Sémantické kategorie (patterny) a Pattern Dictionary of English Verbs: zkušenosti s manuální anotací

Silvie Cinková

UFAL, 25.1.2012

Corpus Pattern Analysis (P. Hanks)

- Slov(es)a **nemají významy** sama o sobě, ale **mají významový potenciál**, jehož různé **složky se aktivují v různých kontextech**.
- Korpus ukazuje frekventovaná normální použití
- Můžeme z nich vysledovat PATTERNS – typické kombinace morfosyntaktických a taxonomických ukazatelů, které říkají totéž
- Existují lepší a horší příklady (normy a exploatace)

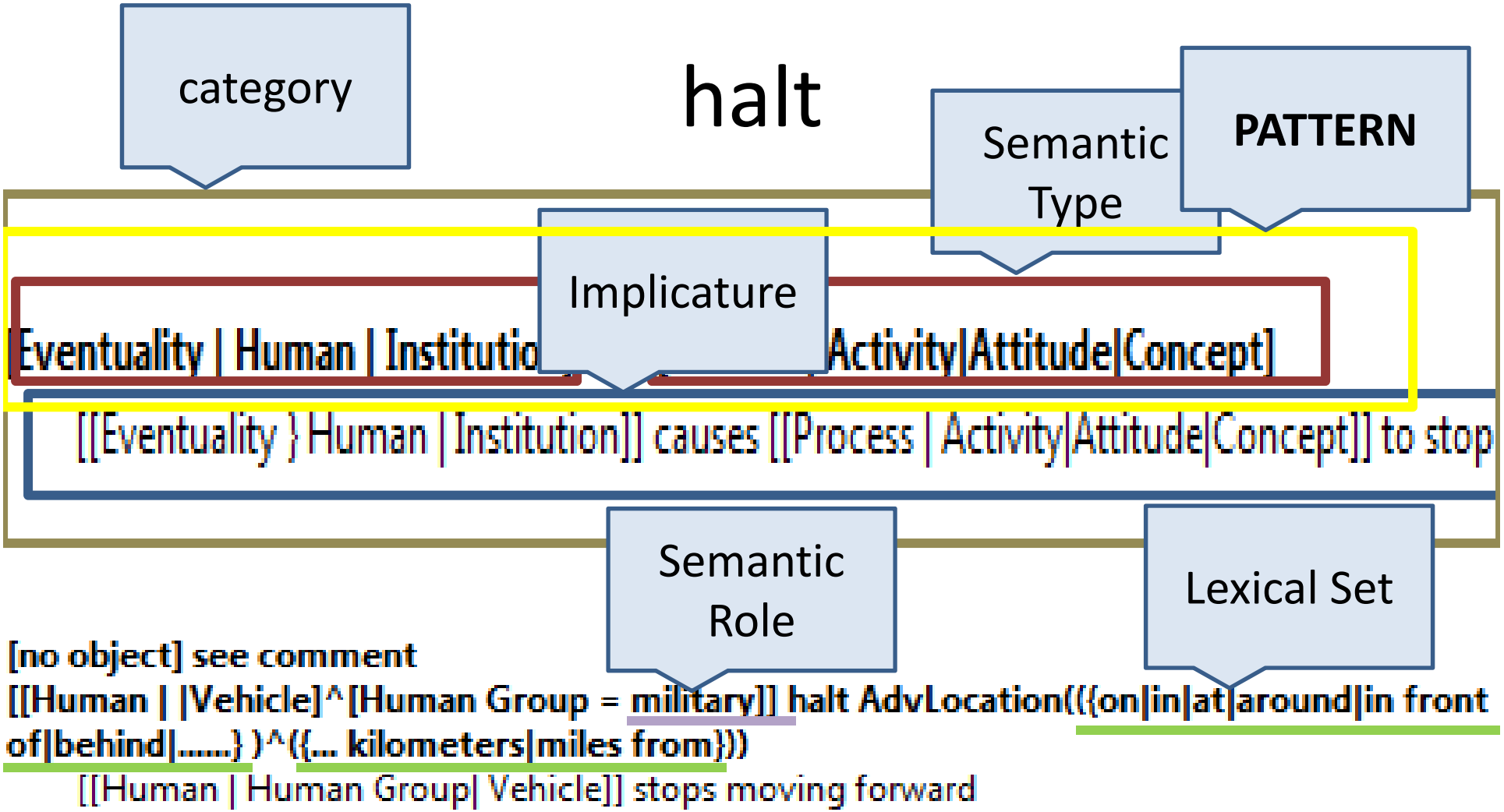
ARCHITEKTURA PDEV

Sémantická konkordance

- termín vymyšlený pro SemCor – korpus anotovaný WordNetem
  - vnáší do jaz. dat znalost světa *(„cukrář je druh člověka“)*
- PDEV je také sémant. konkordance
  - info o vzájemné sémant. podobnosti konkordancí

PDEV - *halt*

llor and the Bank of England were forced to step in to	halt	1	a further dangerous slide by the pound yesterday morni
the already considerable pressure on the university to	halt	1	the sale plan. Yesterday, The Scotsman revealed that th
he government has warned that unless the violence is	halted	1	, the transition from white rule to democracy might be
a good job.' Bill Clinton hoarse after victory Army	halts	1	Bosnian evacuation MORE than 6,000 Bosnians trying
ngineered organisms into the environment; has helped	halt	1.c	a BioWar facility in Utah: and has fought successful co
igwall, will succeed Robert Crawford. Guns fail to	halt	1.c	hot news from the front line Little electricity, less nev
p> The current convulsions began a year ago. They	halted	1.s	as the invasion of Kuwait brought home fears of recess
civilians. At the outset of the fighting, which	halted	1.s	on Oct. 28, both Israel and Lebanon mobilized troops a
230 miles north east of Pearl Harbor, where the fleet	halted	2	. At 0600hrs Hawaii time on December 7 Commar
d by Muslim mounted archers, the army was forced to	halt	2	at Hattin, in a waterless region, in the hope of being abl
ba had paid a brief visit to the UN. On his way back he	halted	2	in Accra. On 8 August he and Nkrumah put their signatu
d example of visual counterpoint, when two trains are	halted	3	on opposing lines in a station. As one train starts to mov
on. He knew of course that he would never be able to	halt	3	the train before it reached the bridge and he had his gu


