Klasifikátor pro sémantické vzory užívání anglických sloves

Bc. Vincent Kríž
ÚFAL, MFF UK | vincent.kriz@kamadu.eu

Obsah

- **Úvod**
- **Definovanie úlohy, popis vstupných údajov**
  - Klasifikácia patternov ako úloha strojového učenia
  - Skúmanie vstupných údajov
- **Metódy a experimenty**
  - Návrh rysov pre strojové učenie
  - Príprava údajov
  - Popis experimentov
- **Záver**

Obsah

Corpus Pattern Analysis

- John Sinclair o WSD:
  - v autentickom použití jazyka je väčšina ambiguít riešiteľná pomocou kontextu
  - lexikálna jednotka by mala byť popísaná syntagmaticky i paradigmaticky
- Patrick Hanks: Corpus Pattern Analysis (**CPA**)
  - poloformálny lexikálny popis, ktorý konzistentne implementuje Sinclairov koncept zachytávania významu vo vzoroch použitia v jazyku namiesto lexikálnych jednotiek, ktoré sú používané v tradičnej lexikografii

Pattern Dictionary of English Verbs

- Pattern Dictionary of English Verbs (PDEV)
  - zachycuje ”normálne” použitia daného slovesa zotriedené do vzorov (**patternov**)
  - pattern pozostáva z lematizovaného slovesa a relevantných **kolokácií**
  - kolokácie sú klasifikovaná pomocou **sémantických typov**, **sémantických rolí** a **lexikálnych množín**
  - každá propozícia je **parafrázovaná** pomocou vety, v ktorej sú označené všetky relevantné kolokácie
  - parafráza predstavuje **implikatúru** alebo **významový potenciál** aktivovaný príslušným patternom

Pattern Dictionary of English Verbs

- Príklad:

2 **[[Human 1 = Student | Disciple]] follow [[Human 2 = Teacher]]**

[[Human 1 = Student | Disciple]] studies and is influenced by or tries to practice the teachings of [[Human 2 = Teacher]]

3 **[[Human | Institution]] follow [[Command | Document | Plan]]**

[[Human | Institution]] acts in accordance with [[Command | Plan]] (expressed in [[Document]])

4 **[[Event 1]] follow ([[Event 2]])**

[[Event 1]] happens after and typically as a consequence of [[Event 2]]

9 **[[Artifact | Proposition]] answer {need | purpose}**

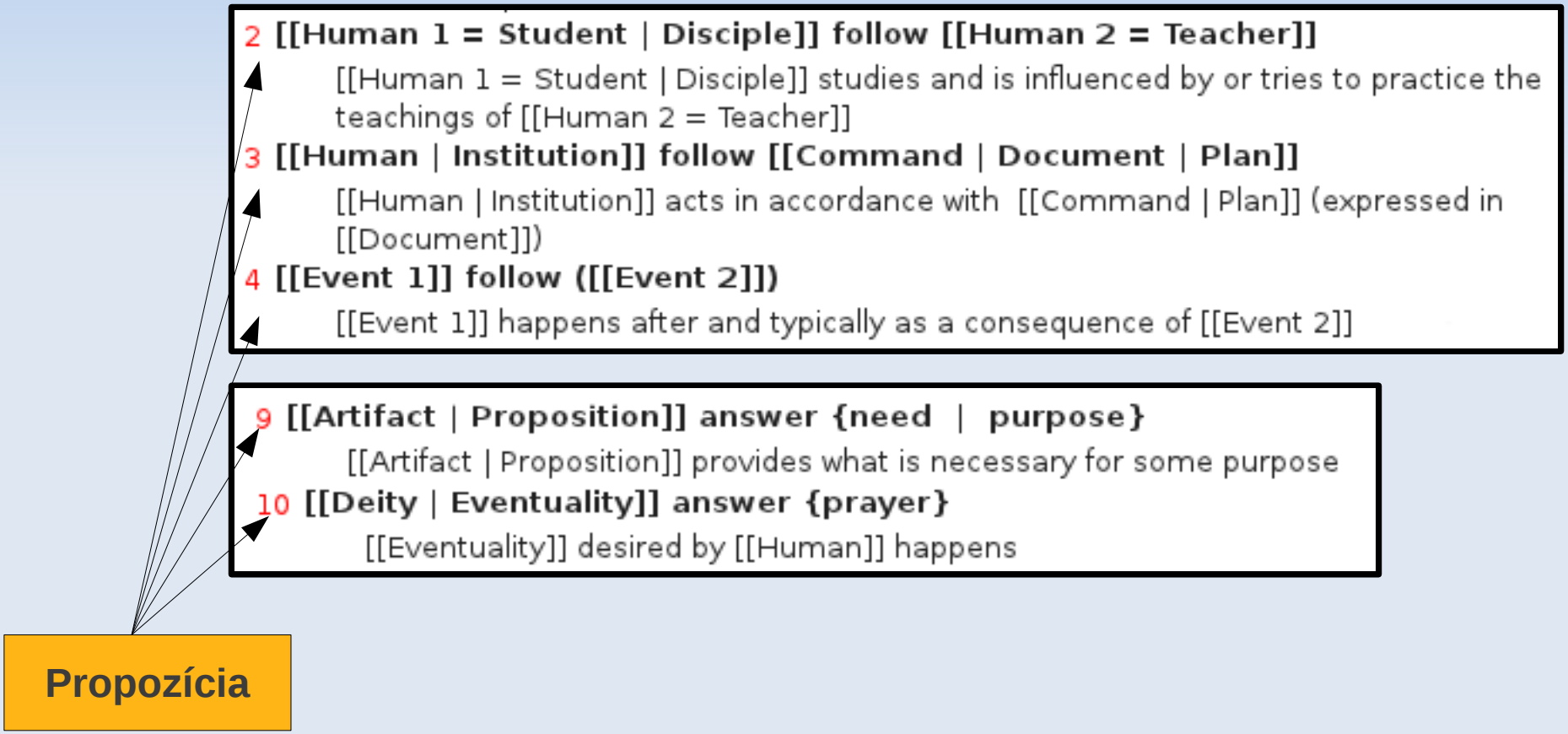
[[Artifact | Proposition]] provides what is necessary for some purpose

10 **[[Deity | Eventuality]] answer {prayer}**

[[Eventuality]] desired by [[Human]] happens

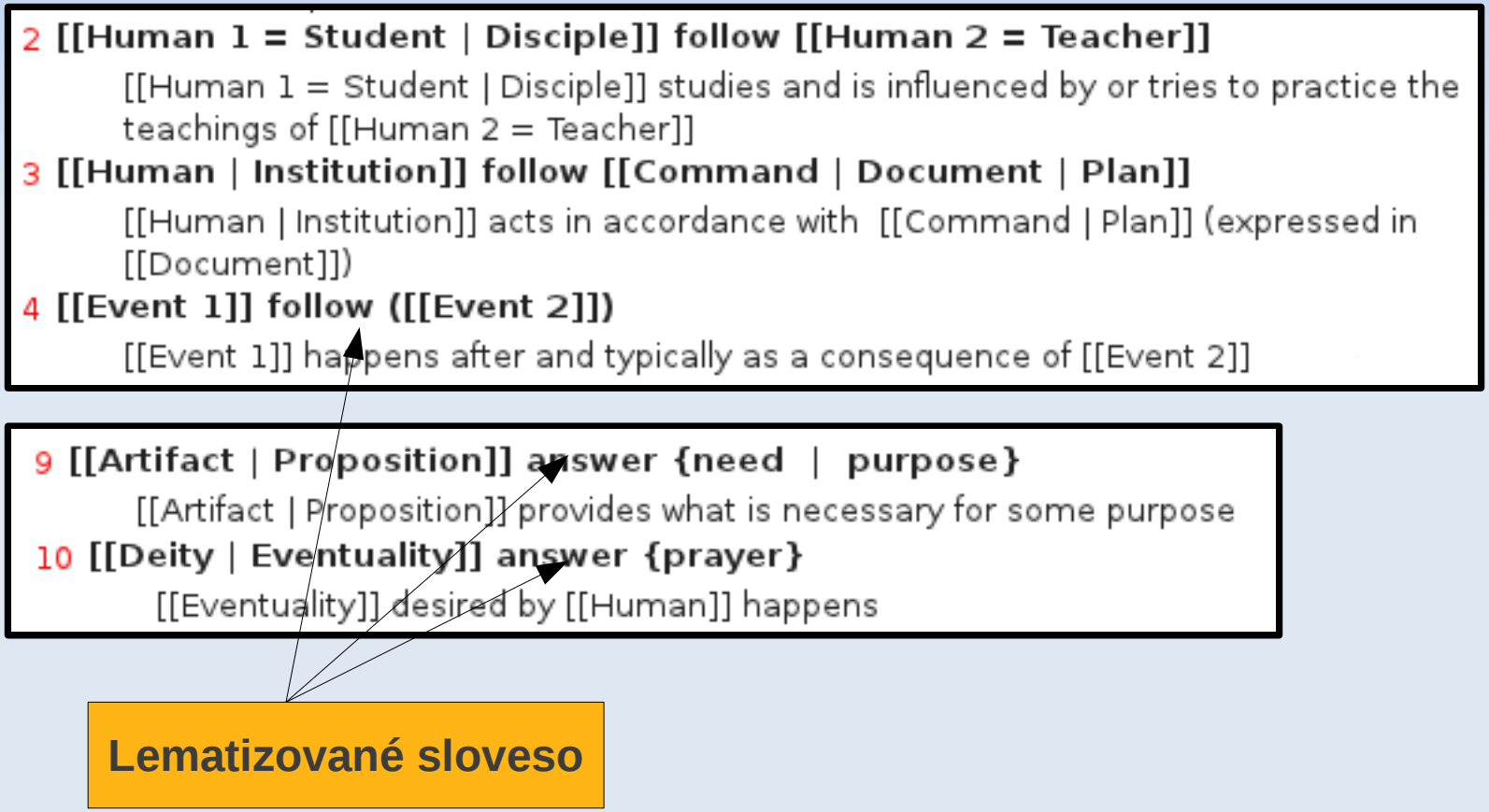
Pattern Dictionary of English Verbs

- Príklad:



