

# Automatická analýza chyb v české deklinaci

Ilustrace využití NLP v aplikacích pro výuku cizích jazyků

**Ivan Šmilauer**

INALCO - Paris 4 Sorbonne, STIH-LALIC



# Využití automatického zpracování jazyka pro CALL

**CALL** (Computer Assisted Language Learning)

specifická subdoména oboru e-learning



**Automatické zpracování jazyka**

1

## Zpracování cílového jazyka

Tvorba a obohacování pedagogických materiálů:

- anotace textů, vizualizace gram. informací
- výběr textů v závislosti na formě a obsahu
- text-to-speech, etc.

2

## Analýza jazykových produkcí studentů

Identifikace, oprava a diagnostika chyb

- feedback
- learner-model:
  - výběr aktivit
  - adaptace feedbacku

# Využití automatického zpracování jazyka pro CALL

## ICALL : Intelligent Computer-Assisted Language Learning

*Nerbonne (2003), Heift & Schulze (2007), Meurers (2009), Meurers (2012)*

- původ v Intelligent Tutors v rámci **AI** v šedesátých letech
- opatrný **rozvoj** v posledních zhruba 20 letech
- **pluridisciplinarita** (lingvistika, NLP, osvojování cizího jazyka, pedagogika, didaktika)
- řada víceméně funkčních aplikací
- minimum komerčních aplikací (náročný vývoj a nespolehlivost)
- techniky jdoucí od jednoduchého **pattern matching** omezených textů až po **parsing** volných produkcí

# Problémy ICALL

- Komplikace s spojené s analýzou **chybných textů** :
  - \* **pravopisné chyby a morfologické chyby** (\**matke*, \**hůle*, \**pesa*)
  - \* **syntaktické chyby** (shoda, rekce, předložky)
  - \* **slovosled**
  - \* **lexikální a sémantické chyby**
- adaptace parserů: **mal-rules**, **constraint relaxation** - nepřiliš uspokojivé výsledky



**Nedokonalost** automatického zpracování může být tolerováno rodilým mluvčím, ale na studenta cizího jazyka má vyloženě **negativní vliv**.

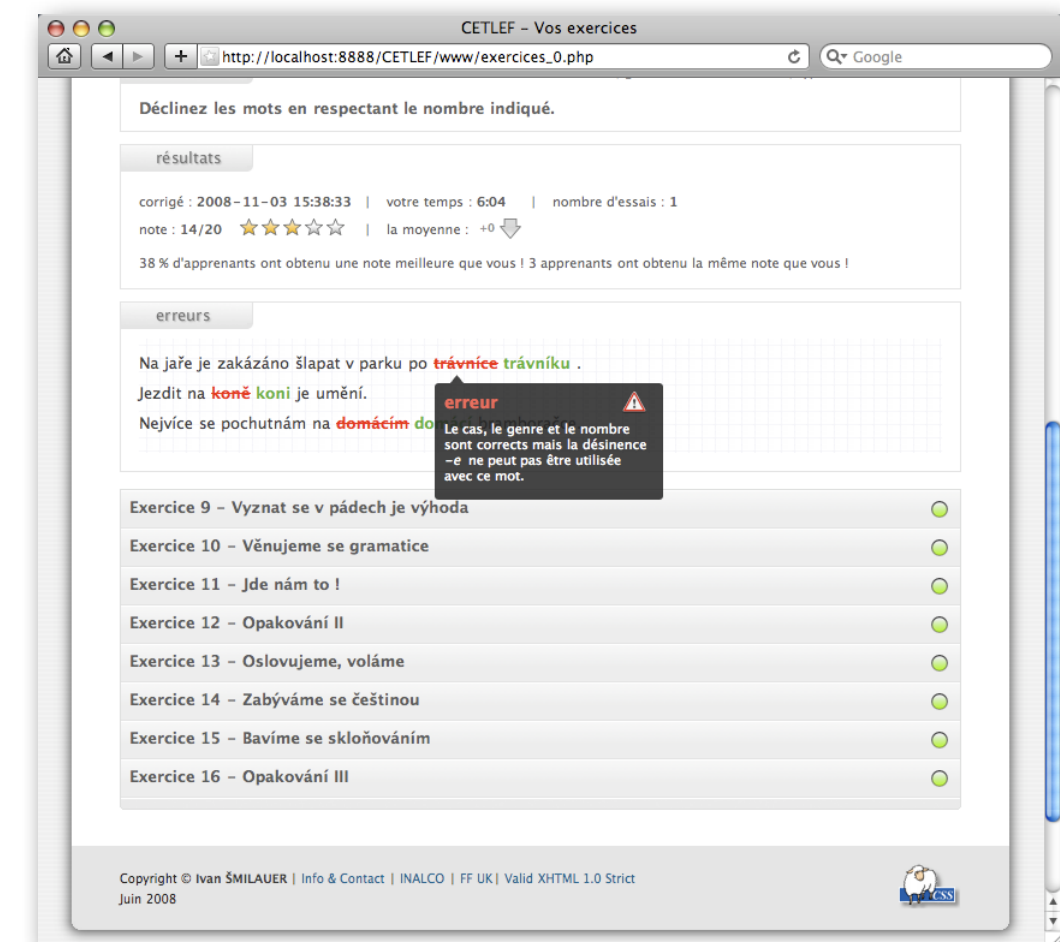
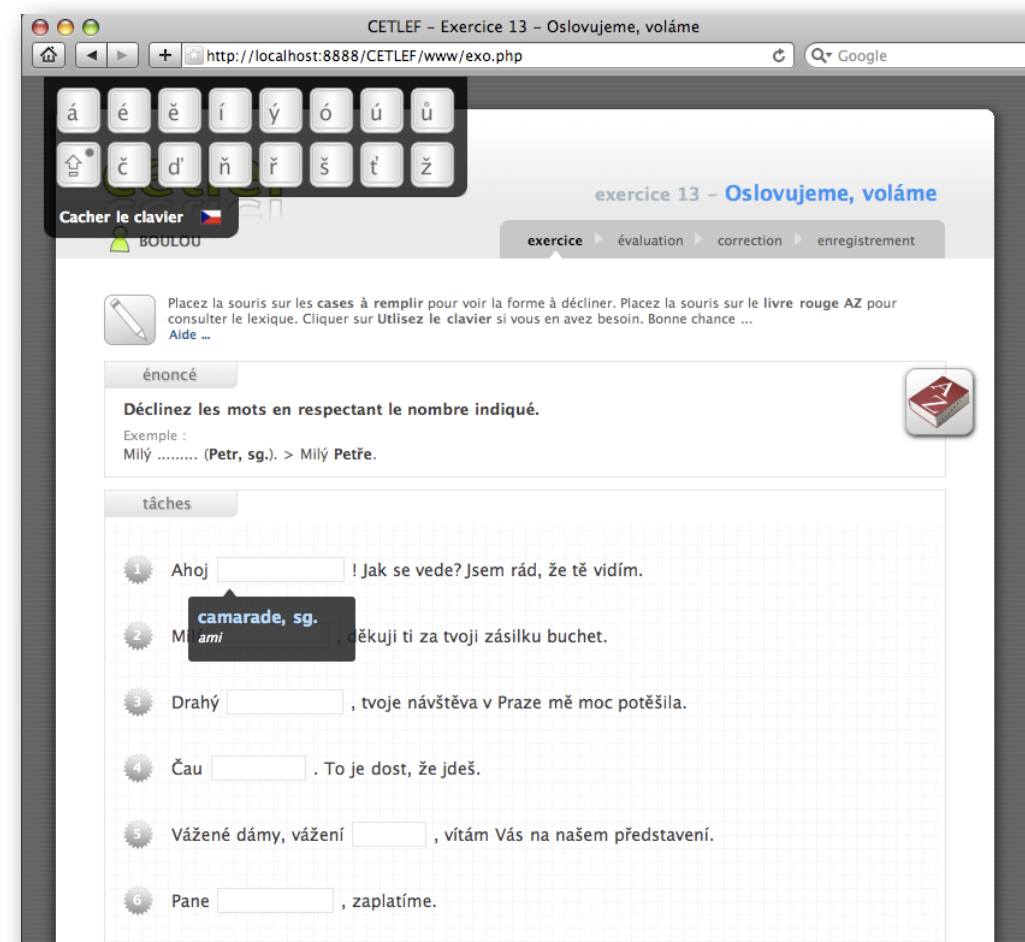
- V praxi se často zaujímá "**pedagogicky zodpovědný**" přístup:
  - **základní roviny** zpracování (pravopis, morfologie)
  - preference analýzy **omezených produkcí** před volným textem

# Platforma CETLEF

<http://www.cetlef.fr>

*Corriger / Comprendre les **Erreurs** en **Tchèque** - Langue **Étrangère** pour les **Francophones***

Webová aplikace určená frakofonním studentům češtiny, která nabízí **doplňovací cvičení zaměřená na českou deklinaci** s automatickou **diagnostikou chyb**.



# Doplňovací deklinační cvičení

## Uživatelské rozhraní

- 5 Vedle našeho  je hezká zahrada.
- 6 Jirka dává  růži
- 7 Dej mi kousek , nebuď lakomý.
- 8 Jdeme s kamarády na výlet do .
- 9 Do  vstoupil nějaký člověk.

matka, sg.

mère

*Jirka dává*

.....