category
halt
Semantic Type
PATTERN
Implicature
[Eventuality | Human | Institutio Activity|Attitude|Concept]
[[Eventuality } Human | Institution]] causes [[Process | Activity|Attitude|Concept]] to stop
Semantic Role
Lexical Set
[no object] see comment
[[Human | |Vehicle]^[Human Group = military]] halt AdvLocation(({on|in|at|around|in front of|behind|.....})^({... kilometers|miles from}))
[[Human | Human Group| Vehicle]] stops moving forward

halt

1 **[Eventuality | Human | Institution] halt [Process | Activity|Attitude|Concept]**

[[Eventuality } Human | Institution]] causes [[Process | Activity|Attitude|Concept]] to stop

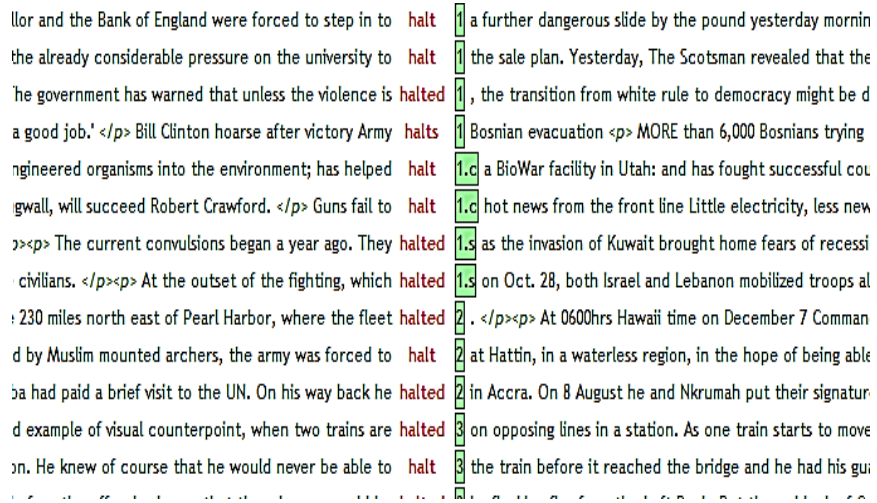
2 **[no object] see comment**
[[Human | |Vehicle]^[Human Group = military]] halt AdvLocation(({on|in|at|around|in front of|behind|.....})^({... kilometers|miles from}))

[[Human | Human Group| Vehicle]] stops moving forward

3 **[[Eventuality]^[Human Group 1 = Military]] halt [[Human]^[Human Group 2 = Military]^[Vehicle]]**

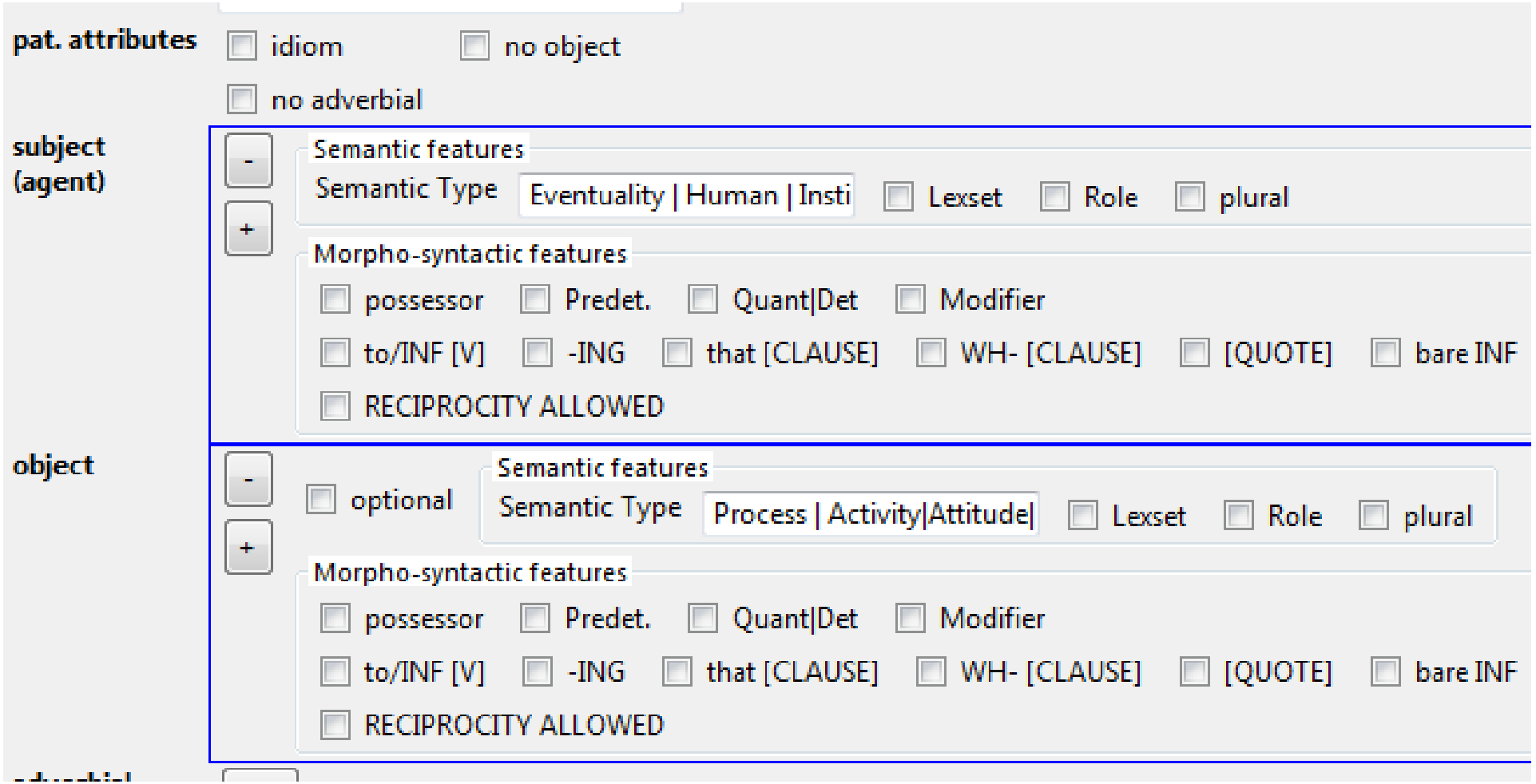
[[Eventuality | Human Group]] causes [[Human | Human Group | Vehicle]] to stop moving forward