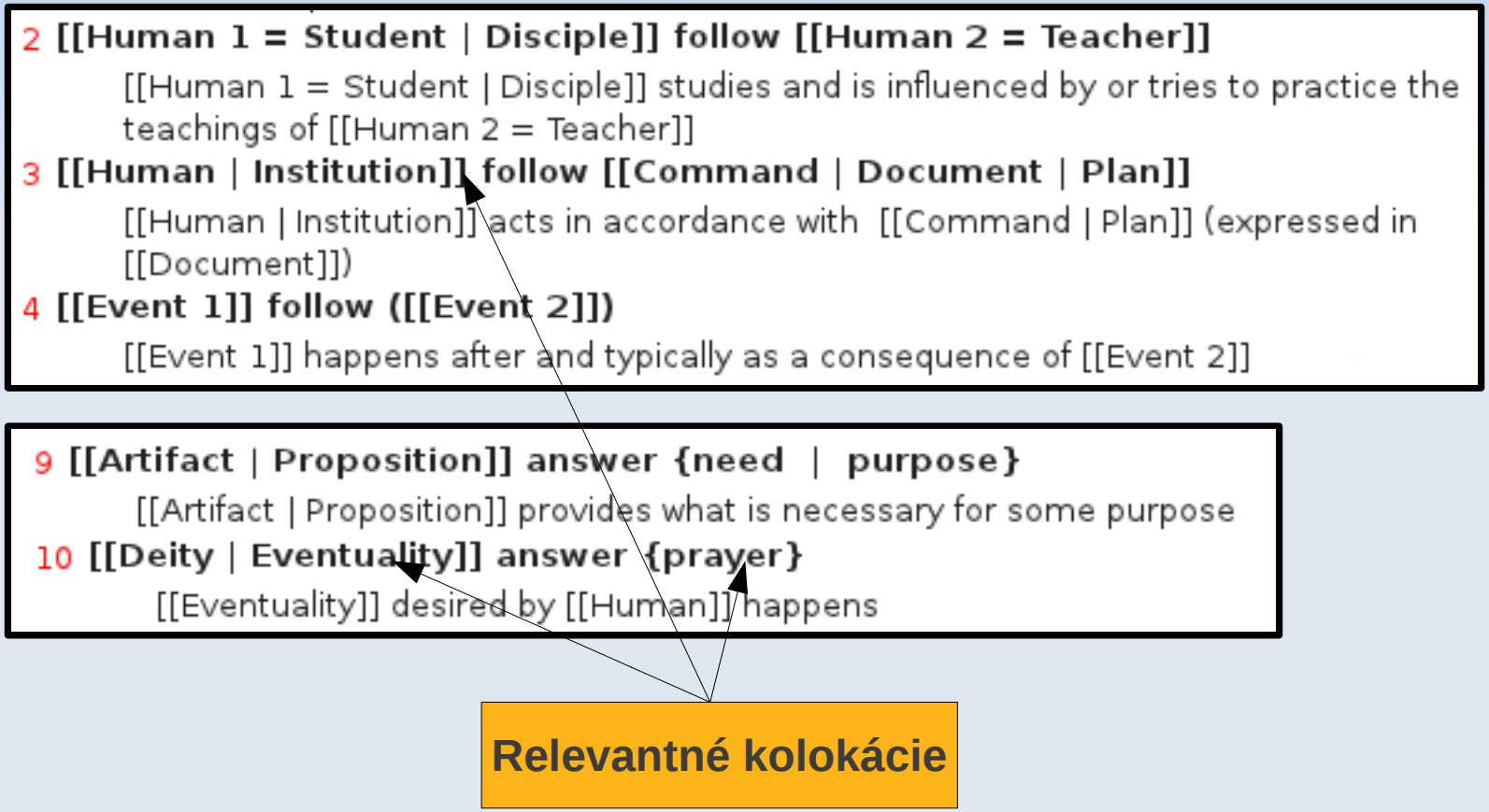
Pattern Dictionary of English Verbs

- Príklad:



Pattern Dictionary of English Verbs

- Príklad:



Pattern Dictionary of English Verbs

- Príklad:

2 **[[Human 1 = Student | Disciple]] follow [[Human 2 = Teacher]]**

[[Human 1 = Student | Disciple]] studies and is influenced by or tries to practice the teachings of [[Human 2 = Teacher]]

3 **[[Human | Institution]] follow [[Command | Document | Plan]]**

[[Human | Institution]] acts in accordance with [[Command | Plan]] (expressed in [[Document]])

4 **[[Event 1]] follow ([[Event 2]])**

[[Event 1]] happens after and typically as a consequence of [[Event 2]]

9 **[[Artifact | Proposition]] answer {need | purpose}**

[[Artifact | Proposition]] provides what is necessary for some purpose

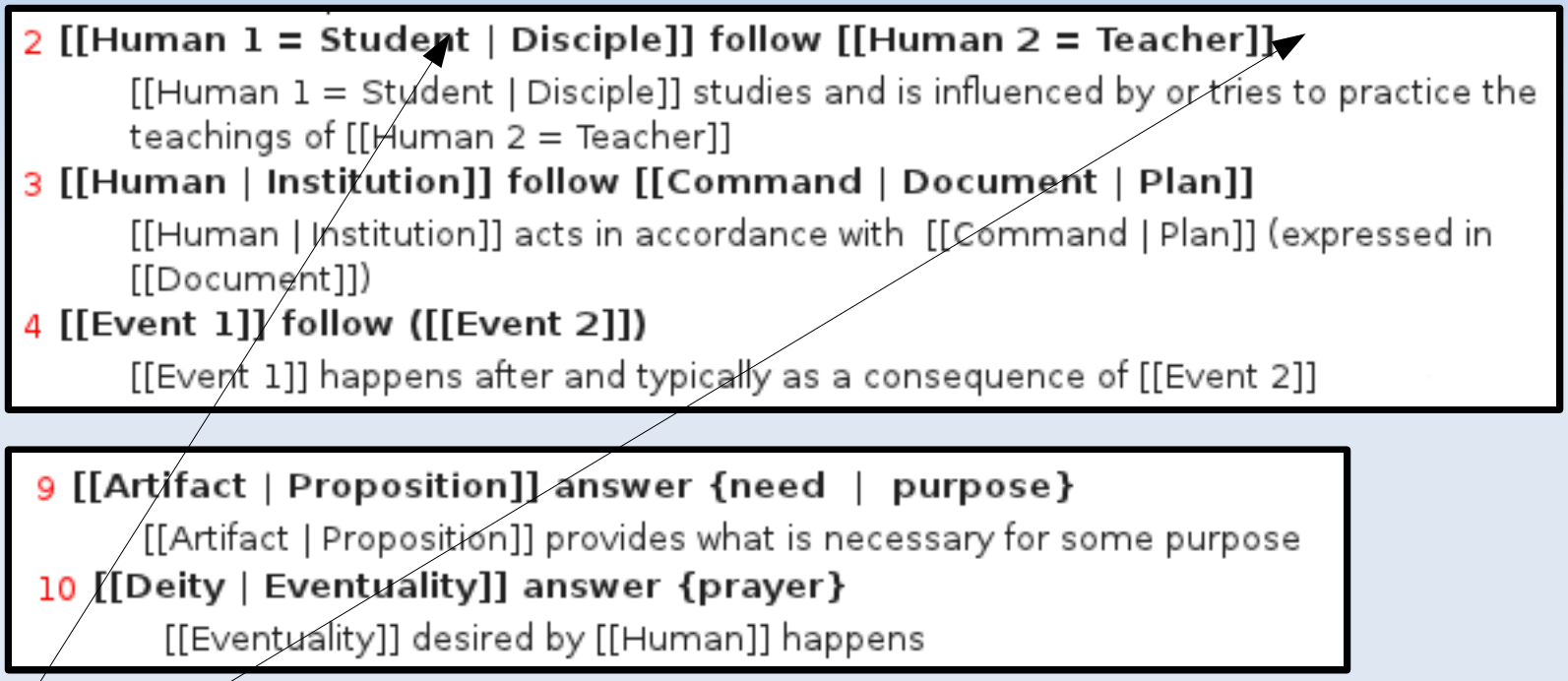
10 **[[Deity | Eventuality]] answer {prayer}**

[[Eventuality]] desired by [[Human]] happens

Parafráza

Pattern Dictionary of English Verbs

- Príklad:



Sémantická rola

Pattern Dictionary of English Verbs

- Príklad:

2 **[[Human 1 = Student | Disciple]] follow [[Human 2 = Teacher]]**

[[Human 1 = Student | Disciple]] studies and is influenced by or tries to practice the teachings of [[Human 2 = Teacher]]

3 **[[Human | Institution]] follow [[Command | Document | Plan]]**

[[Human | Institution]] acts in accordance with [[Command | Plan]] (expressed in [[Document]])

4 **[[Event 1]] follow ([[Event 2]])**

[[Event 1]] happens after and typically as a consequence of [[Event 2]]