*růži.*

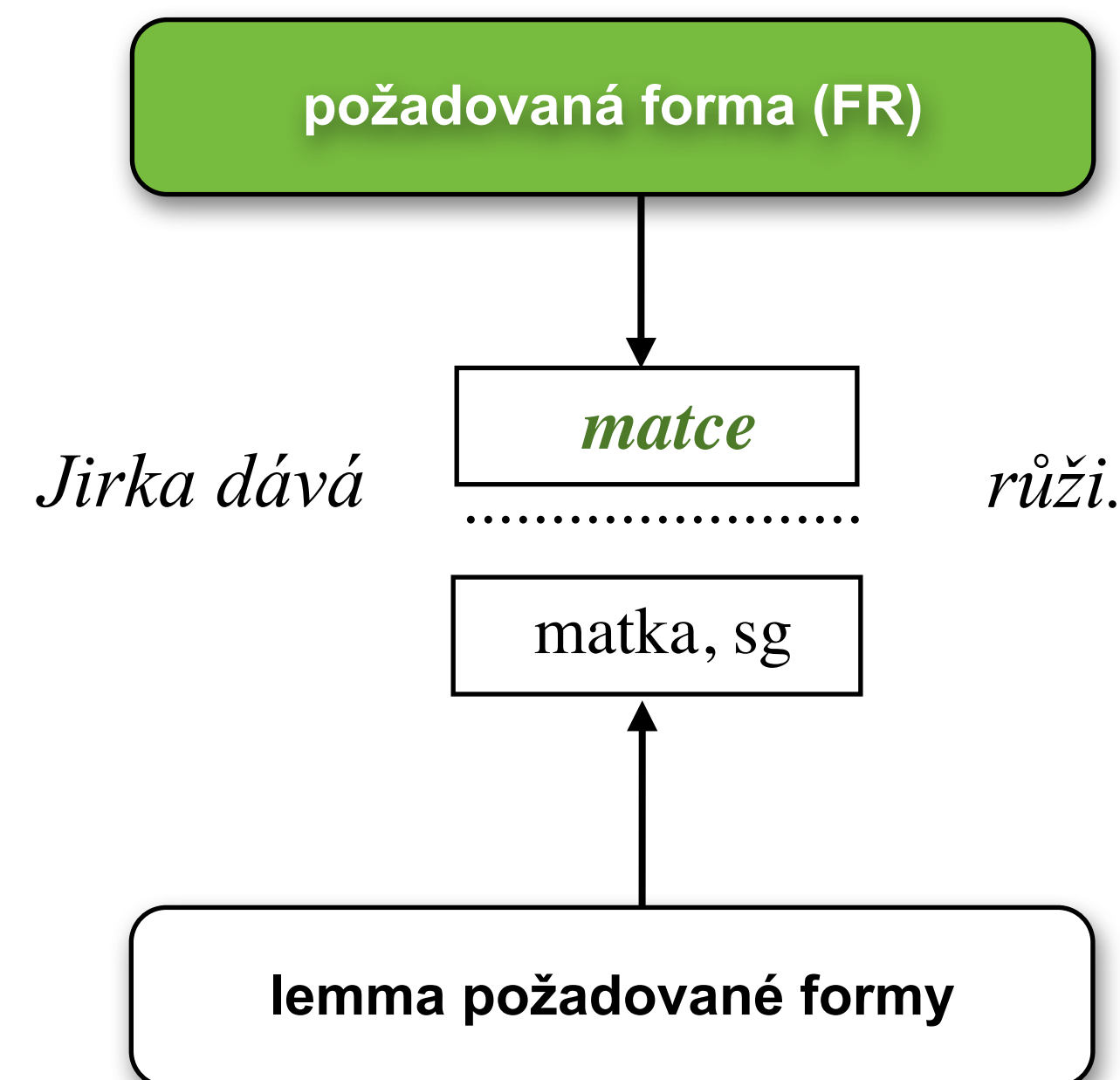
matka, sg

lemma požadované formy

# Doplňovací deklinační cvičení

**omezený** počet požadovaných forem:

- **varianta koncovky** (např. *páni* x *pánové*, *jazyku* x *jazyce*, atd.)
- **varianta zájmena** (*ho* x *jej* x *jeho*)
- **různé pádové formy** (např. *Petr je učitel* x *učitelem*; *smést prach pod skříň* x *skříní*)

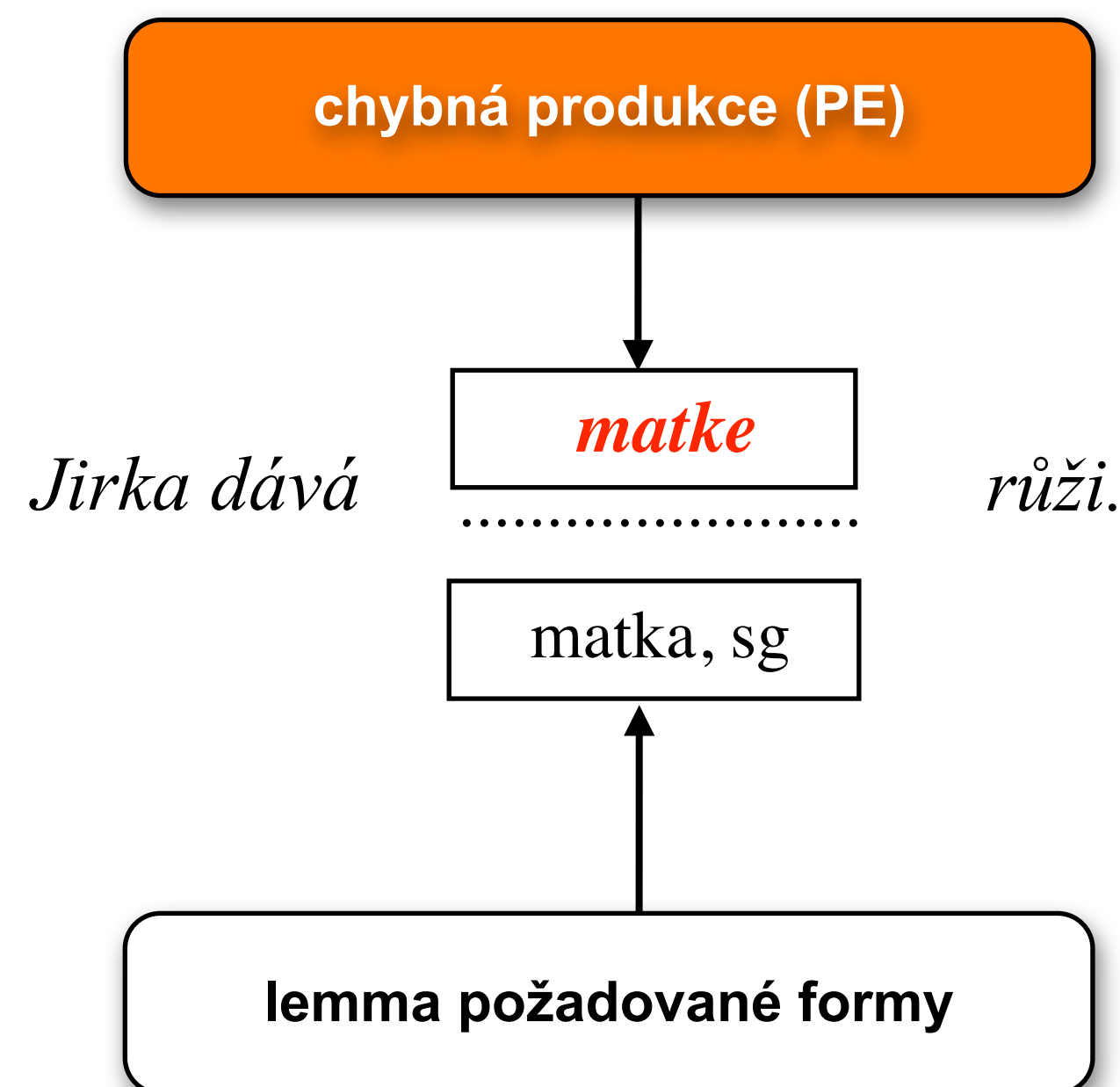


# Doplňovací deklinační cvičení

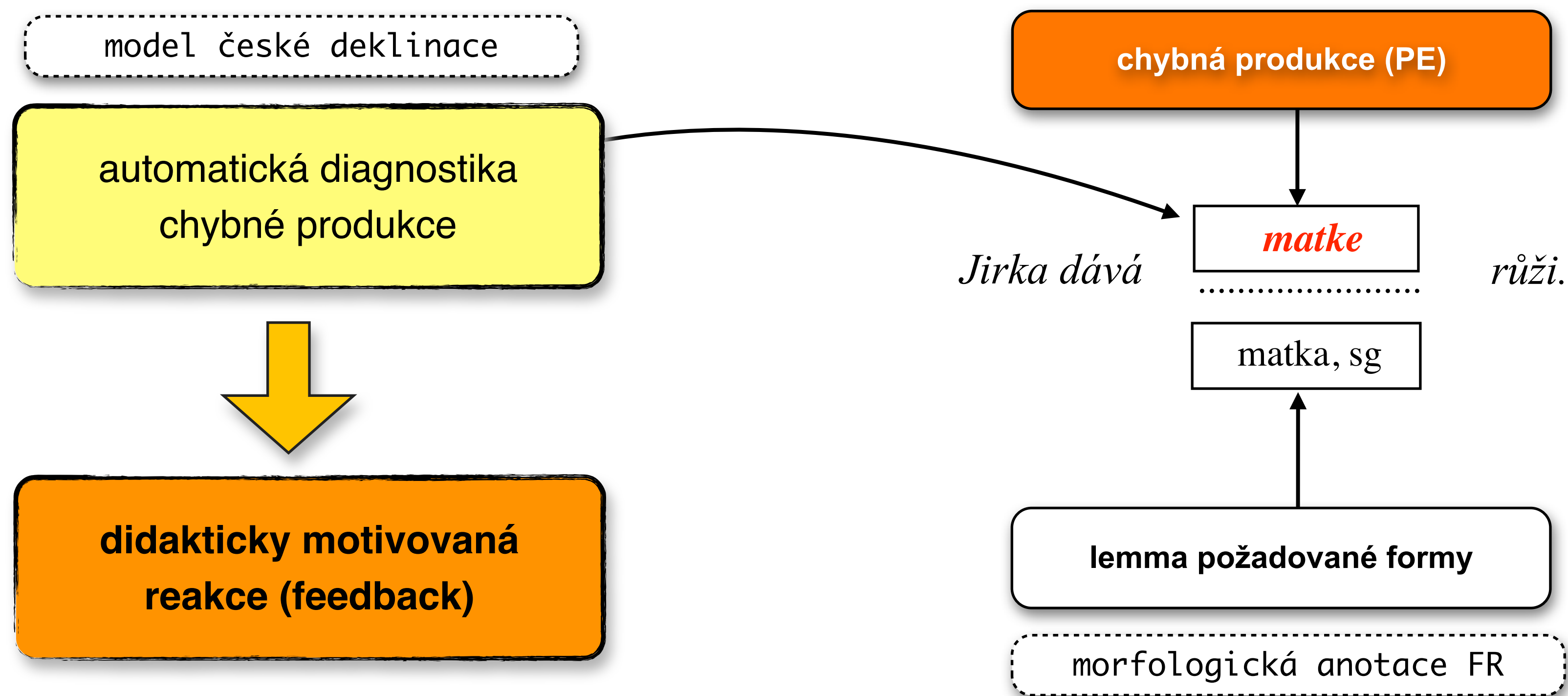


teoreticky **neomezený** počet chybných produkcí:

- obzvlášť pro deklinaci je obtíž s **explicitním vyjmenováním** možných odpovědí, které by sloužily jako podklad pro feedback.