Druhy značek

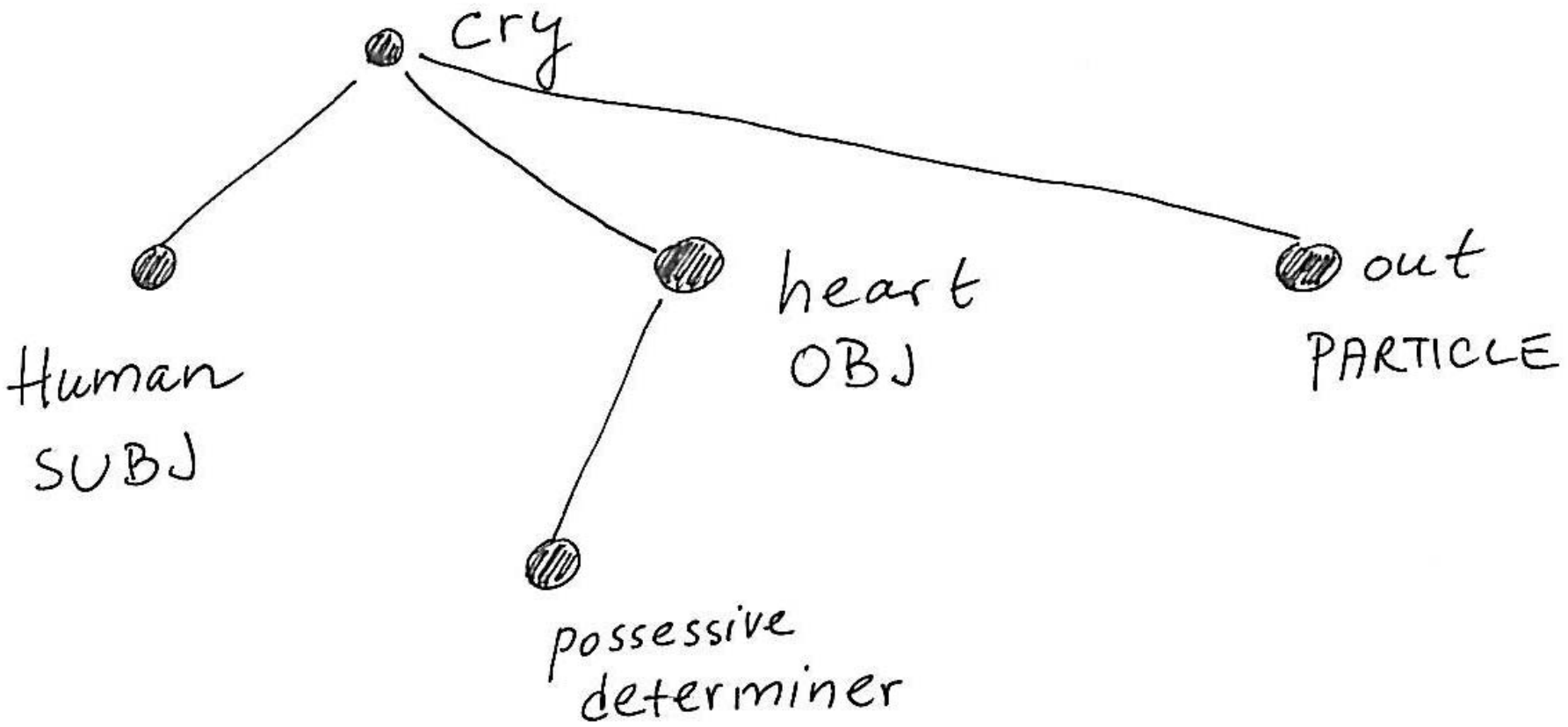


- číslo patternu
- *číslo patternu*.**e = exploatace**
  - metaforický význam, ironie, atd.
  - odchylka v syntaxi, elipsa generického participanta
  - netypický účastník děje (*ride a caterpillar*)
- **x** = toto není sloveso
  - chyba v tagování *(a box of matches)*
  - metaužití *(„read“ means acquire knowledge through written text)*
  - adjektivní/substantivní platnost participia *(-ing, -ed)*
- **u =** je to smysluplně použité sloveso, ale nehodí se žádný existující pattern

halt


pat. attributes
idiom
no object
no adverbial
subject (agent)
-
+
Semantic features
Semantic Type
Eventuality | Human | Insti
Lexset
Role
plural
Morpho-syntactic features
possessor
Predet.
Quant|Det
Modifier
to/INF [V]
-ING
that [CLAUSE]
WH- [CLAUSE]
[QUOTE]
bare INF
RECIPROCITY ALLOWED
object
-
+
optional
Semantic features
Semantic Type
Process | Activity|Attitude|
Lexset
Role
plural
Morpho-syntactic features
possessor
Predet.
Quant|Det
Modifier
to/INF [V]
-ING
that [CLAUSE]
WH- [CLAUSE]
[QUOTE]
bare INF
RECIPROCITY ALLOWED

[Human] cry [POSS {heart}] out



Je CPA užitečné pro NLP?