9 **[[Artifact | Proposition]] answer {need | purpose}**

[[Artifact | Proposition]] provides what is necessary for some purpose

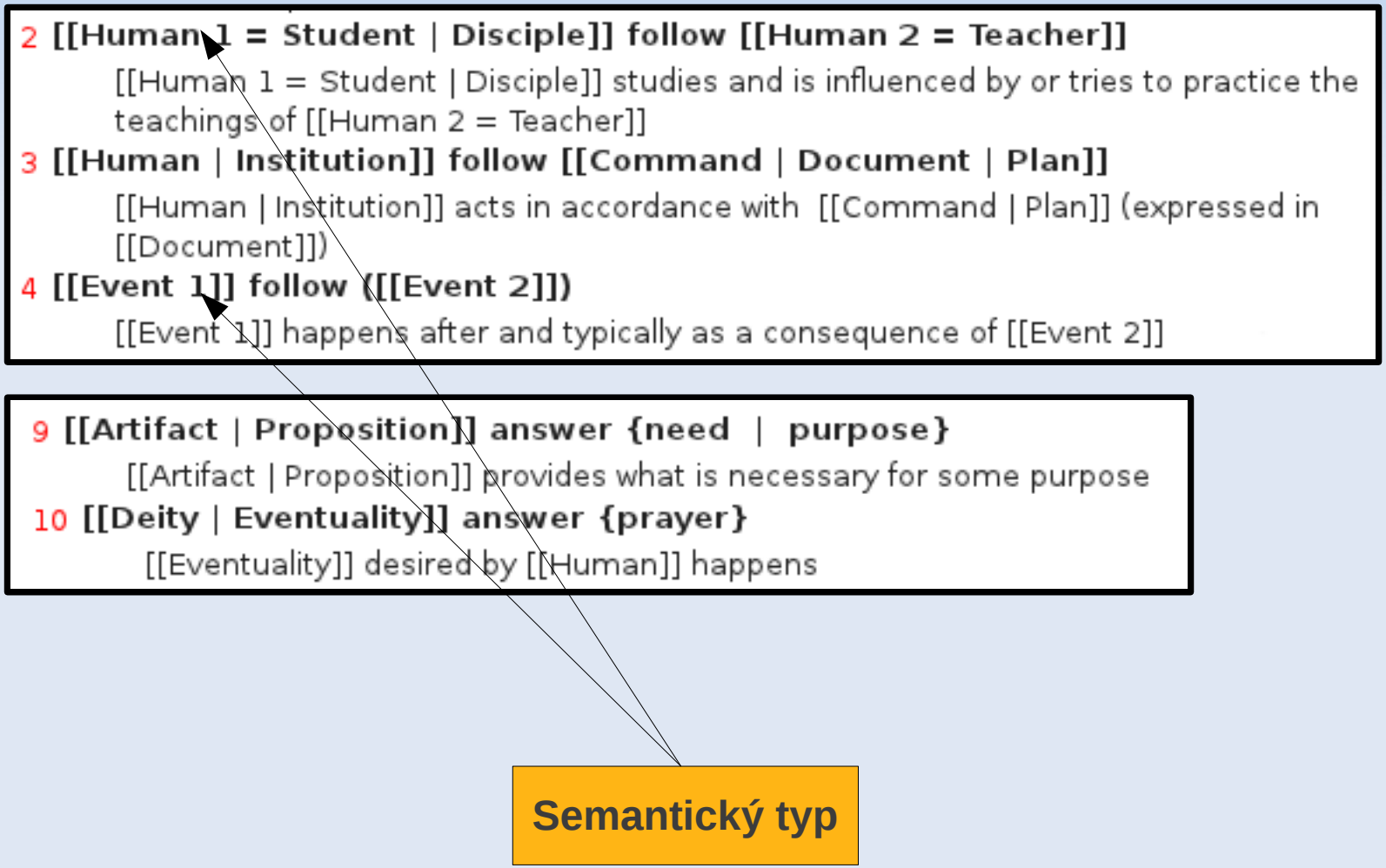
10 **[[Deity | Eventuality]] answer {prayer}**

[[Eventuality]] desired by [[Human]] happens

Lexikálna množina

Pattern Dictionary of English Verbs

- Príklad:



Sémantické značky

- **Normálne použitia** and **exploatácie**
- Anomálny argument (.a)
  - [[Human | Vehicle | Animal]] arrive [NO OBJ] {at [[Location]]}
  - The plot had arrived at Beirut.
- Coercion (.c)
  - [[Human]] drink [[Beverage]]
  - She drank 8 glasses of spirits
- Figuratívne použitie (.f)
  - [[Human | Institution]] arrive [NO OBJ] {at [[Concept = Considered Opinion]]}
  - Nobody arrives at ICI board level without some steel in his character.

Sémantické značky

- Anomálna syntax (.s)
  - [[Human 1 | Institution 1]] legally imposes a fine on, imprisons, or inflicts harm on [[Human 2 | Institution 2]] for [[Action]]
  - We punish too much—and in particular, we imprison too much
- Špeciálne značky **X** a **U**
- Chyby v taggovaní a ďalší šum (x)
  - **Ally** McCoist salutes Rangers' second goal
- Neklasifikovateľné (u)
  - Značka pre inštancie, ktorým nie je možné priradiť ani jeden z definovaných patternov

Obsah

- **Úvod**
- **Definovanie úlohy, popis vstupných údajov**
  - Klasifikácia patternov ako úloha strojového učenia
  - Skúmanie vstupných údajov
- **Metódy a experimenty**
  - Návrh rysov pre strojové učenie
  - Príprava údajov
  - Popis experimentov
- **Záver**

Zadanie úlohy

- **Motivácia**
  - CPA považujeme za užitočnú a pochopiteľnú metódu, doteraz ale chýba dôkaz, že CPA je vhodná aj na strojové spracovanie jazyka
- **Ciele práce**
  - analyzovať a najlepšie využiť údaje v PDEV
  - navrhnúť, implementovať a evaluovať klasifikátory pre rozpoznávanie patternov
  - rozpoznávanie lexikálnych jednotiek realizujúcich jednotlivé sémantické typy v PDEV
  - využitie automatického parsingu angličtiny a metód strojového učenia

Dostupné údaje

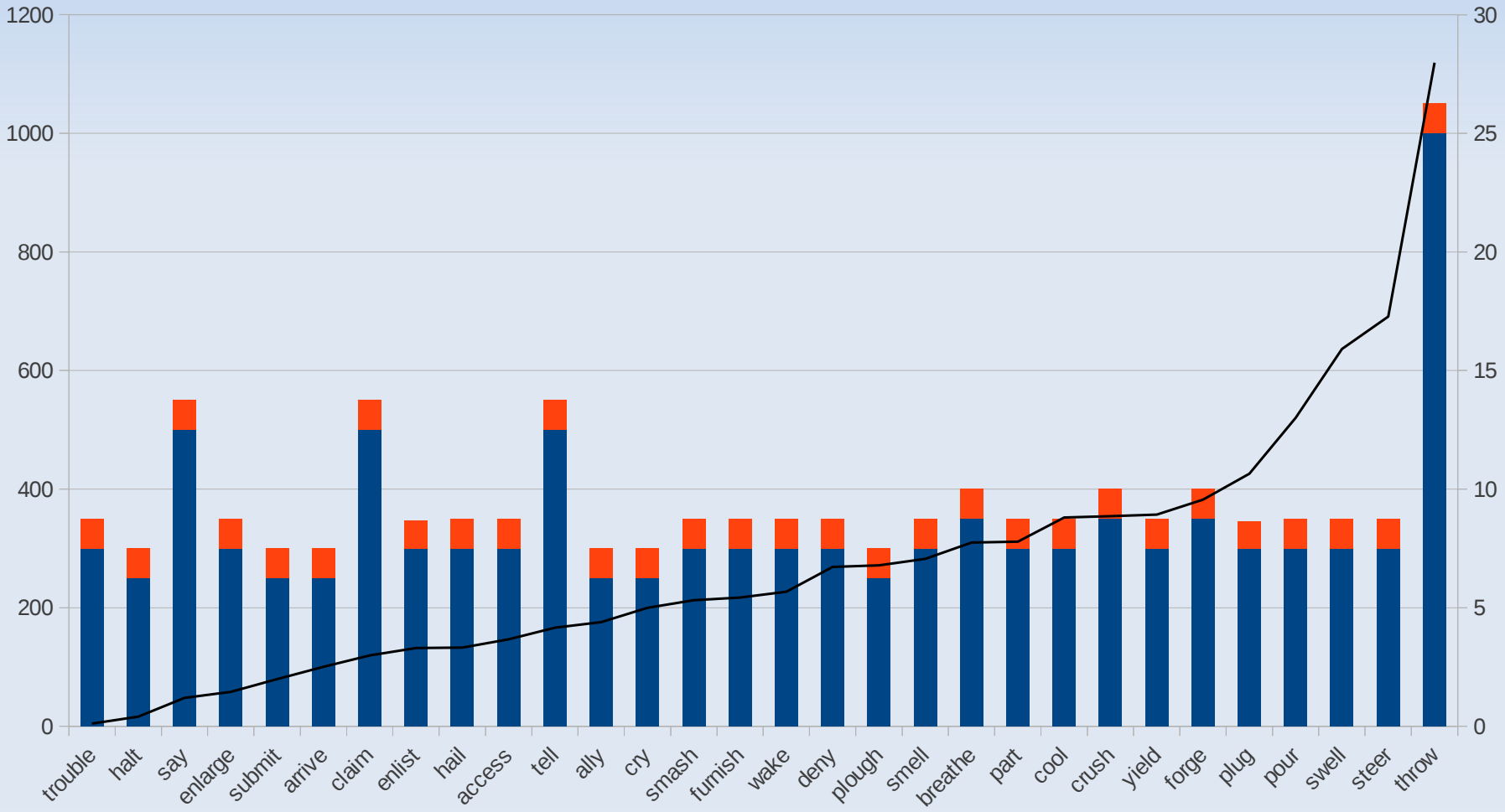
- **Kolekcia VPS-30-En**
  - 30 slovies vybraných zo slovníka PDEV
  - pilotná verzia nového lexikálneho zdroja
  - prečistené a multianotované konkordancie
  - snaha o čo najčistejšiu vzorku použiteľnú pri strojových experimentoch