# Doplňovací deklinační cvičení



# Doplňovací deklinační cvičení

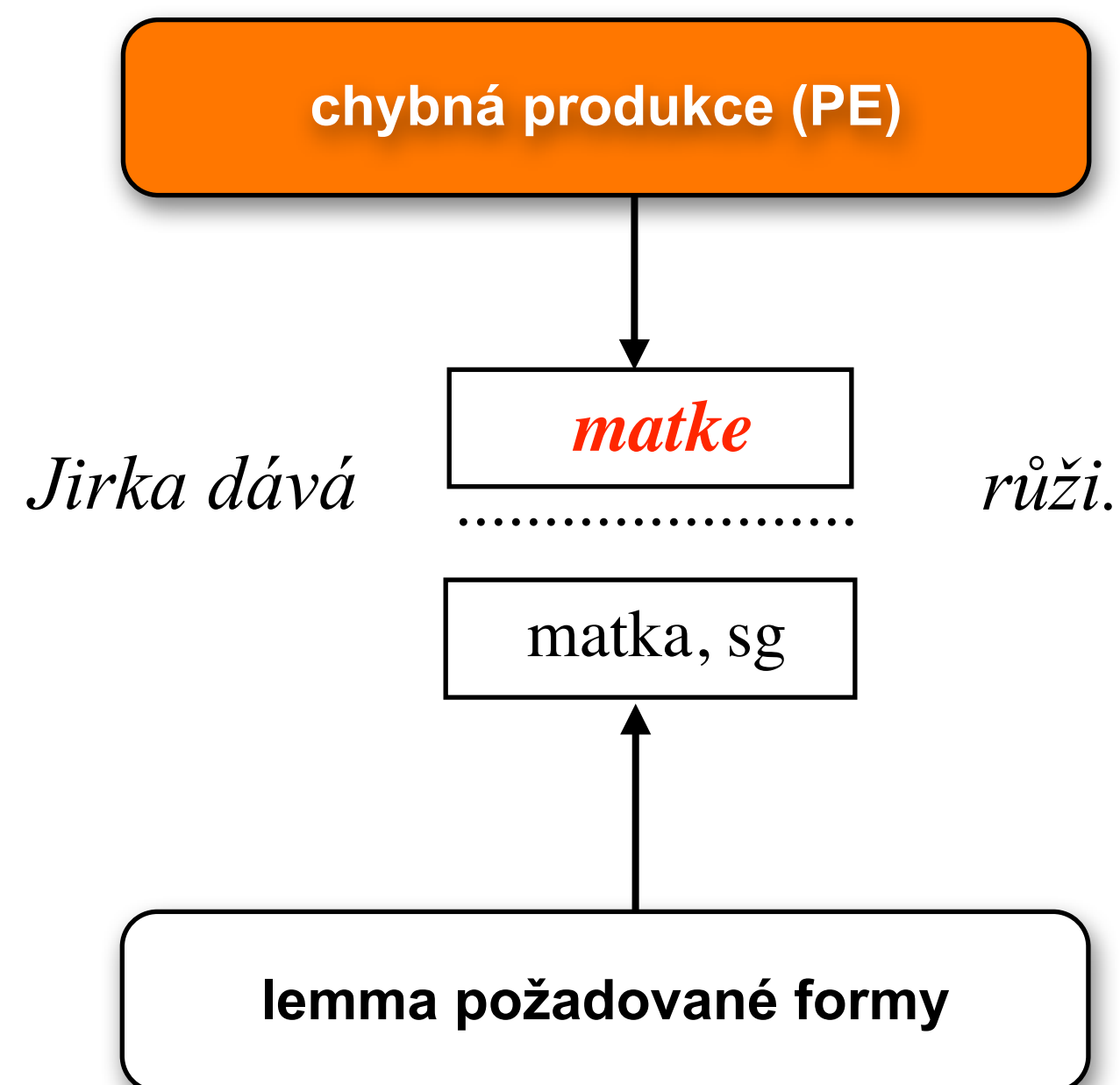
## Anotace chyby

matce@matke[k>c@sans]

## Uživatelské rozhraní

- 5 Vedle našeho ~~důmu~~ domu je hezká zahrada.
- 6 Jirka dává ~~matke~~ matce růži
- 7 Dej mi kous ~~leš~~ leše, nebud' lakomý.  

**erreur**   
 Aucune erreur de cas, de genre  
 ni de nombre mais vous n'avez  
 pas effectué l'alternance  $k > c$ .
- 8 Jdeme s kamarády na vylet do ~~lesu~~ lesa.
- 9 Do ~~místnosti~~ místnosti vstoupil nějaký člověk.



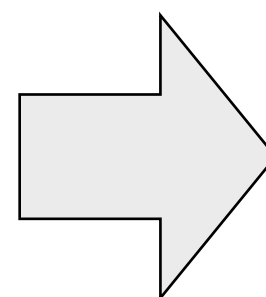
# Automatická diagnostika chybné produkce



**matce**

(matka | subst | N | zn | dat | sg | f | k>c)

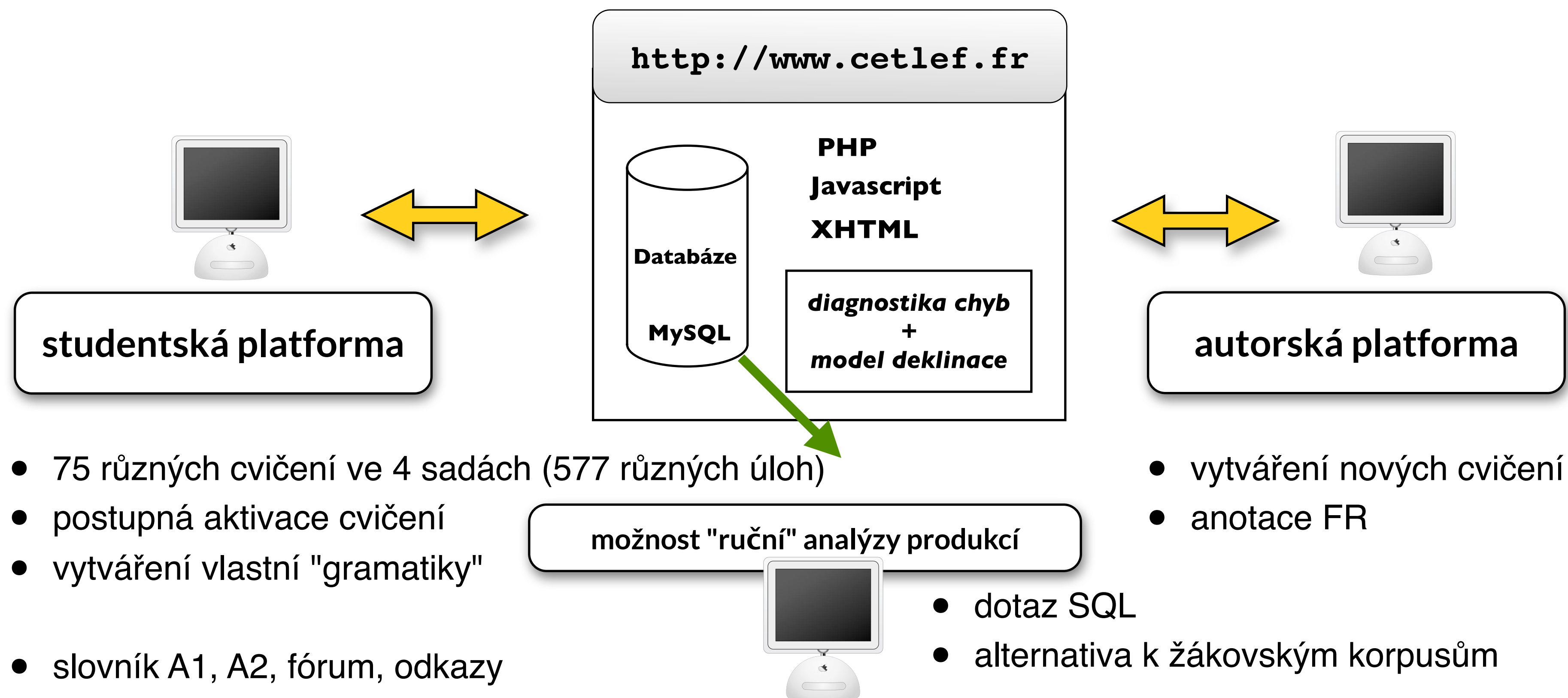
**matke**



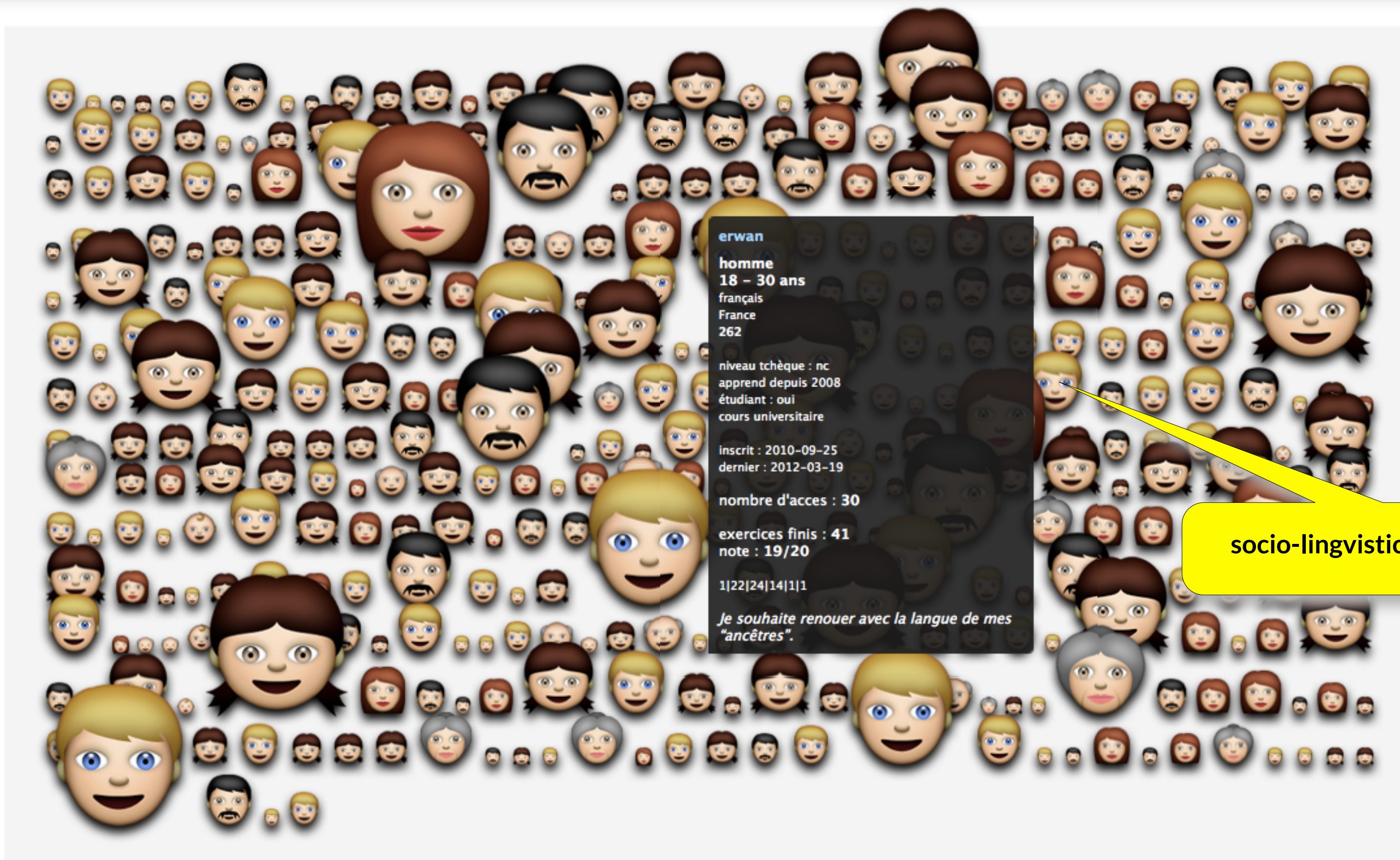
**LOC | zn | e | dat | sg | f | k>c@sans**