- Je možné na takto intuitivním rozhodování dosáhnout shody použitelné pro strojové učení?
- Jak vysoká je použitelná shoda?
- Existují typy neshod, které ničemu nevadí?
- Má schopnost vytvářet sémanticky a morfosyntakticky podobné shluky vůbec nějaký přínos pro analýzu/syntézu textu?

Problémy v PDEV

– nízké pokrytí (PH preferoval vzácná slovesa)
– nikdy neměřená shoda
– nesystematické revize dat po revizích hesel
– pravděpodobně nízká konzistence
– málo diferencovaný zápis morfosynt. ukazatelů

SÉMANTICKÁ KONKORDANCE VD-30-EN

Validation database of 30 English verbs (VD-30-En)

Postup

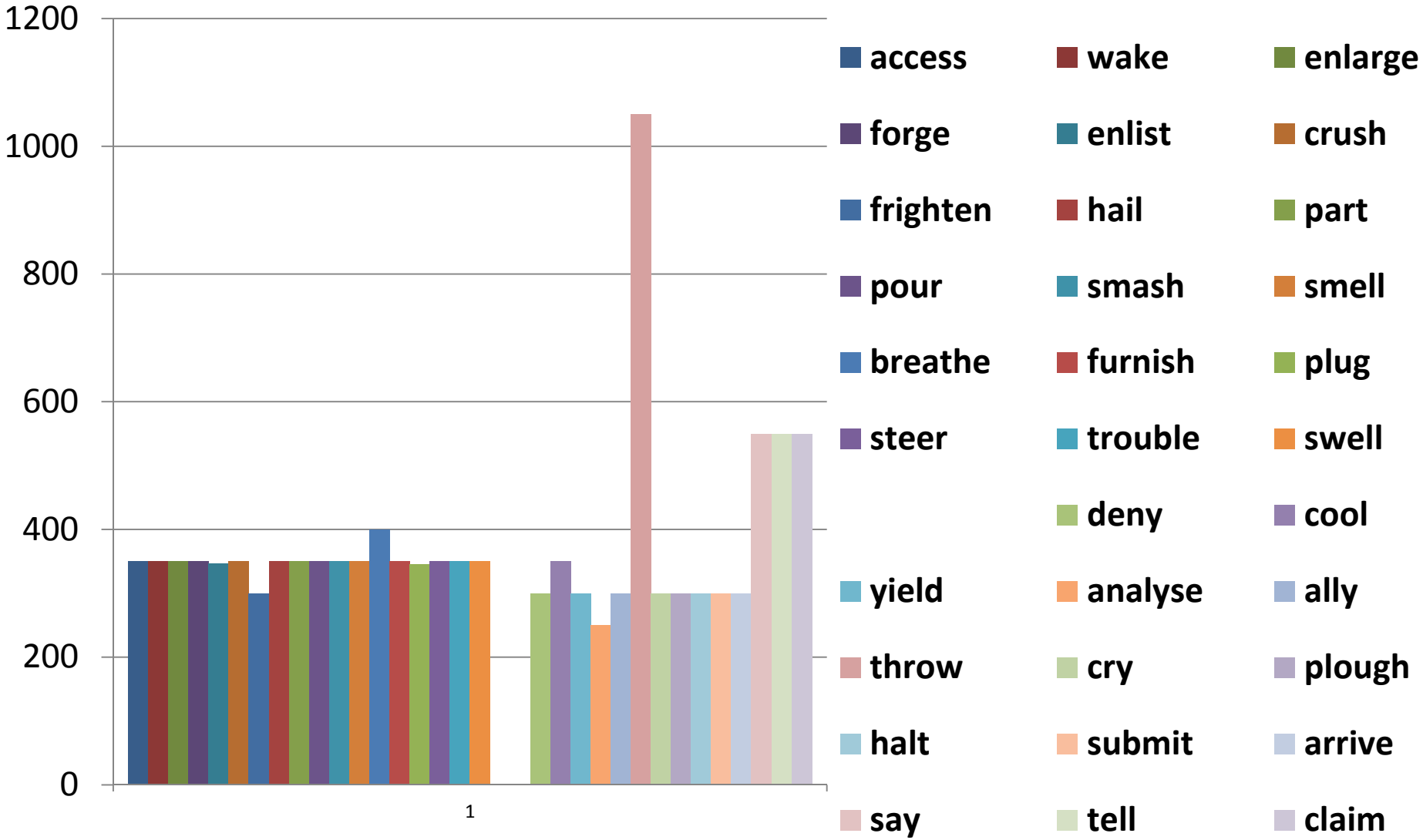
- revize hesla + referenčního vzorku (250+ vět)
- 3 anotátoři anotují náhodný vzorek (50 vět)
- měření shody, matice neshod pro každý pár
- adjudikace nebo další revize hesla a vzorků, nový vzorek k vícenásobné anotaci