Obsah

- **Úvod**
- **Definovanie úlohy, popis vstupných údajov**
  - Klasifikácia patternov ako úloha strojového učenia
  - Skúmanie vstupných údajov
- **Metódy a experimenty**
  - Návrh rysov pre strojové učenie
  - Príprava údajov
  - Popis experimentov
- **Záver**

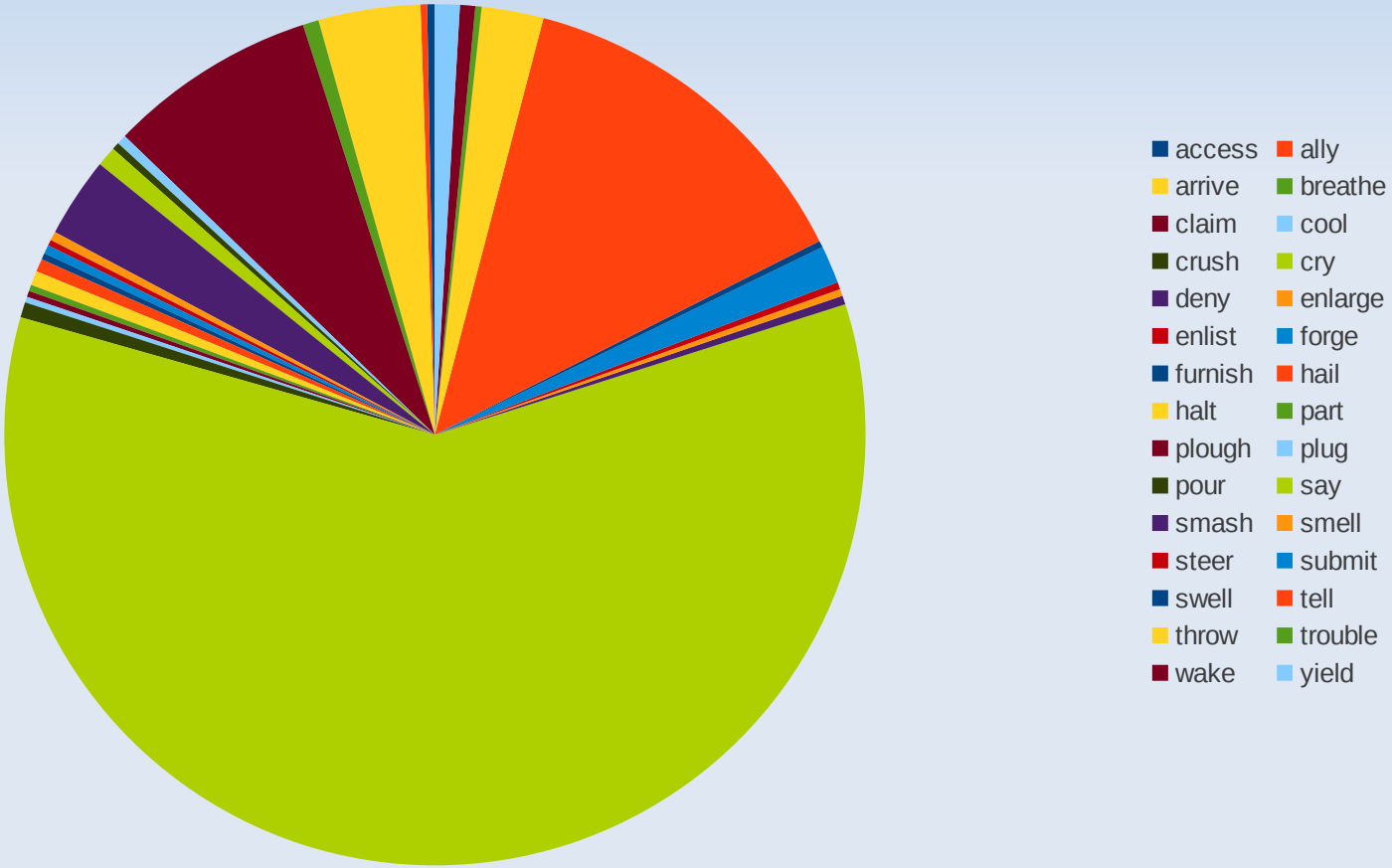
Čo vieme o vstupných údajoch

- **Počet dostupných údajov**



Čo vieme o vstupných údajoch

- **Pokrytie v BNC50**



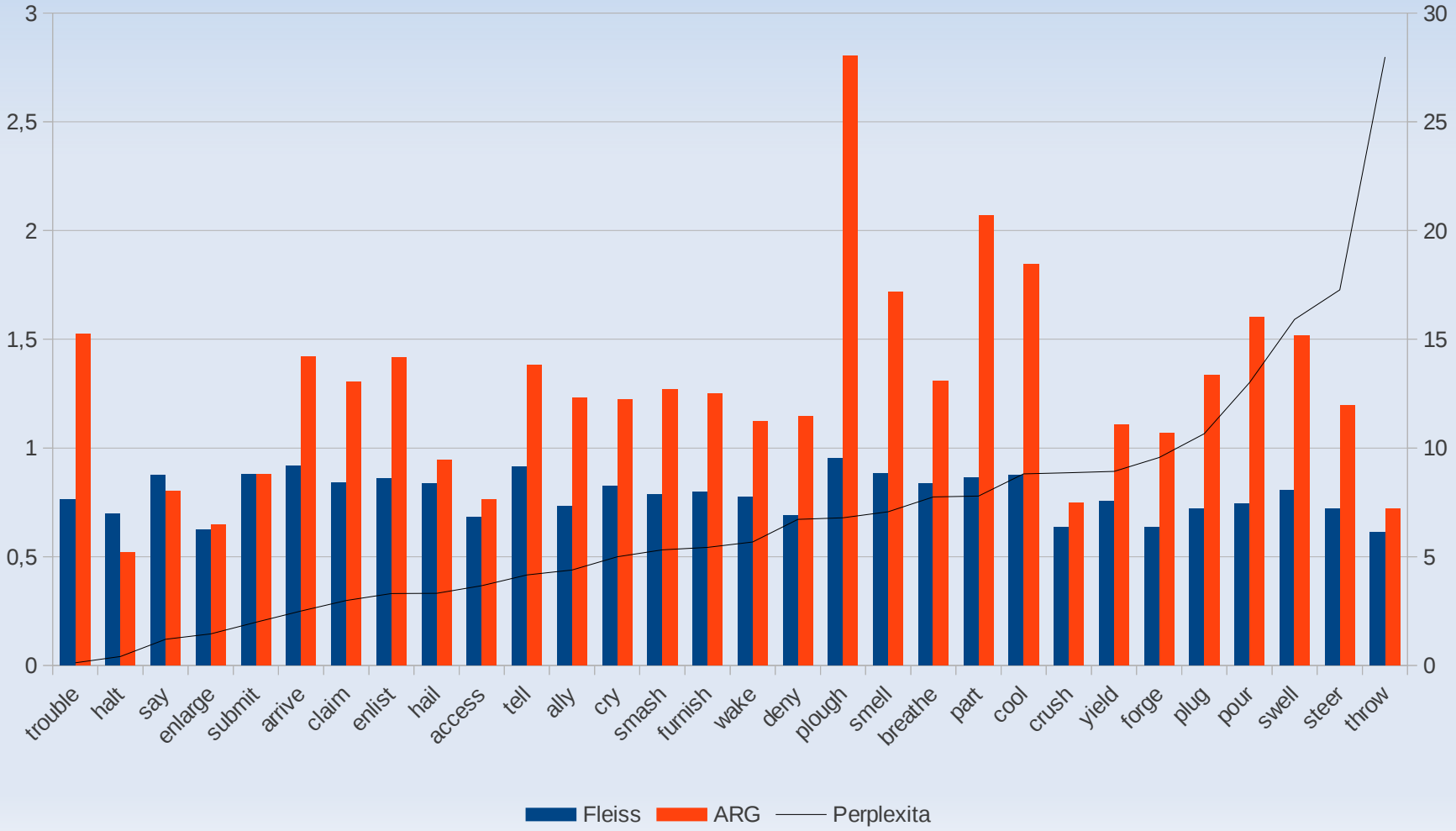
Čo vieme o vstupných údajoch

- **Rozdelenie slovies do skupín**

Skupina	Počet slovies	Počet konkordancií	Váha skupiny
A	3	128675	80.66%
B	6	19571	12.27%
C	21	11272	7.07%
Celkom	**30**	**159518**	**100.00%**

Čo vieme o vstupných údajoch

- **Medzianotátorská zhoda**



Obsah

- **Úvod**
- **Definovanie úlohy, popis vstupných údajov**
  - Klasifikácia patternov ako úloha strojového učenia
  - Skúmanie vstupných údajov
- **Metódy a experimenty**
  - Návrh rysov pre strojové učenie
  - Príprava údajov
  - Popis experimentov
- **Záver**