Aucune erreur de cas, de genre ni de nombre mais vous n'avez pas effectué l'alternance **k > c**.

# Architektura platformy CETLEF a její využití



# Přehled uživatelů od roku 2008

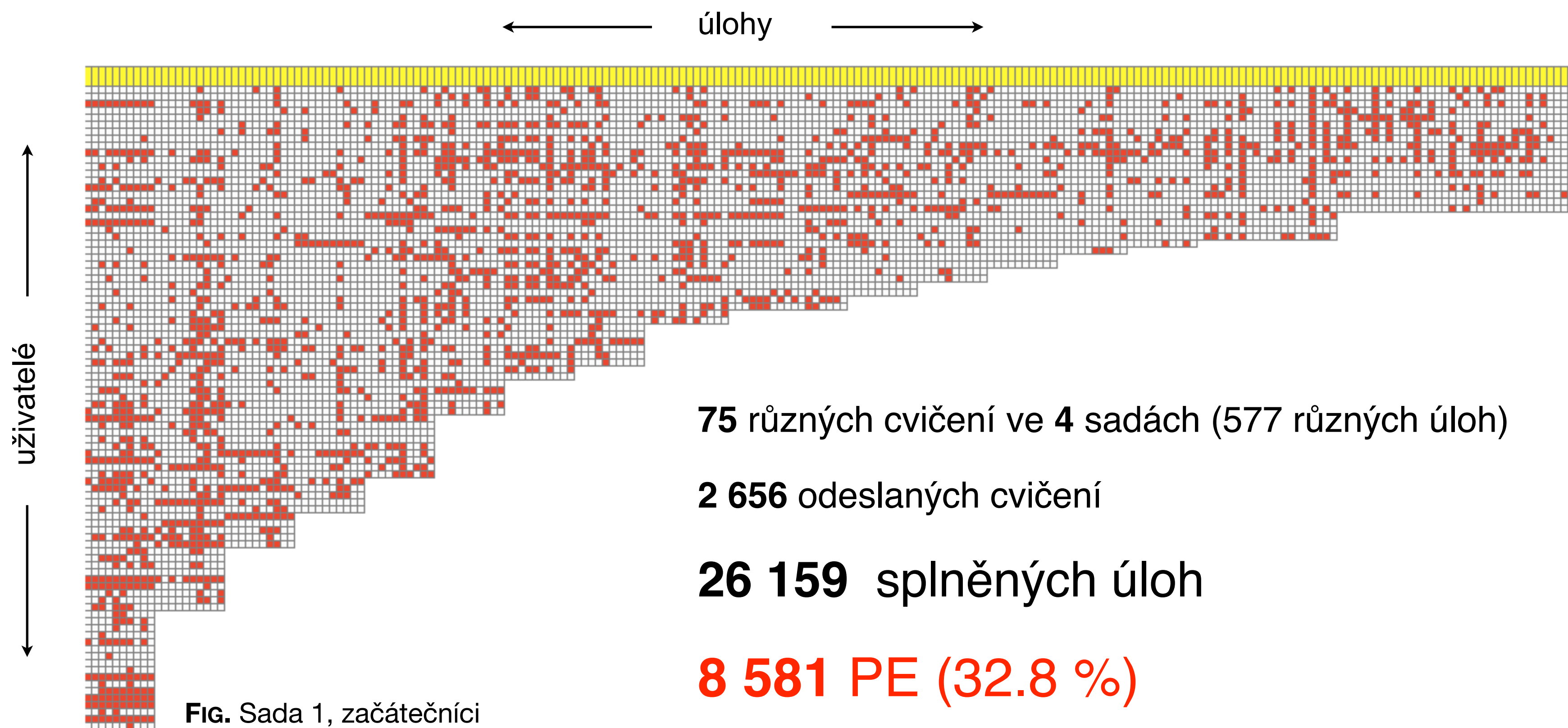


**395** zapsaných studentů

*velikost hlavičky  
vyjadřuje počet  
splněných cvičení*

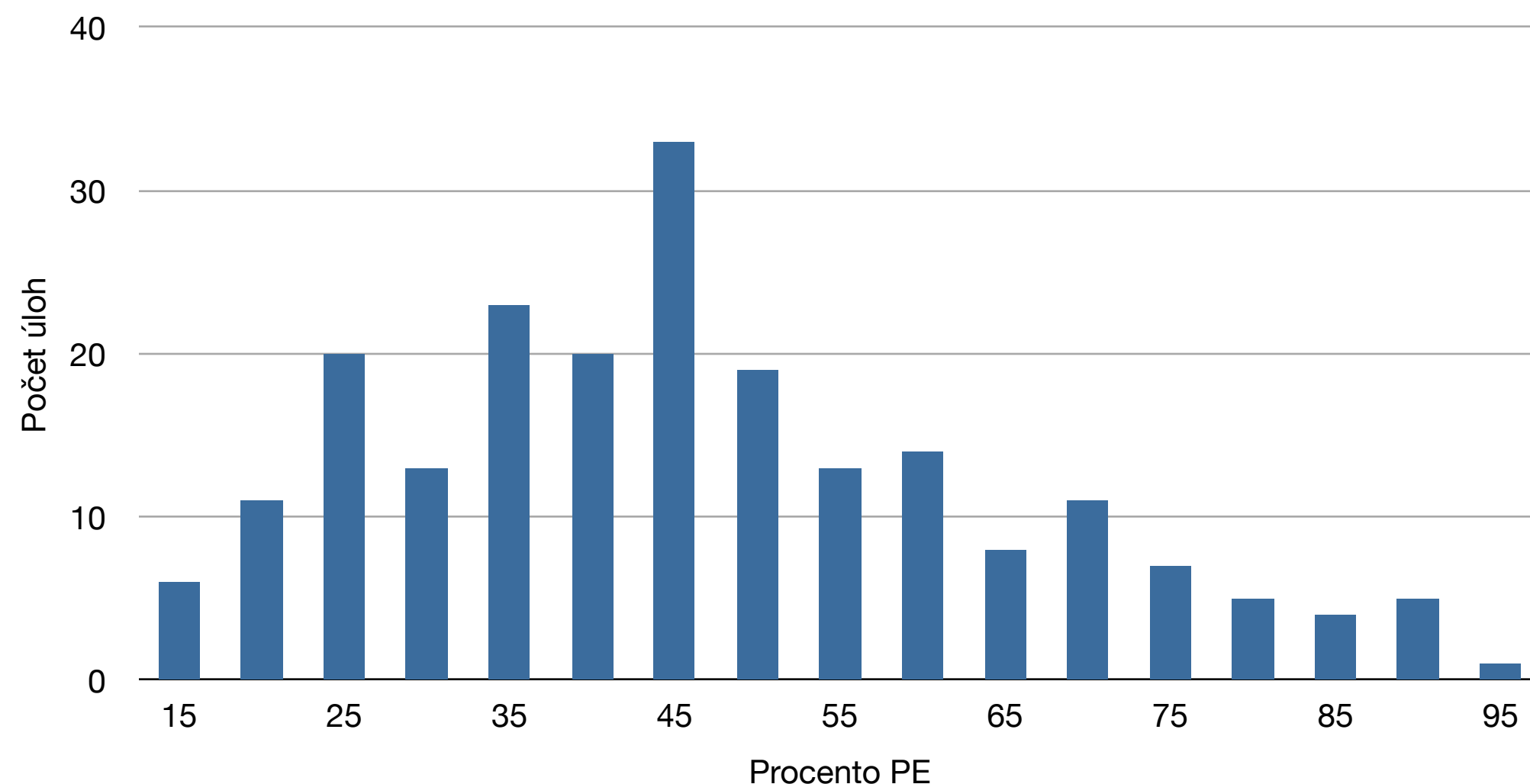
socio-lingvistické informace o uživateli