Validation database of 30 English verbs (VD-30-En)
Výsledek

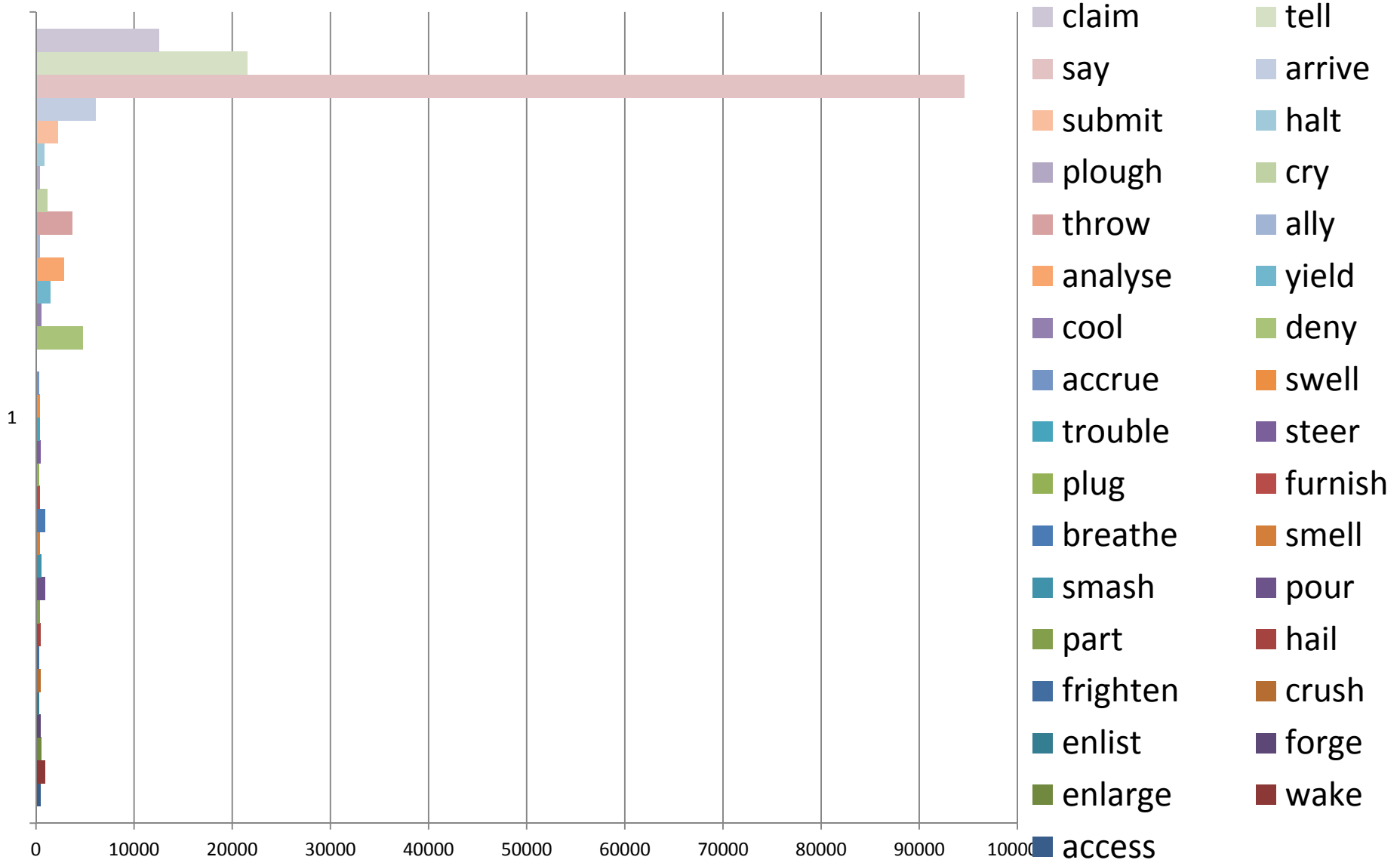
Pro každé sloveso:

- alespoň 300 anotovaných vět, z toho 50 třikrát
- adjudikace posledních 50
  - eliminace chyb
  - výběr „nejlepší značky“
- = ručně CPA-anotovaný korpus 11 993 vět propojených se slovníkem 30 sloves