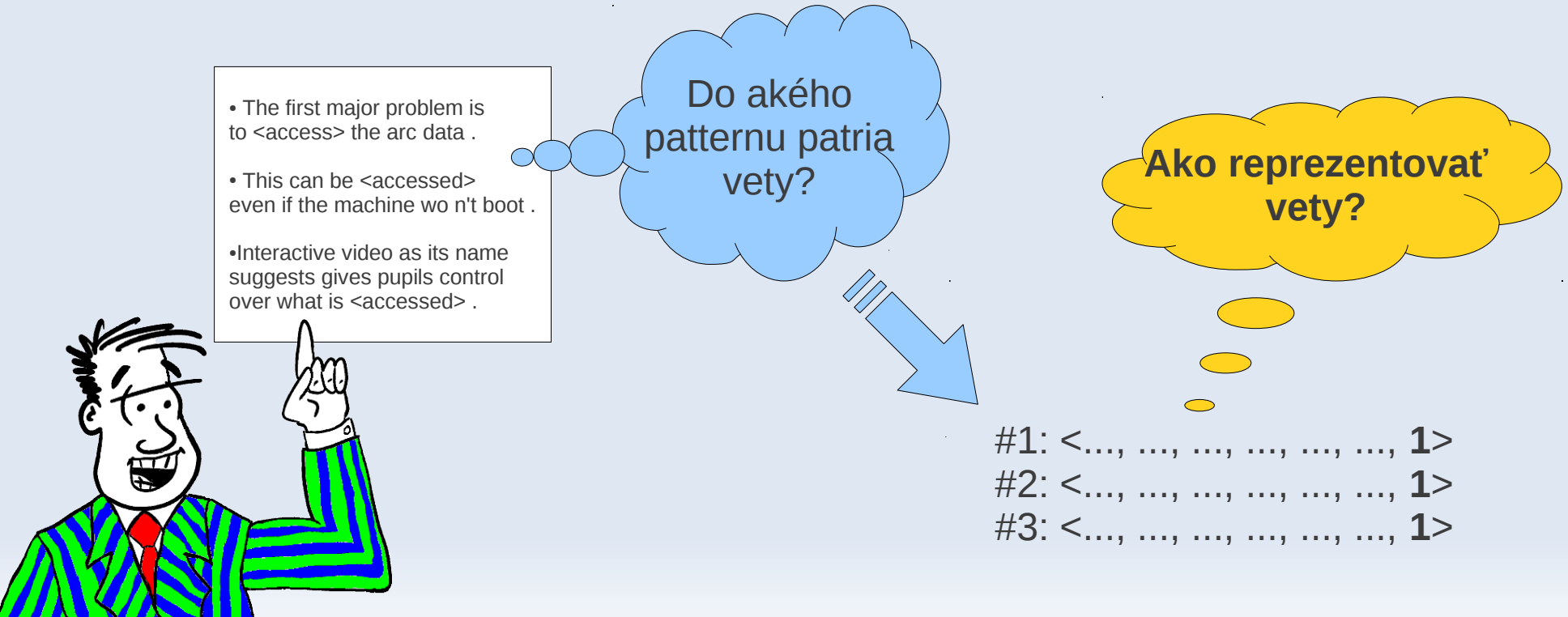
Strojové učenie

- **Reprezentácia objektov z reálneho sveta**
  - vytvoríme množinu rysov, ktorými budeme objekty charakterizovať
  - **inštancie** budeme reprezentovať **vektormi rysov**

Kráľovná morí: <1500t, 5kW, 100m, 1976>
Victoria II: <1600t, 4kW, 105m, 1876>
Enez Euza: <100t, 1kW, 50m, 2006>

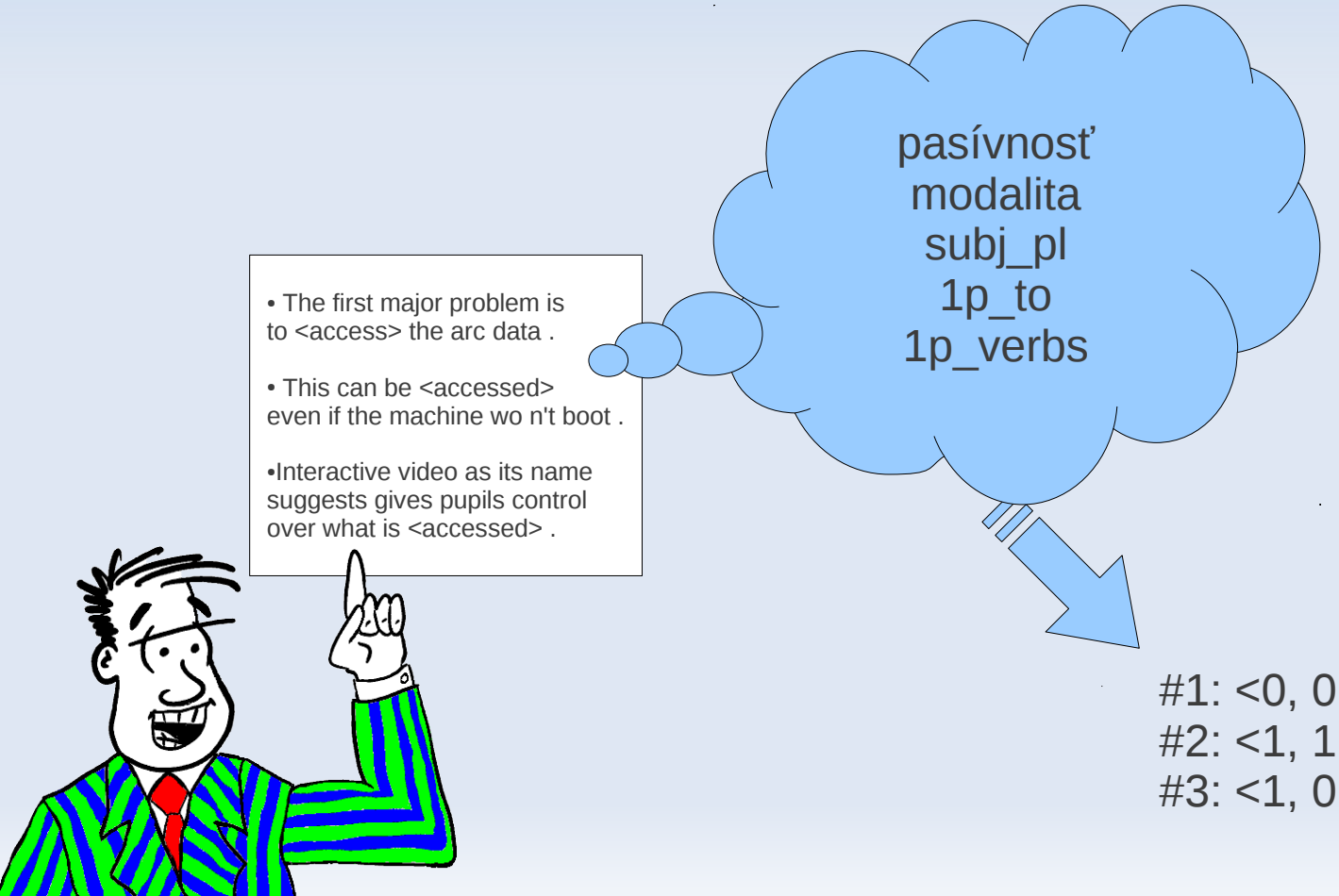
Strojové učenie

- **Reprezentácia objektov z reálneho sveta**
  - vytvoríme množinu rysov, ktorými budeme objekty charakterizovať
  - **inštancie** budeme reprezentovať **vektormi rysov**



Defaultná sada rysov

- **Morfologicko-syntaktické rysy**
  - binárne rysy budú charakterizovať sloveso, syntaktické členy a kontext



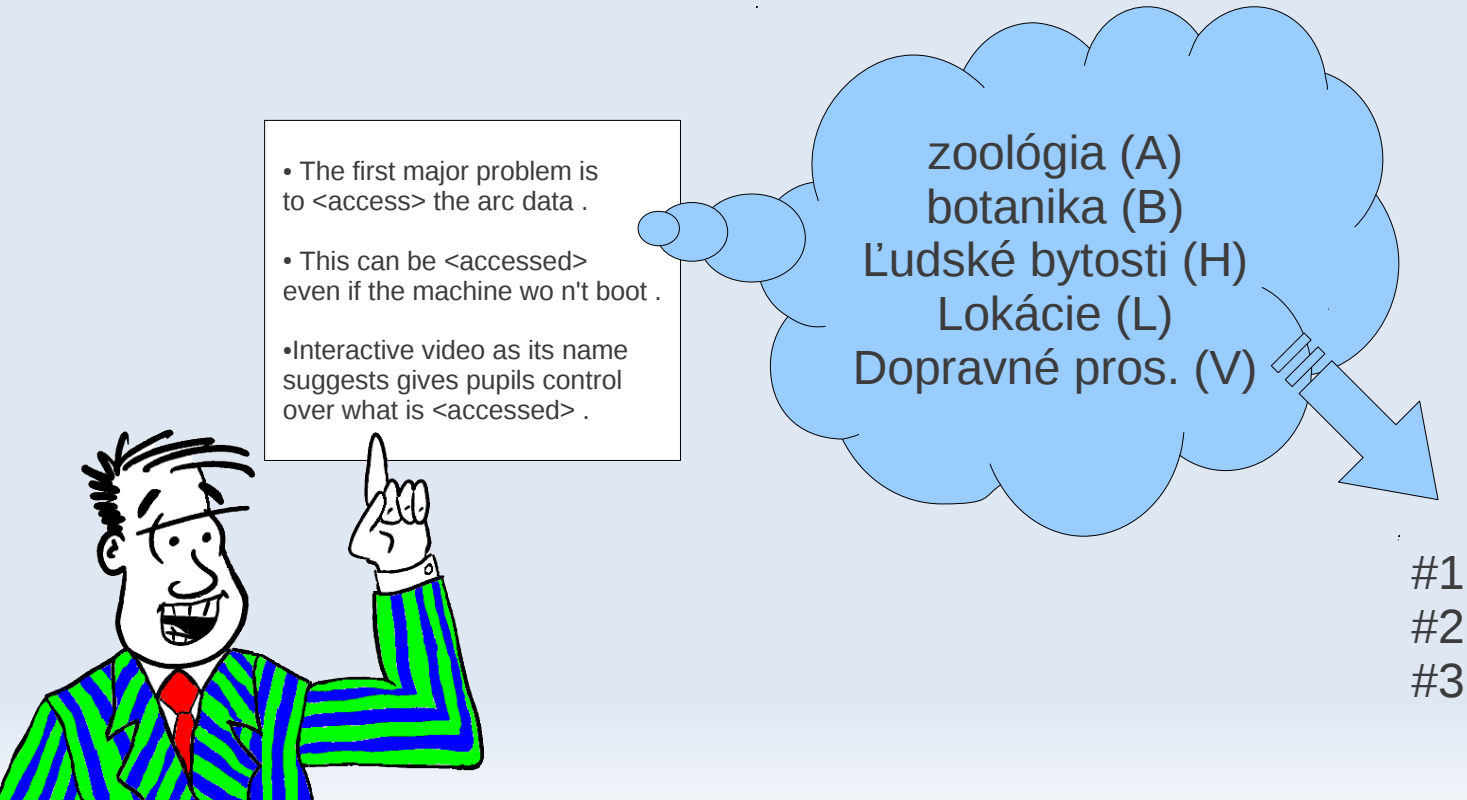
#1: <0, 0, 0, 1, 0, ..., **1**>
#2: <1, 1, 0, 0, 0, ..., **1**>
#3: <1, 0, 0, 0, 1, ..., **1**>