# Korpus produkcí - obsah databáze (prosinec 2012)



# Poměr PE vůči všem produkcím v rámci dané úlohy

Distribuce počtu PE v dané úloze

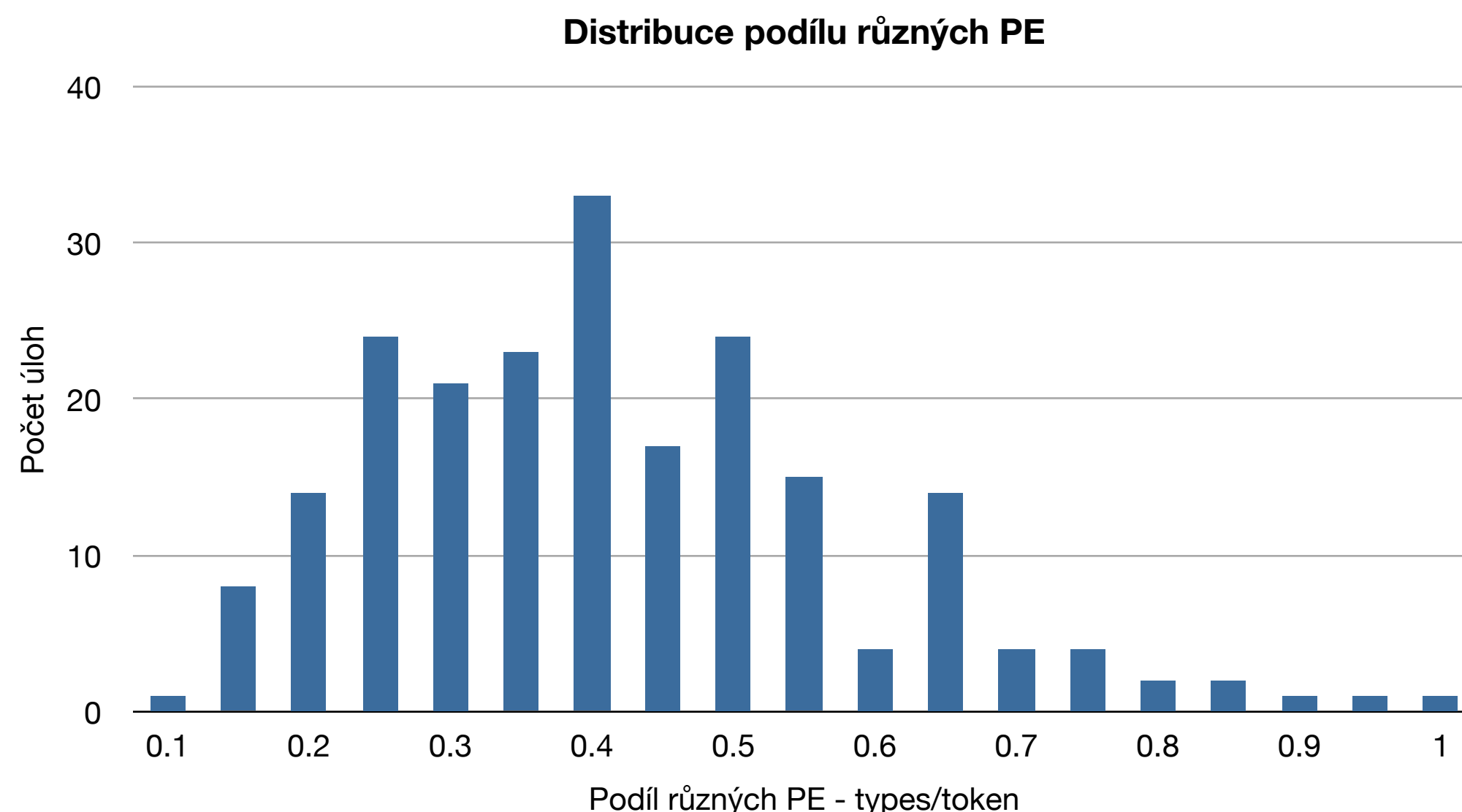


Nejčastější je případ, kdy je mezi všemi produkcemi v určité úloze zhruba **třetina chybná**.

údaje pouze pro úlohy, ve kterých je **minimálně 10 PE** (tzn. 213 z 577 úloh)

- Průměr pro všechny úlohy : **26 %**  
*vliv úloh s méně než 10 PE*

# Variabilita PE v rámci dané úlohy (PE\_types/PE\_tokens)



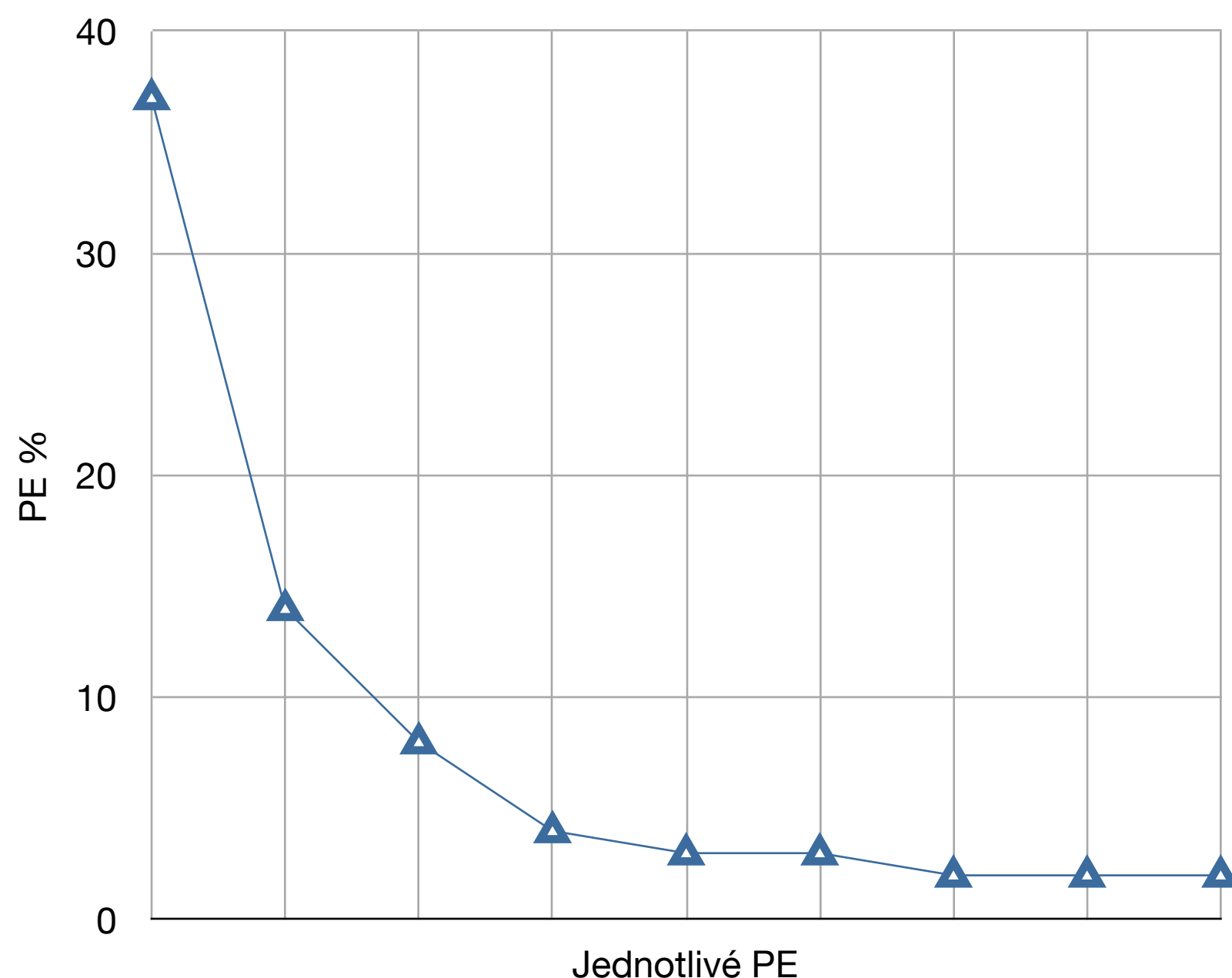
Nejčastější je případ, kdy je mezi PE v určité úloze zhruba tolik různých PE, kolik odpovídá **polovině počtu všech PE**.

údaje pouze pro úlohy, ve kterých je **minimálně 10 PE** (tzn. 213 z 577 úloh)

- Průměr pro všechny úlohy : **0.55**

*čím víc je PE pro danou úlohu, tím je zjevnější, že studenti inklinují k jisté PE.*

# Variabilita PE v rámci dané úlohy



Jirka má rád svoji ... (manželka, sg.)

**manželku** subst | N | zn\_Re | acc | sg | f | sans

99 PE (37 %) z 263 produkcí

33 PE (type), variabilita : **0.333**

**manzelku** 37 (37 %)

**manželka** 14 (14 %)

**manželkou** 8 (8 %)

**manželke** 4 (4 %)

**manželkú** 3 (3 %)

**manželkě** 3 (3 %)

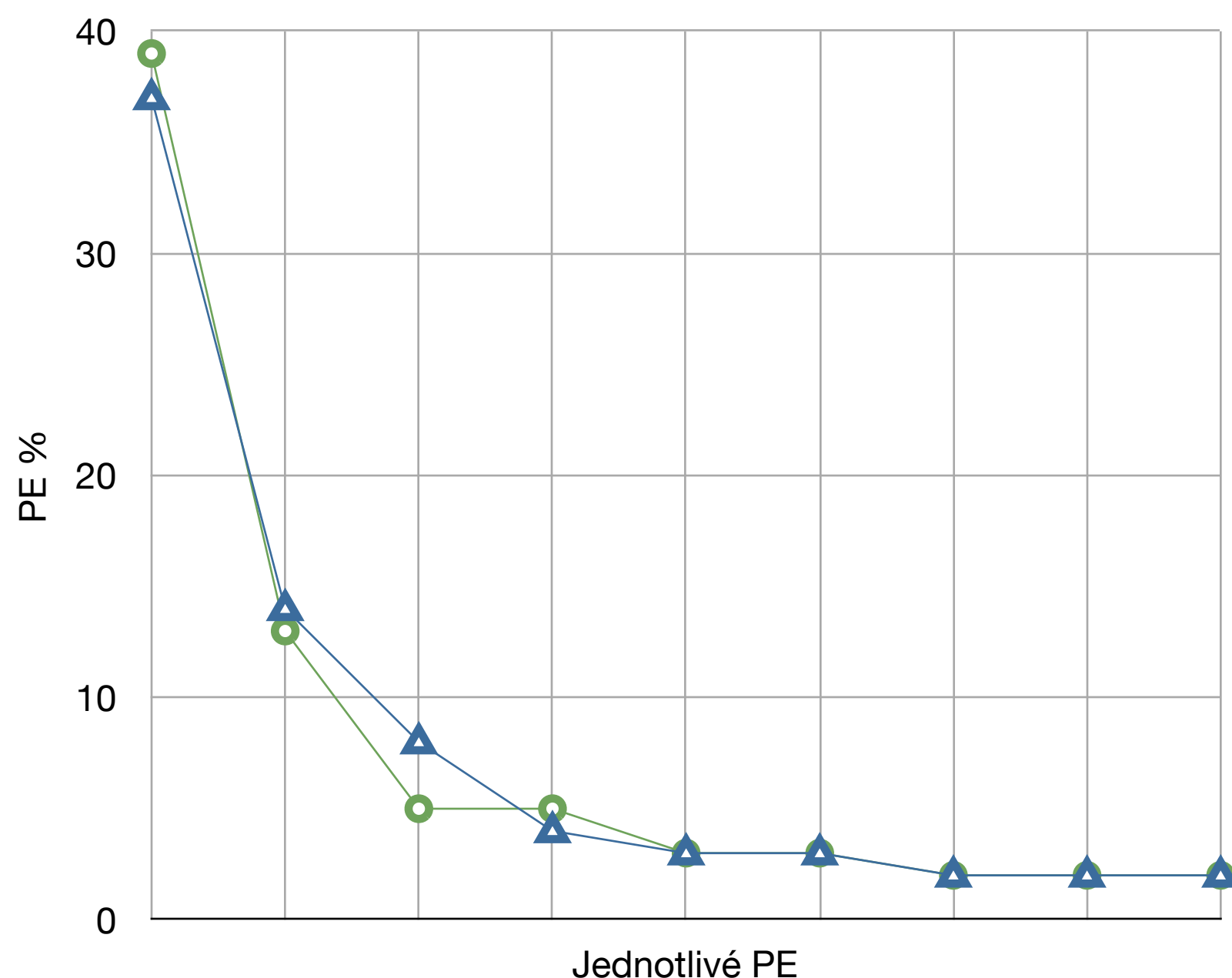
**manzelka** 2 (2 %)