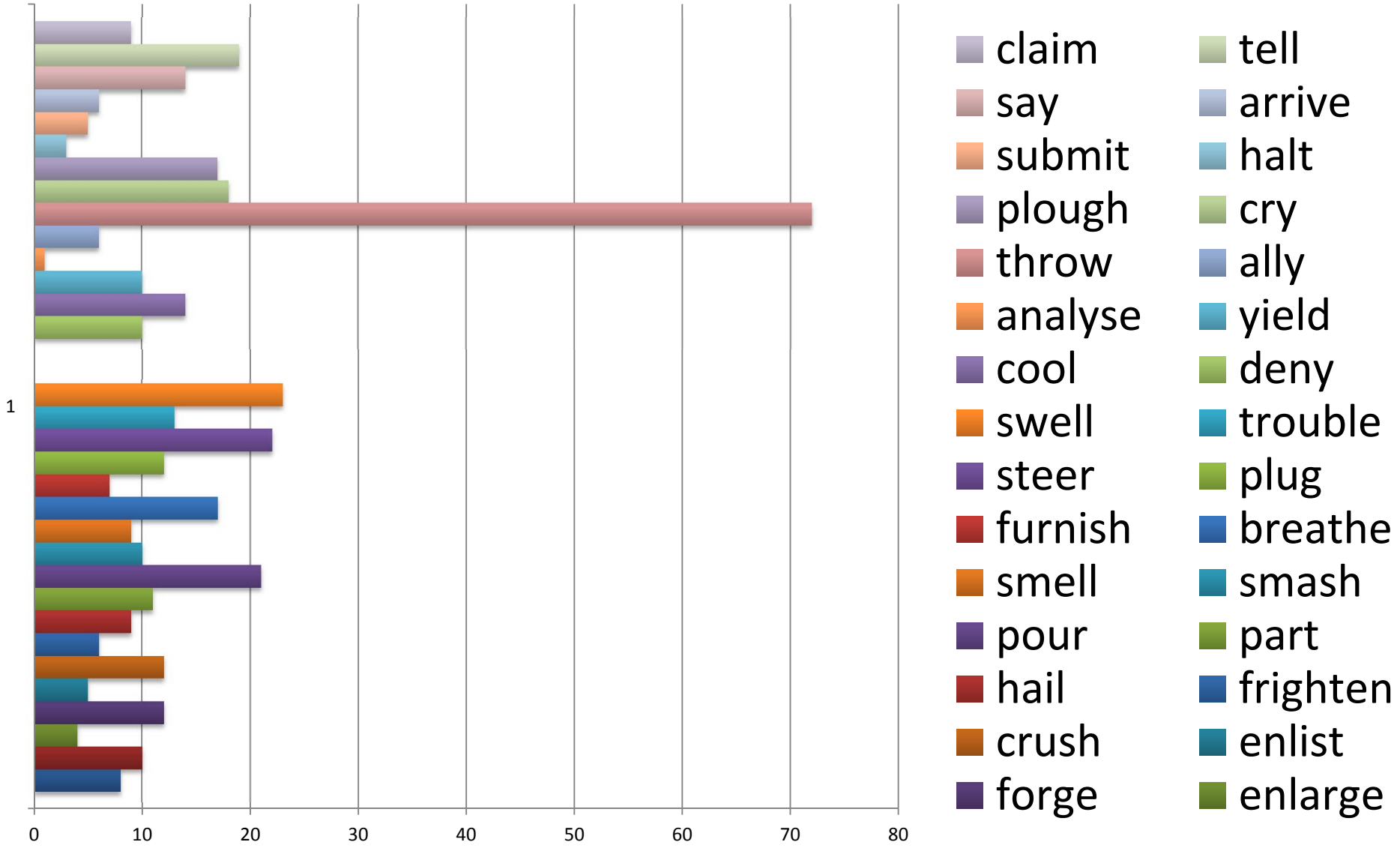
Anotované věty



Frekvence sloves v BNC



Počet patternů v hesle



Lepší a horší příklady v konkordancích

„Norma“

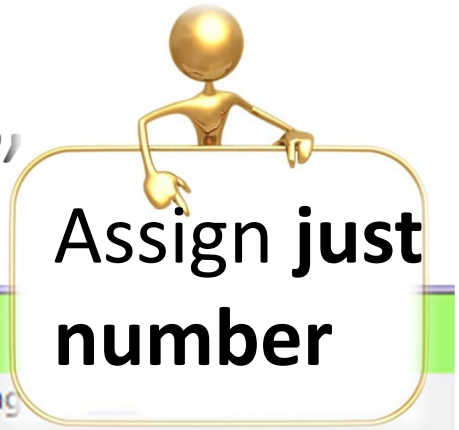
- má stejnou implikaturu jako pattern
- má stejný počet argumentů
- používá stejné předložky
- argumenty odpovídají předepsaným sémant. typům (a rolím)
- nesubstantivní argumenty mají předepsanou formu (např. THAT-CLAUSE)
- argument je v gram. koreferenci
- elipsu/anaforu lze snadno rekonstruovat z kontextu
- argumentem je neurčité zájmeno

„Exploatace“

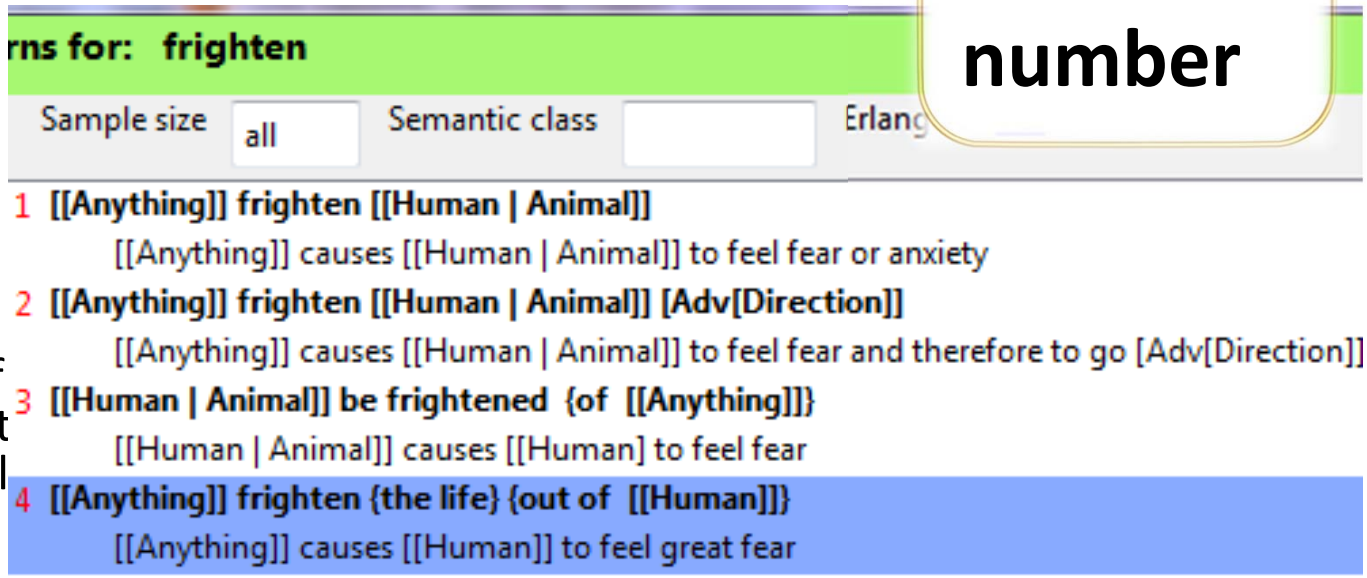
- **semantic coercion** a **transparent heads** *(she drank a glass, enjoyed the soup)*: **.c**
- **anomalous argument .a**
- odlišná syntax: **.s**
  - odlišné předložky a další pomocná slova
  - alternace/diateze
- figurativní použití: **.f**
  - konkrétní se stává abstraktním
  - kreativní metafora