Defaultná sada rysov

- **Morfologicko-syntaktické rysy**
  - Charaketistika cieľového slovesa
    - napr. pasívnosť, modalita, negácia, čas
  - Charakteristika najbližšieho kontextu slovesa (± 3 slová)
    - zaradenie slova do jednej z definovaných skupín (napr. substantíva, adjektíva, slovesá, modálne slovesá, príslovky, ...)
  - Charakterizácia syntakticky závislých slov
    - existencia subjektu, objektu, adverbiálov

Defaultná sada rysov

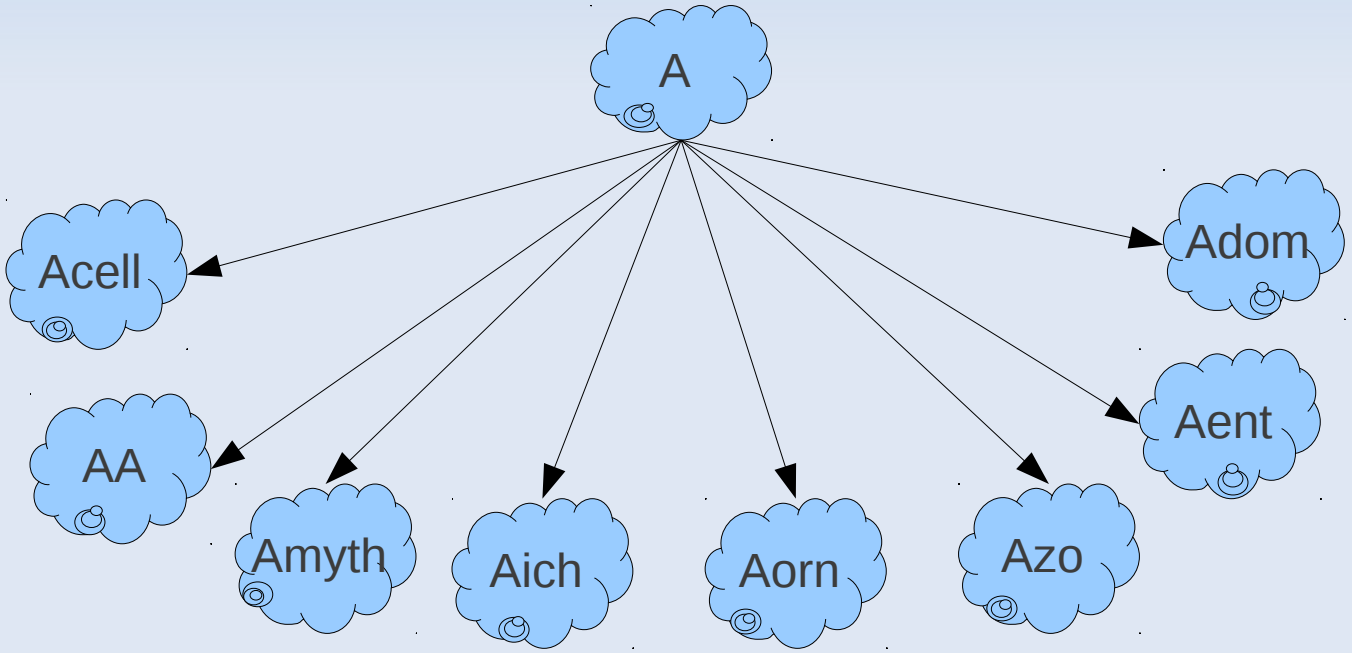
- **Sémantické rysy**
  - Lexikón sémantických prototypov (LSP)
  - E. Bick
  - prvé použitie LSP pre ML a WSD



#1: <0, 0, 0, 0, 0, ..., **1**>
#2: <0, 0, 0, 0, 0, ..., **1**>
#3: <0, 0, 0, 0, 0, ..., **1**>

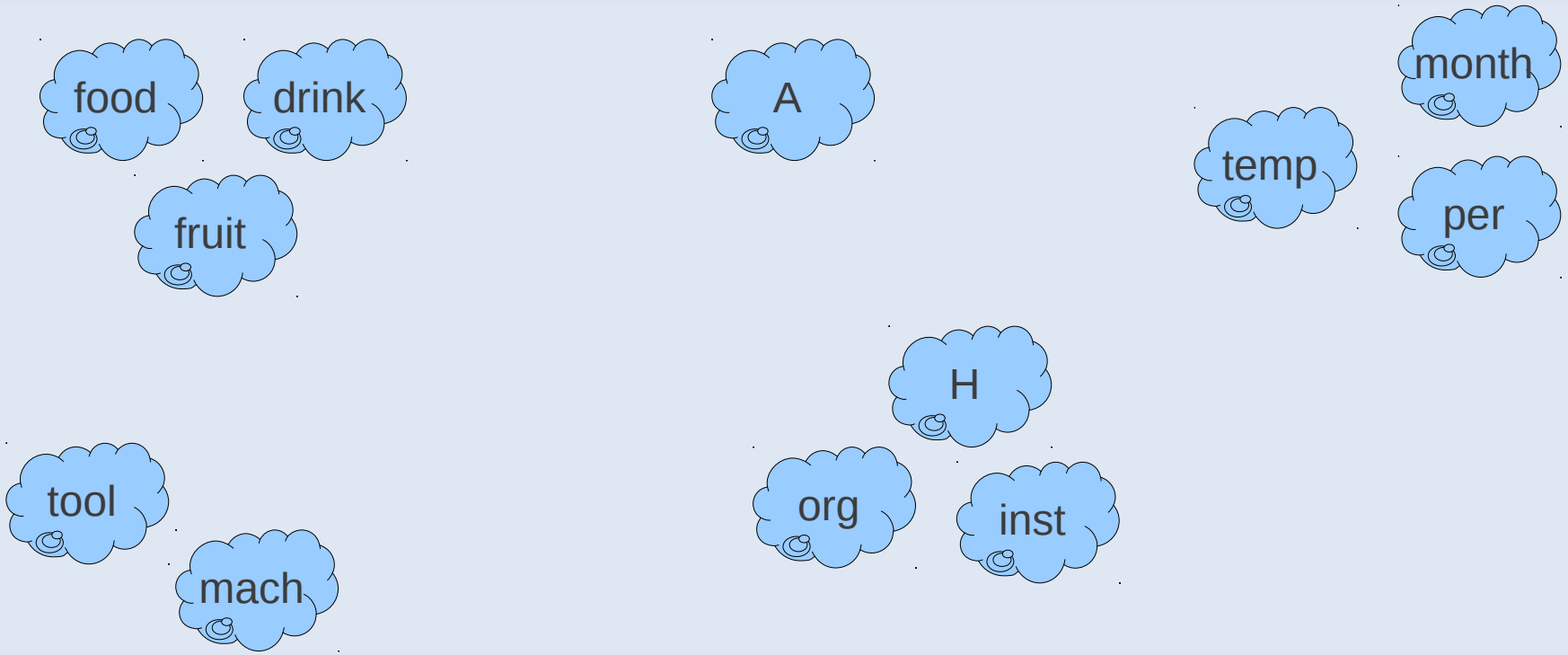
Defaultná sada rysov

- **Lexikón sémantických prototypov (LSP)**
  - hierarchický systém prototypov



Defaultná sada rysov

- **Lexikón sémantických prototypov (LSP)**
  - hierarchický systém prototypov



Defaultná sada rysov

- **Lexikón sémantických prototypov (LSP)**
  - 2 skupiny sémantických rysov
  - hlavné zastrešujúce prototypy
  - zastrešujúce prototypy

Zoológia	Celky a kolektívy
Botanika	Domény
Ľudia	Vlastnosti
Lokácie	Jedlo
Dopravné prostriedky	Pocity a vnímanie
Abstraktné koncepty	Produkty ľudskej kreat.
Akcie, udalosti	Okolnosti dejov
Anatómia	Čas
Konktrétne koncepty	Nástroje a stroje
Oblečenie	Jednotky
Materiály	Počasie

Obsah

- **Úvod**
- **Definovanie úlohy, popis vstupných údajov**
  - Klasifikácia patternov ako úloha strojového učenia
  - Skúmanie vstupných údajov
- **Metódy a experimenty**
  - Návrh rysov pre strojové učenie
  - Príprava údajov
  - Popis experimentov
- **Záver**