**manzelkou** 2 (2 %)

**manželky** 2 (2 %)

.... atd.

# Variabilita PE v rámci dané úlohy



Studenti poslouchají ... (profesor, sg.)

**profesora** subst | N | pn | acc | sg | m | sans

94 PE (35 %) z 263 produkcí

29 PE (type), variabilita : **0.309**

**profesor** 37 (39 %)

**profesore** 12 (13 %)

**profesori** 5 (5 %)

**profesoru** 5 (5 %)

**profesorka** 3 (3 %)

**profesorí** 3 (3 %)

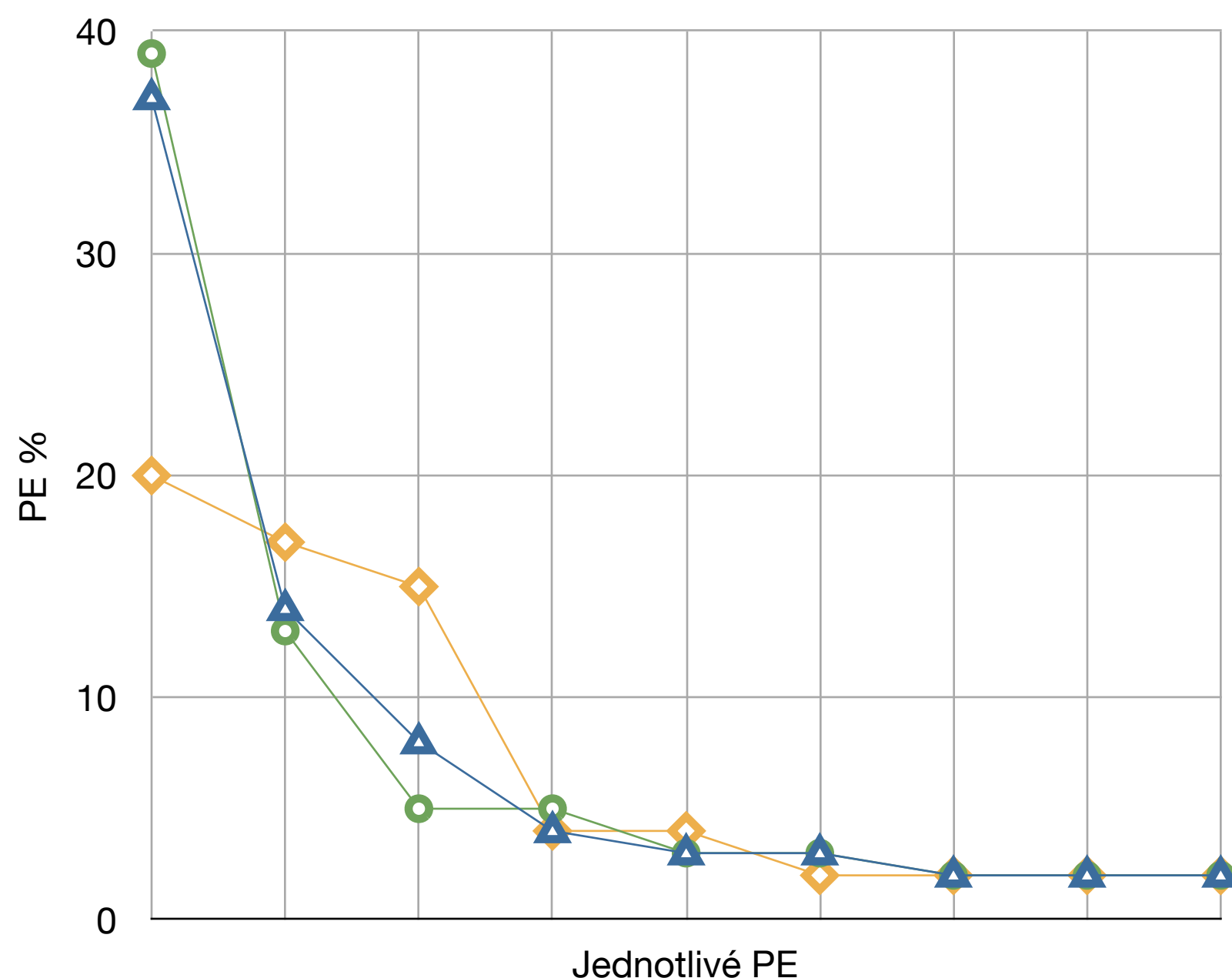
**PROFESORA** 2 (2 %)

**profera** 2 (2 %)

**profesoram** 2 (2 %)

.... atd.

# Variabilita PE v rámci dané úlohy



Petr píše ... (dopis, sg.) pro kamaráda.

**dopis** subst | N | hd\_Re | acc | sg | i | sans

46 PE (17 %) z 261 produkcí

23 PE (type), variabilita : 0.5

**dopise** 9 (20 %)

**dopisu** 8 (17 %)

**dopisa** 7 (15 %)

**DOPIS** 2 (4 %)

**dopís** 2 (4 %)

**Dopis** 1 (2 %)

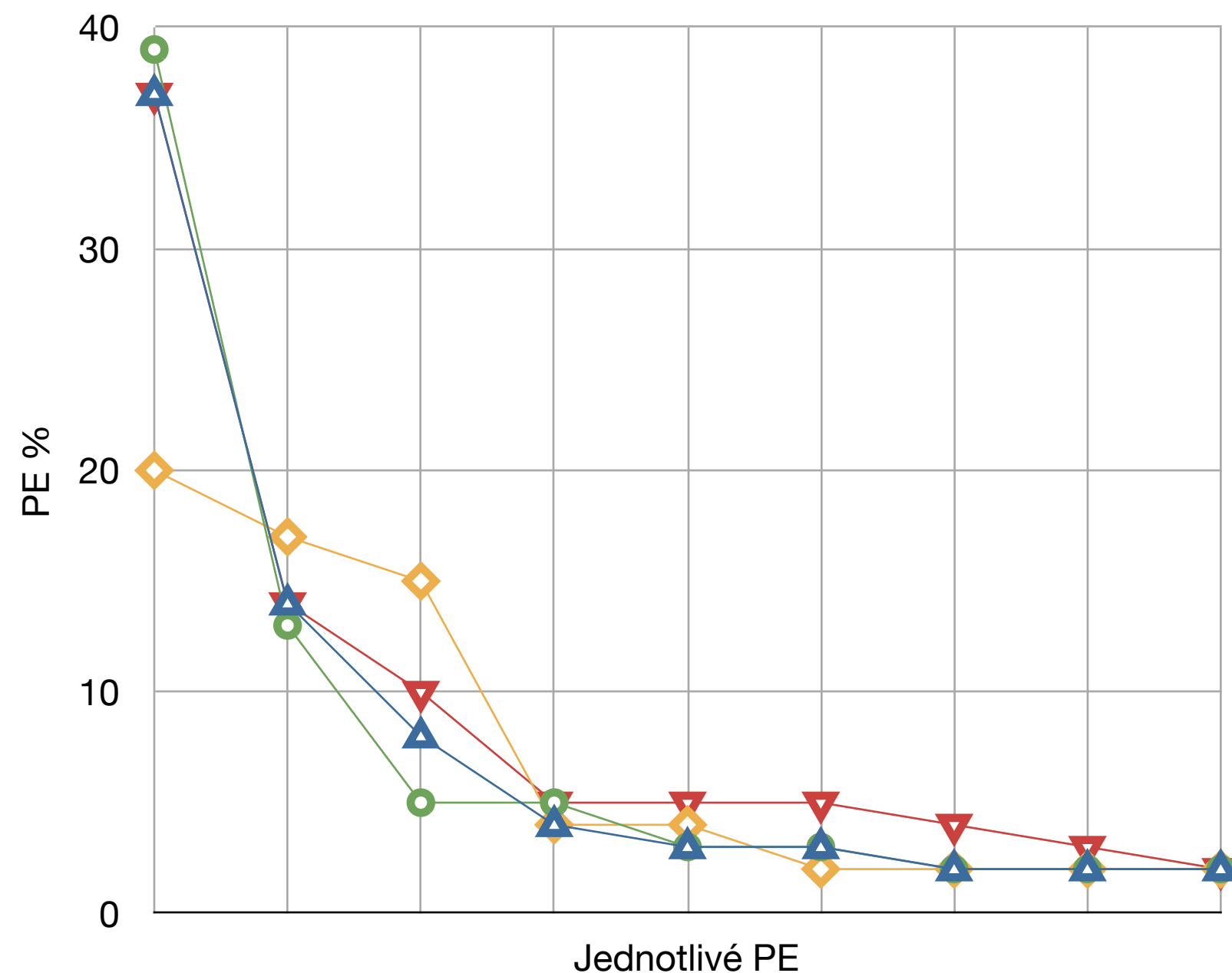
**aze** 1 (2 %)

**dopi** 1 (2 %)

**dopis sg** 1 (2 %)

.... atd.