ANALÝZA NESHOD

“Phraseological adverbial”



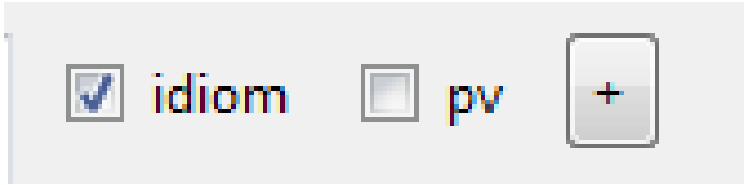
- The construction matches the implicature of a regular (non-idiom) pattern as well as its syntax (having the same configuration of subject/object(s)), but contains an additional phraseological element, typically an adverbial.
- The meaning is slightly modified (augmented, diminished, style register changes)



I said , ` Do you know , you <frightened> me ***to death*** *? ‘*
The Gothic horror <frightens> us ***out of our skin****.*

Verb does matches Pattern 1, as both the concordance and the pattern require agent (who frightens) and patient (who gets frightened).

Idioms not captured by lexicographer (I)



1. Check idiom patterns

- Implicature does not match
- Different number of objects
- Different distribution of thematic roles
- Different Lexical Sets/Semantic Types
- Different function words
- Additional optional elements

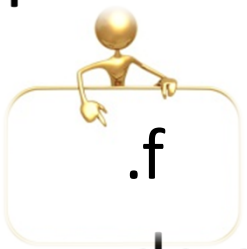
Idioms not captured by lexicographer (II)

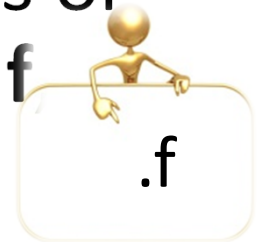
2. **Check regular patterns**

Implicature does not match the **literal** meaning at all

Implicature is too literal, but explains the motivation of the metaphor

Anything wrong about Lexical Sets/Semantic Types, function words or Thematic roles? With **idioms** assign f anyway.





Norma, nebo exploatace ve stejném patternu?

- Některý anotátor si všimne, že argument/pomocné slovo nesedí na předepsaný pattern, a přiřadí **.a** nebo **.s**, ale ostatní si nevšimnou, nebo je to sporné
- **[Human | Institution | Proposition] deny [[Entity | Eventuality]^[Property | THAT-CL = Characteristic feature]^[]]**

= [[Human|Institution|Proposition]] asserts that [Entity | Eventuality| Property = Salient Fact] does not exist, has not occurred, is not true or is irrelevant

Beauty that <denies> its opposite becomes shallow and artificial .

norma, a, f

Chyba anotátora

- záměna tranzitivního patternu za intranzitivní, když je věta v pasívu
- přehlédnutí

Participia

- participium
  - tranzitivní vs. intranzitivní vs. x
  - *The soup was already cooled*: *X cools* vs. *X cools Y*, vs. *X is cooled*.
  - *The high-radiation area was enlarged* vs. *The northern wing of the mansion was enlarged*
- neshodu mezi normálním a participiálním patternem už nepokládáme za neshodu
  - kappu jsme ale nepřepočítávali
  - např. u *enlarge* máme jenom **0,46/65%**, ale 10 z 22 neshod dvou anotátorů s jedním je participiálních. Kdybychom je nepočítali, dostaneme shodu **78%.**

Nejasná koreference

[allegation]... The Iranians <deny> that, but in practice provide safe havens and allow the mujaheddin to come and go .

It 's an indication that they are working on it and what may seem to be an accepted fact one day can be vehemently <denied> the next as the bereaved person comes to terms with the loss .

- **[Human | Institution] deny [Proposition | {charge | allegation|proposition|story|rumour|admission|accusation|report|claim|view|...}]**= [[Human | Institution]] says that [[Proposition]] is not true
- **[Human | Institution] deny [Action | THAT-CL = Bad]** = [Human | Institution] says that he or she did not do [Action Bad]

Sémantická modulace

Evening rush-hour traffic in Liverpool city centre was <halted> during the alert , in which the gunman was spotted looking through the sights at potential targets .

- je *traffic* Activity|Process, nebo Vehicle?

