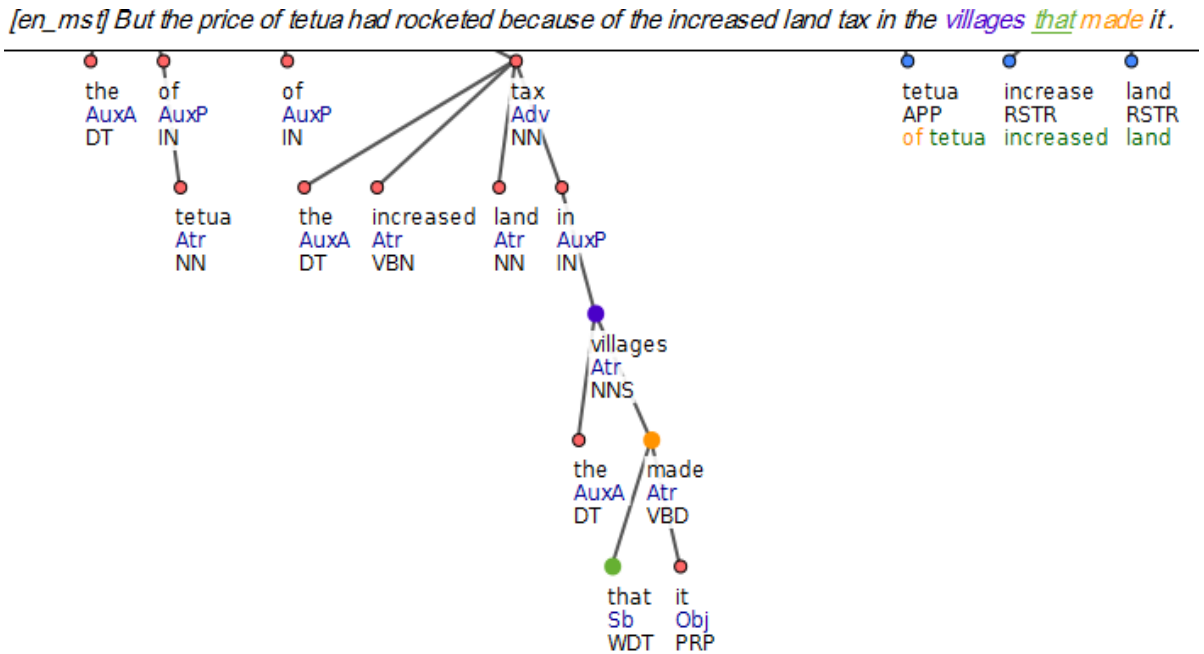
- A list of tags and lemmas not governed by a preposition that are not objects
  - Dates, weekdays, names of months, holidays
  - Measuring units, e.g. *km, mile*
- Of course not complete!

[en_mst] , Samuel Sanders (1812-1873) , architect , was born 2 March 1812 in Greenwich , the eldest of the four sons of Samuel Teulon , cabinet-maker and upholsterer (later a surveyor) of Greenwich , and his wife Louisa Sanders from Rotherhithe .