# Variabilita PE v rámci dané úlohy

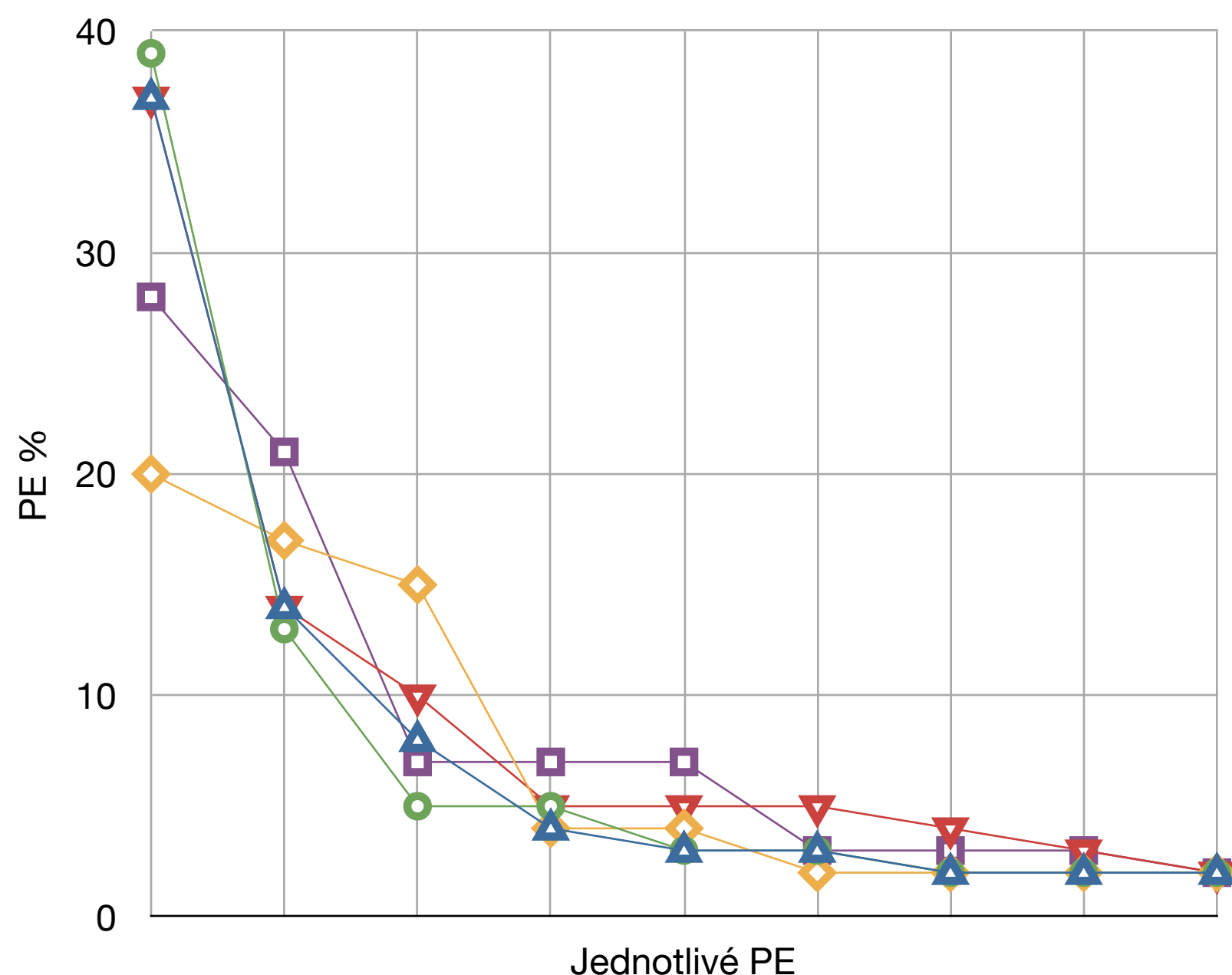


Každou neděli dávám ženě ... (růže, sg.)  
**růži** subst | N | rz | acc | sg | f | sans

153 PEs (58 %) z 260 produkcí  
 33 PE (type), variabilita : **0.216**

**růže** 56 (37 %)  
**ruze** 21 (14 %)  
**ruzi** 15 (10 %)  
**růžu** 8 (5 %)  
**růze** 7 (5 %)  
**růža** 7 (5 %)  
**růží** 6 (4 %)  
**ruži** 4 (3 %)  
**ruzy** 3 (2 %)  
 .... atd.

# Variabilita PE v rámci dané úlohy



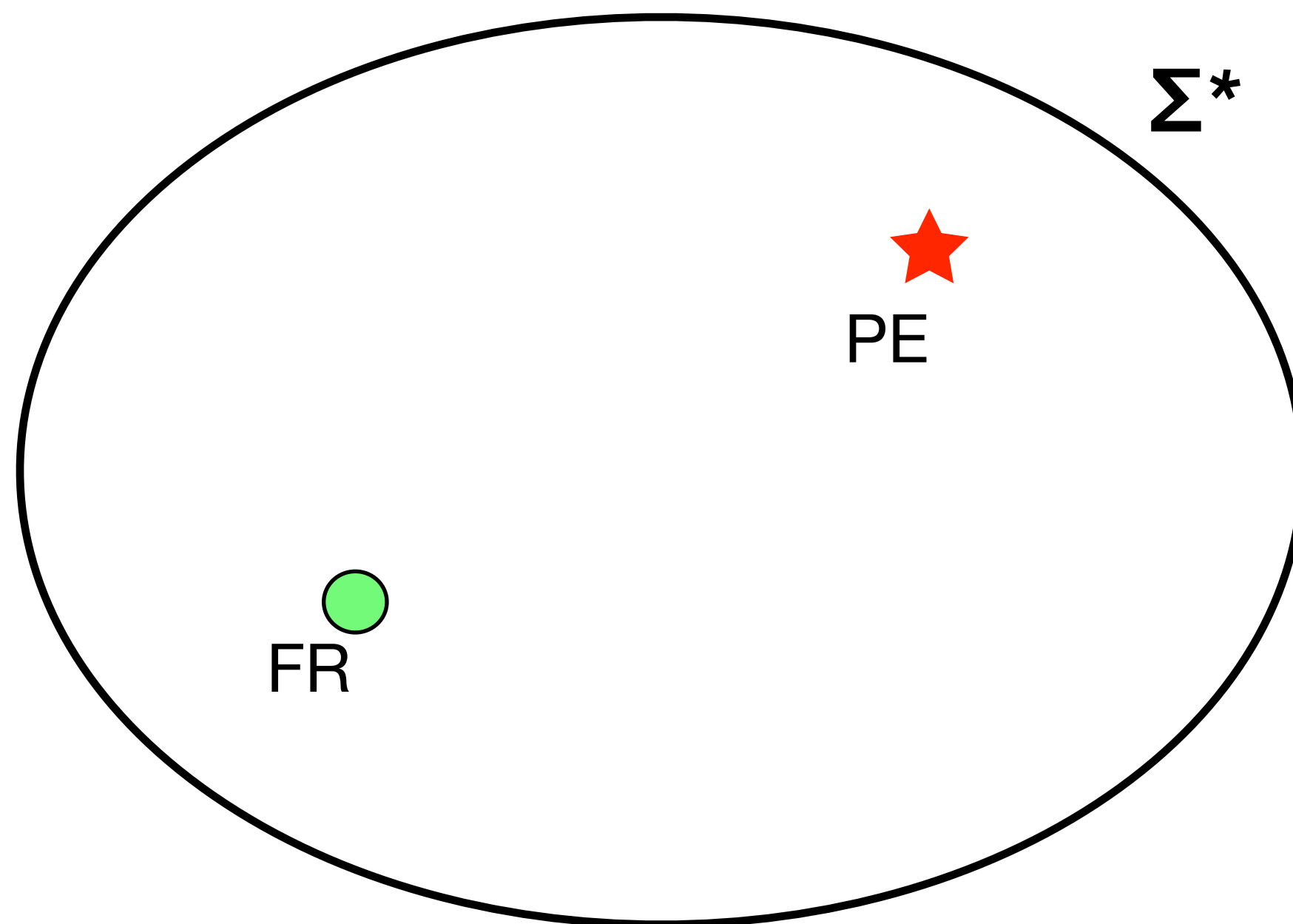
Mám ... (radost, sg.) , že se učíte česky.  
**radost** subst | N | kt | acc | sg | f | sans

58 PEs (22 %) z 254 produkcí  
 20 PE (type), variabilita : **0.345**

**radosti** 16 (28 %)  
**rad** 12 (21 %)  
**radoste** 4 (7 %)  
**radostu** 4 (7 %)  
**rád** 4 (7 %)  
**rada** 2 (3 %)  
**radosta** 2 (3 %)  
**radostě** 2 (3 %)  
**radost sg** 1 (2 %)  
 .... atd.

# Formální vymezení chybných izolovaných forem

- každý řetězec nad abecedou  $\Sigma$  odlišný od požadované formy FR.