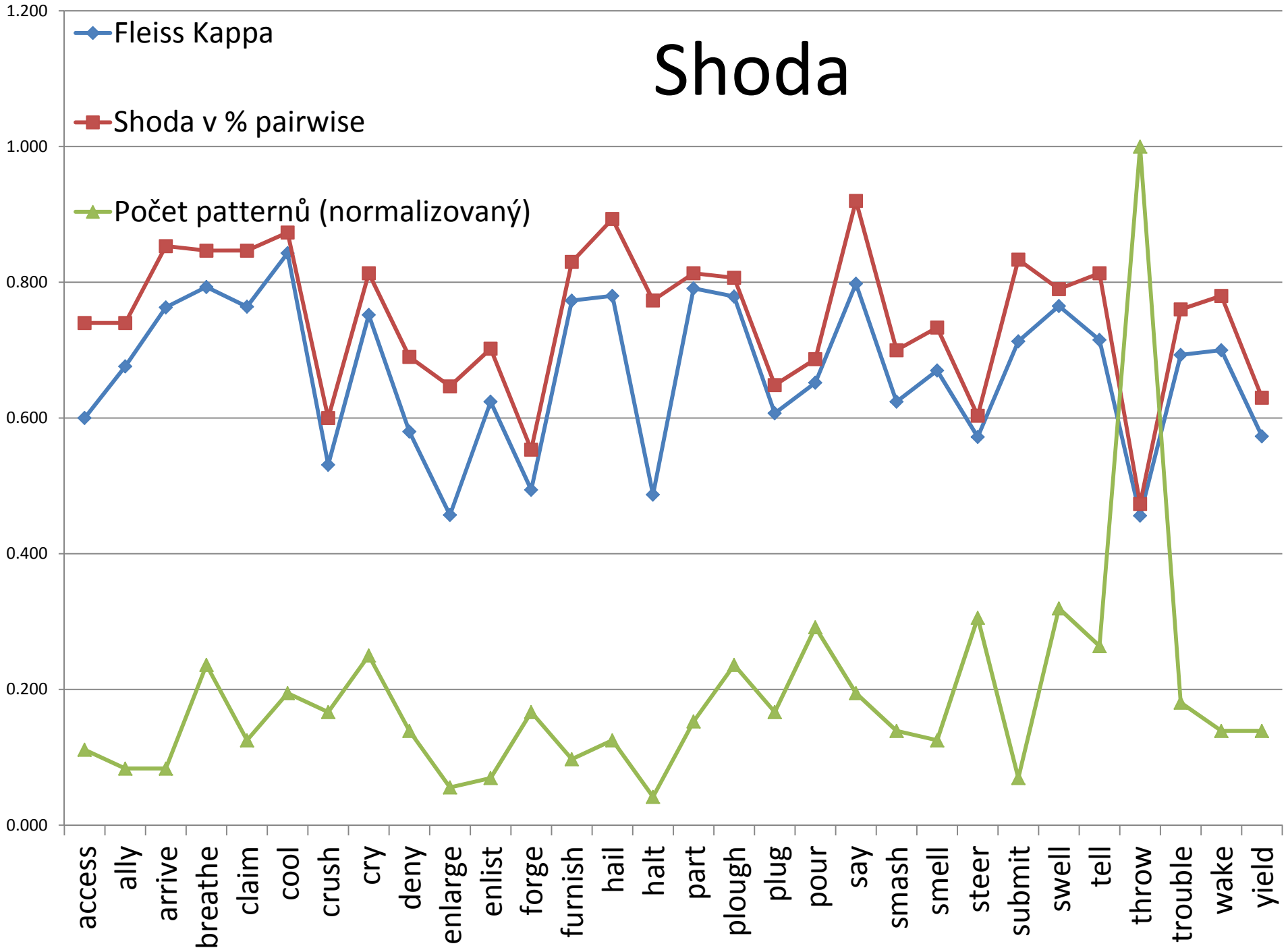
- **[Eventuality|Human | Institution] halt [Process| Activity| Attitude|Concept]**= [[Eventuality } Human | Institution]] causes [[Process | Activity|Attitude|Concept]] to stop
- **[[Eventuality]^[Human Group 1 = Military]] halt [[Human]^[Human Group 2 = Military]^[Vehicle]]**= [[Eventuality | Human Group]] causes [[Human | Human Group | Vehicle]] to stop moving forward

Sémantická modulace

Some of the scope for the criticism of policy that comes to Opposition members is <denied> to government supporters.

[Human 1 | Institution 1] deny [Human2 | Institution 2] {right|freedom|choice|permission|power |privilege|option|liberty|...} to+INF|of [Anything]|COMPOUND |ADJ|...] = [Human 1|Institution] refuses to recognize that [Human 2|Institution 2] has a right to decide whether and how to perform an [Action]

[Human 1 | Institution 1 | Eventuality 1 | Concept 1] deny [Human 2 | Institution 2] [Abstract| {access|entry|support|vote| help} = Service|Information|Right to do something] = [Human 1|Institution 1] refuses to do something for [Human 2|Institution 2] or allow [Human 2|Institution 2] to do something [Human 2 | Institution 2] expect to get it; regardless of who would perform the action (agent or patient)


Shoda
Fleiss Kappa
Shoda v % pairwise
Počet patternů (normalizovaný)
1.200
1.000
0.800
0.600
0.400
0.200
0.000
access
ally
arrive
breathe
claim
cool
crush
cry
deny
enlarge
enlist
forge
furnish
hail
halt
part
plough
plug
pour
say
smash
smell
steer
submit
swell
tell
throw
trouble
wake
yield

Adjudikace

#agrs	id	AV	EK	JT	SC	adjudication	sent_id	sentence
1	35	12	1.a	1	1.a	1.a	35741077	Thus one might say " What A <says> (i.e. the proposition A states)
2	17	1	6	1	6	1	15402751	When consonant clues are fewer , it takes longer to translate what is <said
3	7	11	11	11.s	11	11	5931864	Another was that , while Recommendations 7 and 8 had proposed higher pa
3	12	x	x	u	x	x	10948821	Since January 1990 all governing bodies of schools in Denmark have had co
3	22	1.a	1.a	14.a	1.a	1.a	18879937	` God helps those who help themselves ' , <says> the proverb .
3	39	1	1	1.c	1	1	44406328	A top source <said> in Dublin : ` It would make sense for the President to t
3	48	11	11	11.s	11	11	54453440	Suppose that some disturbance occurs -- <say> , a fall in investment dema
6	1	1	1	1	1		591696	Donald Selkin , head of stock index futures research at Prudential-Bache ,

Výsledky pilotní anotace

- 30 sloves VD-30-En stylem „1 tag na konkordanci“
- anotační manuál
- specifikace anotačního schématu
- Fl. kappa
  - u více než poloviny sloves nad 0,7
  - 8 sloves pod 0,6
- identifikace lingv. jevů, které systematicky znemožňují dosažení shody

Využití pilotní anotace

- Statistické rozpoznávání patternů (Vincent Kríž)
- Odhad informačního zisku z každého tagu (Vincent Kríž)
- Integrace patternů do distribučně sémantického modelu selekčních preferencí (Martin Holub)