Výstupné triedy

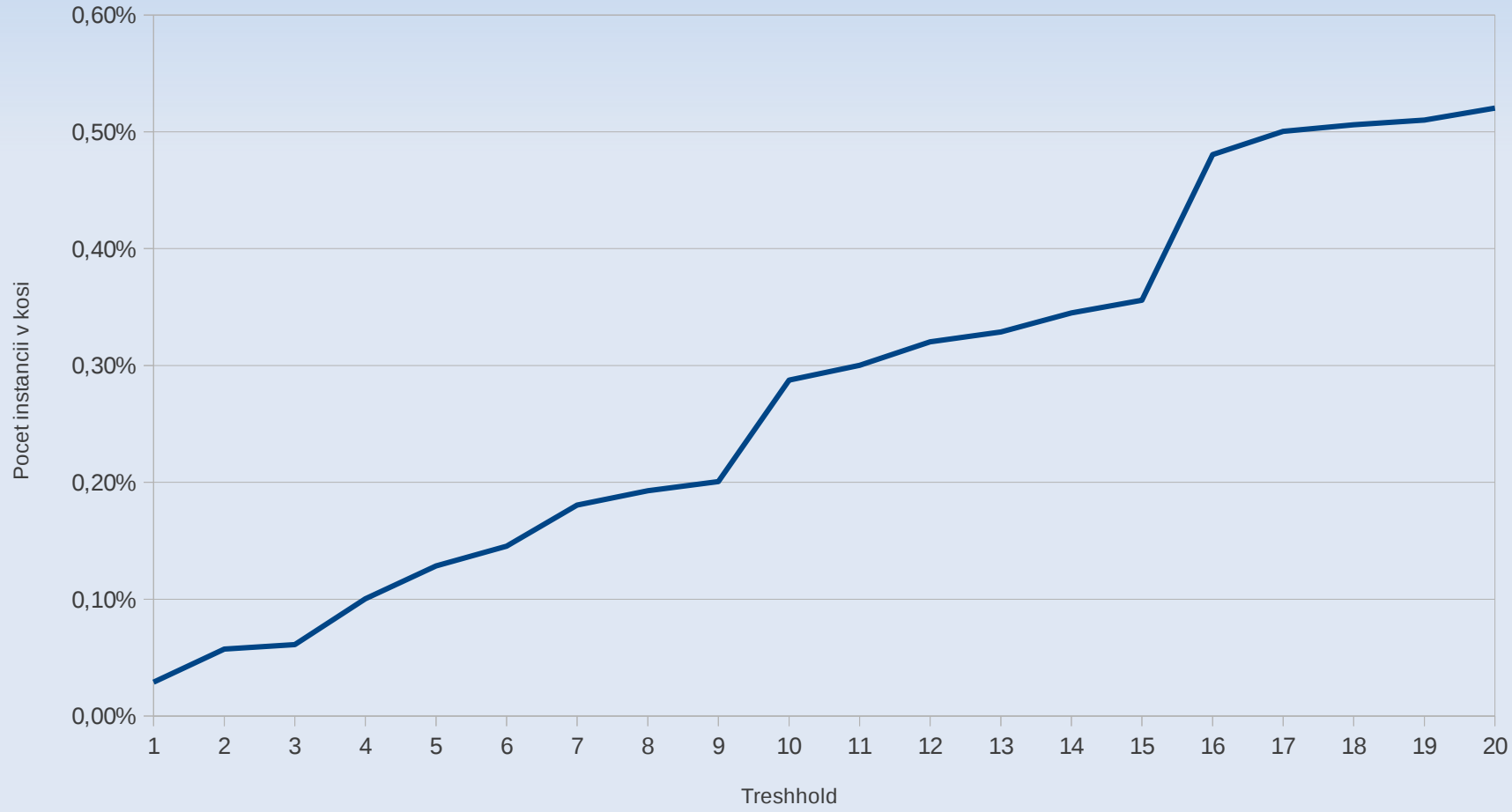
- Patterny by mali byť triedy, ale
  - n patternov = $\mathrm{n} + 4\mathrm{n} + 1 + 1$ tried
- Nemáme dostatok inštancií pre každú triedu
  - potrebujeme aspoň 5-10
- **Aké sú možnosti?**
  - zlúžiť exploatácie s normálnymi použitiami
  - odstrániť triedy s malými frekvenciami
  - zlúčiť podobné patterny
- Ako ošetriť špeciálne triedy X a U?

Defaultná sada rysov

- **Počet inštancií pre jednotlivé patterny**

Defaultná sada rysov

- **Prahová frekvencia patternov**



Defaultná sada rysov

- **Extrakcia kategoriálnych rysov**
- **Definovane cross-validácie**
  - 9-fold cross-validation
- **Rozdelenie údajov na trénovacie a testovacie**
  - 50 multianotovaných inštancií na test
  - zvyšok na trénovanie

Obsah

- **Úvod**
- **Definovanie úlohy, popis vstupných údajov**
  - Klasifikácia patternov ako úloha strojového učenia
  - Skúmanie vstupných údajov
- **Metódy a experimenty**
  - Návrh rysov pre strojové učenie
  - Príprava údajov
  - Popis experimentov
- **Záver**

Defaultná sada rysov

- **Naivný klasifikátor**

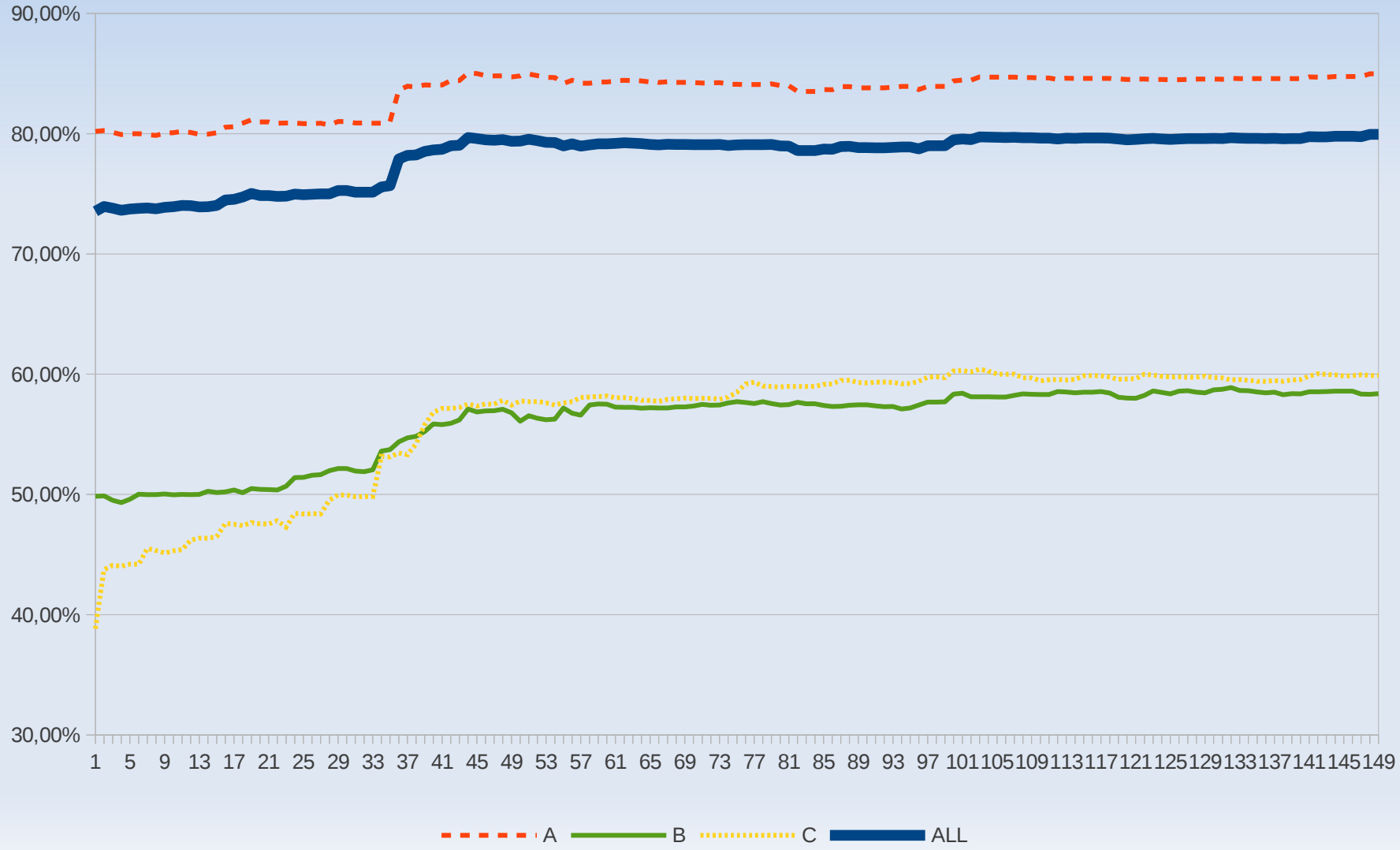
Skupina	Perplexita	Acc.
A	2.320	80.2 ±0.6
B	6.482	50.0 ±1.0
C	5.342	43.9 ±1.1
Celkom	3.044	73.9 ±0.7

Defaultná sada rysov

- **Optimalizácia morfo-syntaktických rysov**
  - rôzne zoradenia rysov
    - úspešnosti rysov pri klasifikovaní každého patternu zvlášť (rebríčky A, W)
    - výber rysov hladovým algoritmom (rebríček G)
  - hladový algoritmus pre každé sloveso