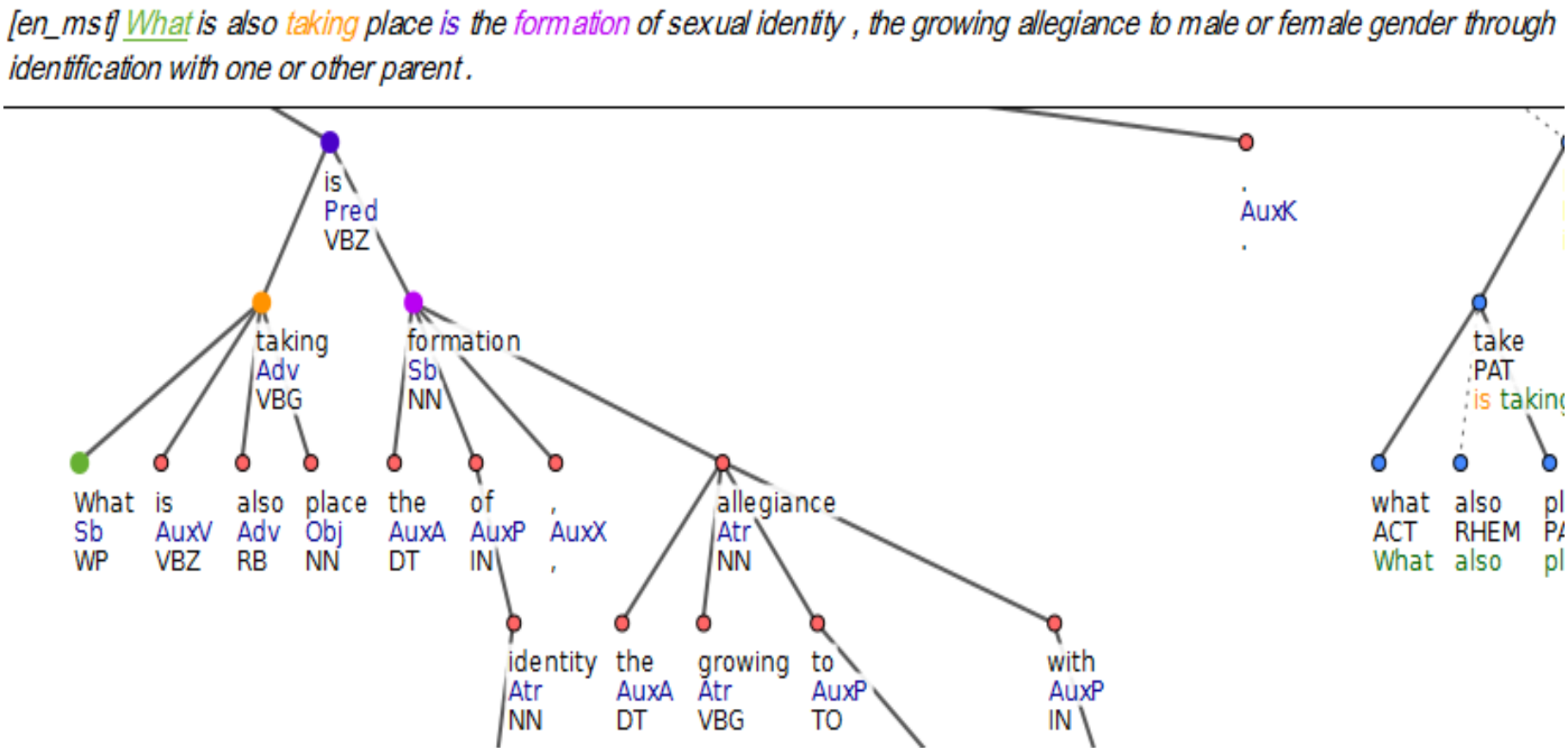
COL_TV_NP_WH_coref_rel

- Relative words with grammatically predictable coreference in an attributive sentence (i.e. governed by a noun)
  - Except *when, how, where, why*, because their antecedents are not always straightforward
- The antecedent word identified and labeled $ANTECEDENT



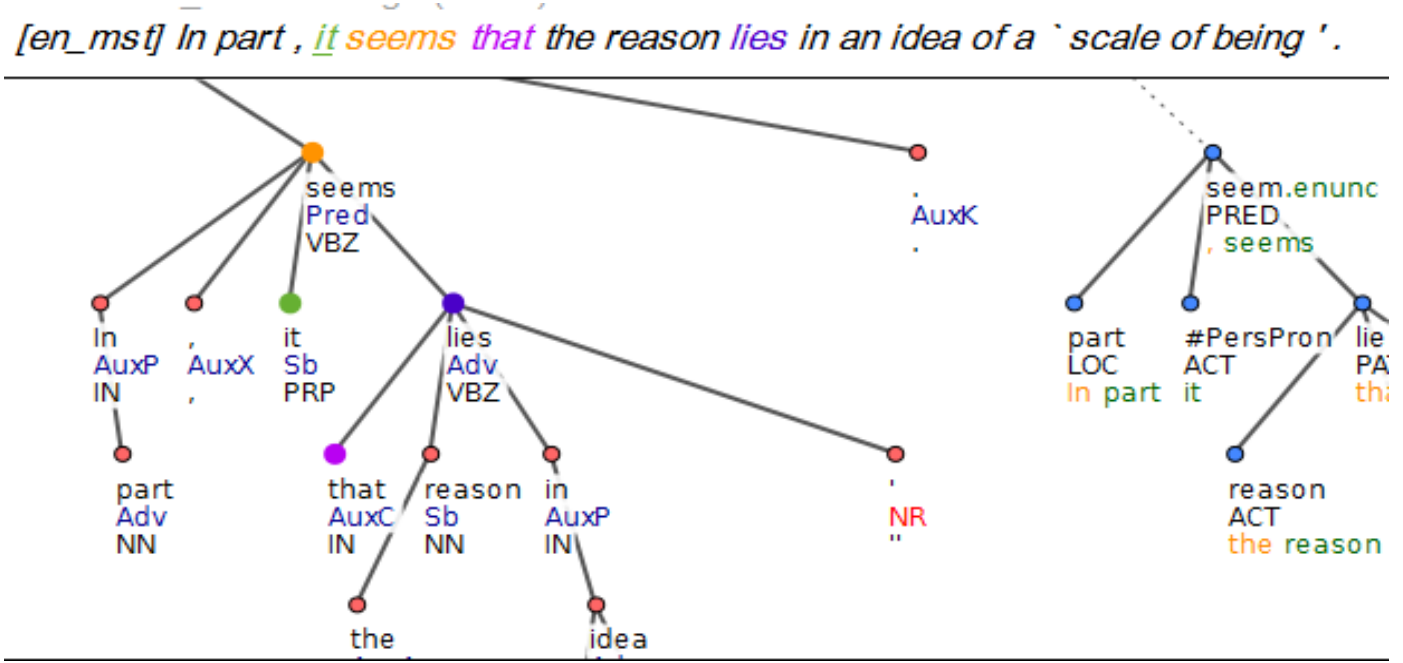
COL_TV_NP_WH_coref_pseudocleft

- Relative pronoun in a pseudocleft sentence and its antecedent



COL_TV_expl_it_issubj_v_act_thatcl

- Expletive *it* as subject of active-voice verb, *that*-clause follows
  - Not universally applicable with all verbs. Therefore the query contains a list of verbs that do this systematically and are frequent in the corpus: *It seems that…* X *It claims that…*
  - This makes the query less powerful – it won't find any other verbs occurring in this structure



COL_TV_VP_FIN_thatcl

- *That*-clause as an argument of TV