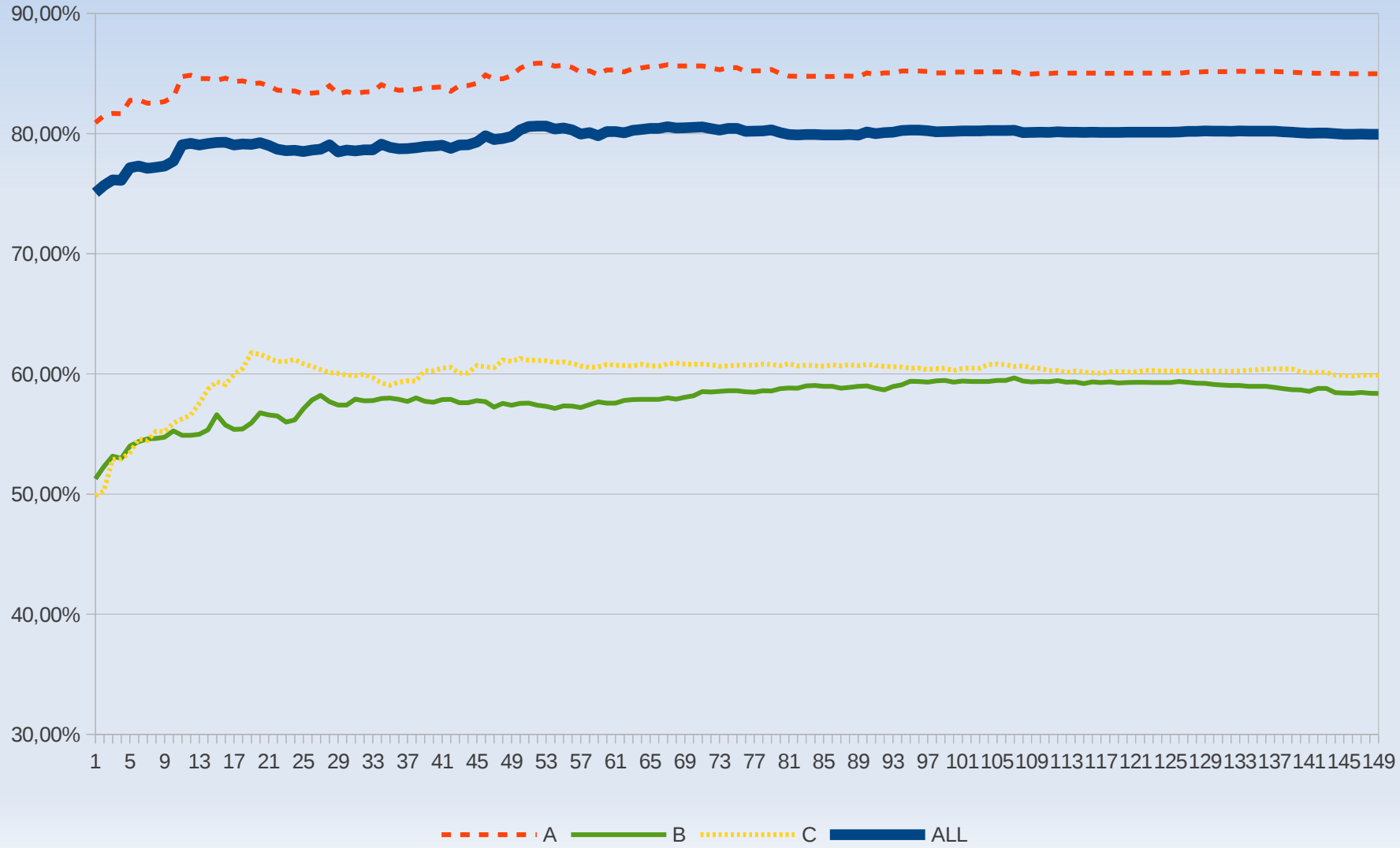
Defaultná sada rysov

- **Rebríček A**



Defaultná sada rysov

- **Rebríček G**

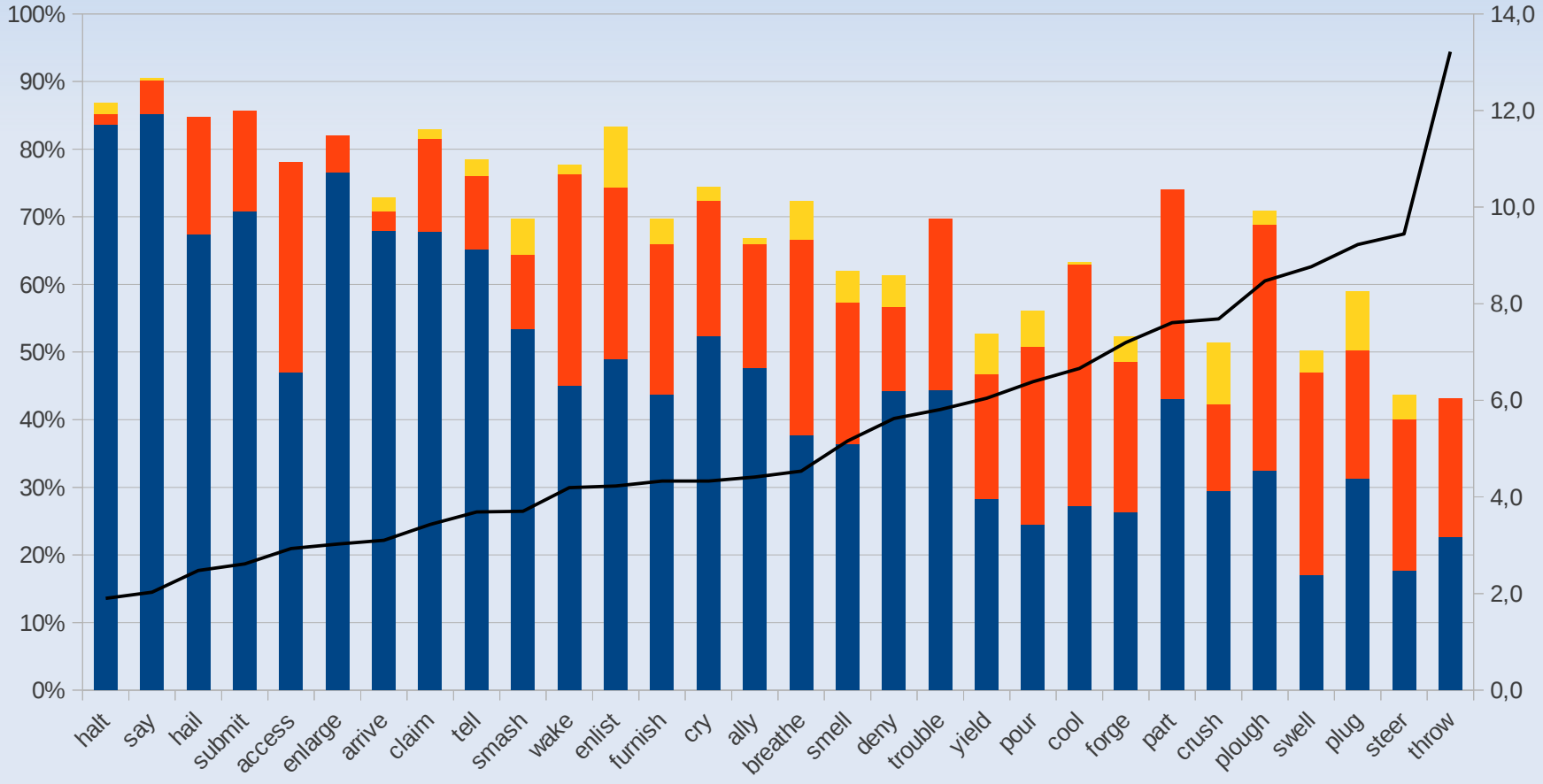


Defaultná sada rysov

- **optimalizácia sémantických rysov**
  - hladový algoritmus
  - výber z Best58 aj zo sémantických rysovň
    - modely BestMU, BestAU
  - zafixovaná Best58 a výber zo sém. rysov
    - model GreedyAU

Defaultná sada rysov

- **Porovnanie najlepších modelov**



Obsah

- **Úvod**
- **Definovanie úlohy, popis vstupných údajov**
  - Klasifikácia patternov ako úloha strojového učenia
  - Skúmanie vstupných údajov
- **Metódy a experimenty**
  - Návrh rysov pre strojové učenie
  - Príprava údajov
  - Popis experimentov
- **Záver**

Záver

- **návrh vhodnej množiny rysov pre ML**
- **určenie 2 množín ako najvhodnejších**
  - čiste morfo-syntaktické rysy
  - morfo-syntaktické a sémantické rysy
- **hlavný limit výkonnosti – nedostatok údajov**
- **experimenty so sémantickými rysmi s rovnakými výsledkami ako skôr publikované štúdie**

Záver

- morfo-syntaktické rysy – najdôležitejšie pre sémantickú desambiguáciu pomocou ML
- pre *niektoré* slovesá hrajú sémantické rysy dôležitú úlohu

Publikácie

- **Cinková S., Holub M., Kríž V. Managing Uncertainty in Semantic Tagging. EACL 2012, Avignon, France.**
  - nová miera ARG pre medzianotátorskú zhodu
- **Cinková S., Kríž V., Holub, M. Optimizing semantic granularity for NLP – report on a lexicographic experiment. To appear. Accepted for the 15th Euralex International Congress, Oslo, Norway, 2012.**
  - použitie ARG pri optimalizácii tagsetov
- **Holub M., Kríž V., Cinková S., Bick E. Automatic Classification of Verb Semantic Patterns. Submitted to the First Joint Conference on Lexical and Computational Semantics (*SEM), 2012.**
  - priamo výsledky experimentov popísaných v DP