$$\mathbf{PE} \in \Sigma^* \wedge \mathbf{PE} \neq \mathbf{FR}$$

matce@matku

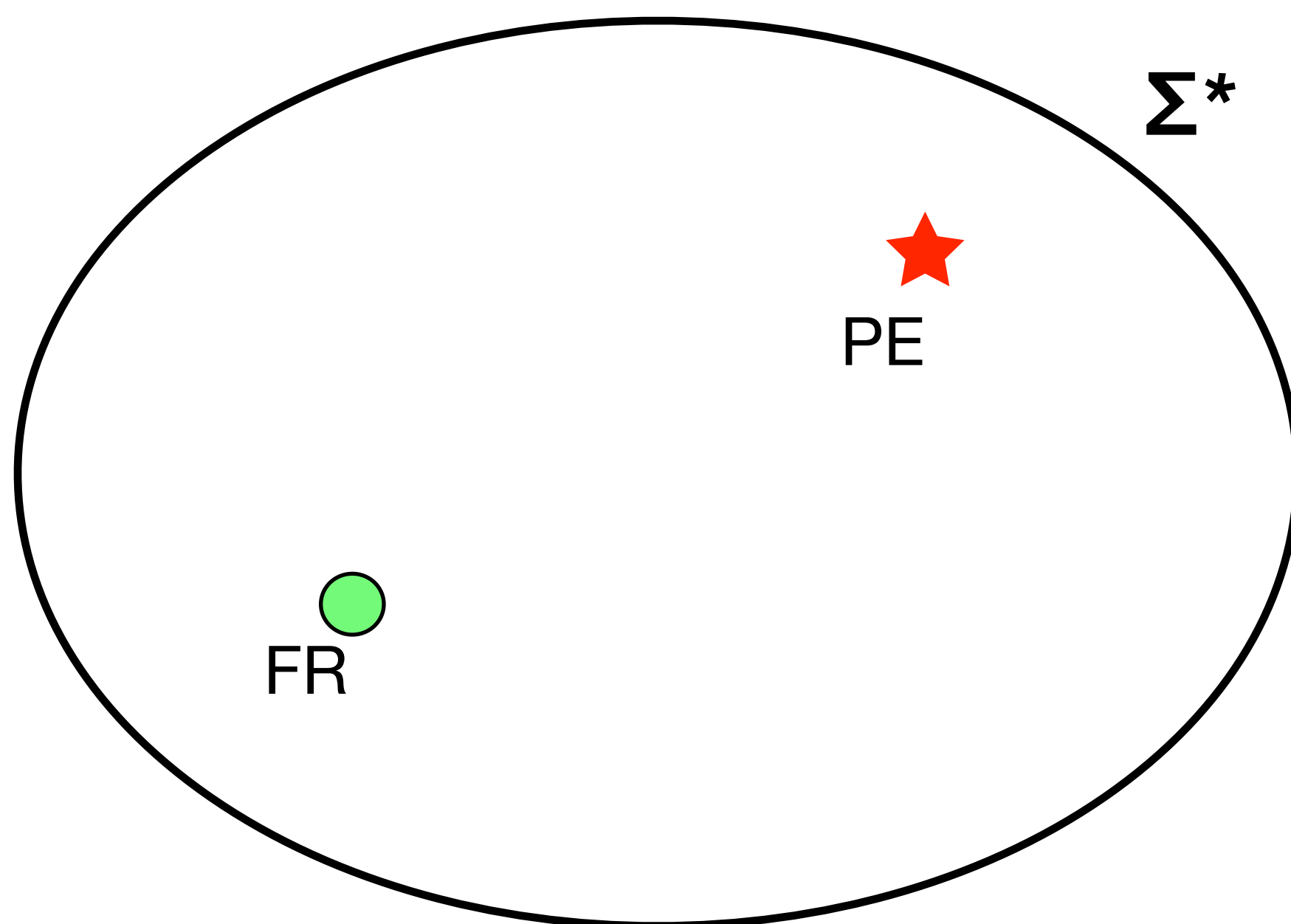
matce@matke

matce@mzrke

matce@mate

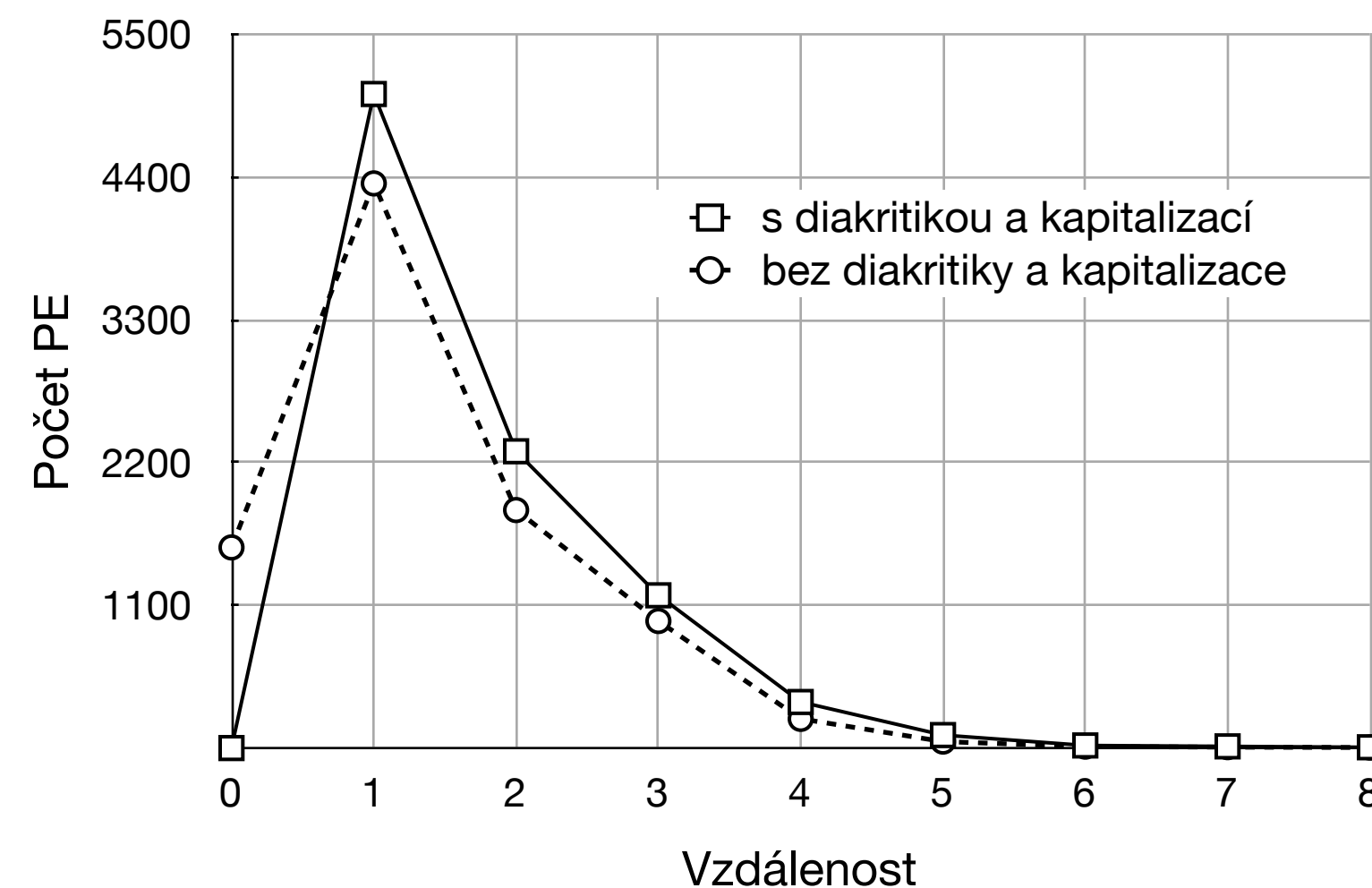
# Levenshteinova vzdálenost

- každý řetězec nad abecedou  $\Sigma$  odlišný od požadované formy FR.



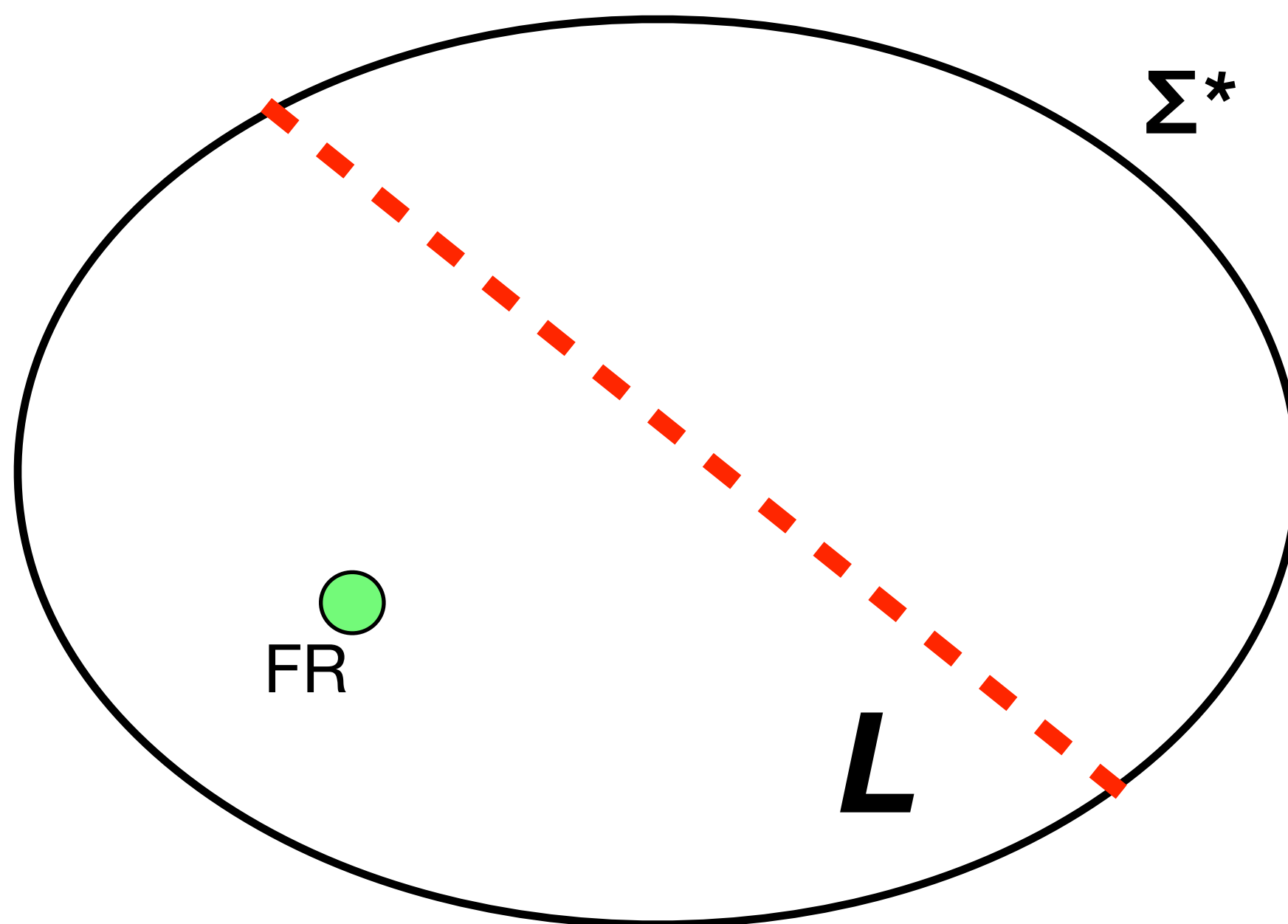
$$PE \in \Sigma^* \wedge PE \neq FR$$

## Levenshteinova vzdálenost



# Existující vs. neexistující formy v daném jazyce

- každý řetězec nad abecedou  $\Sigma$  odlišný od požadované formy FR z jazyka L.

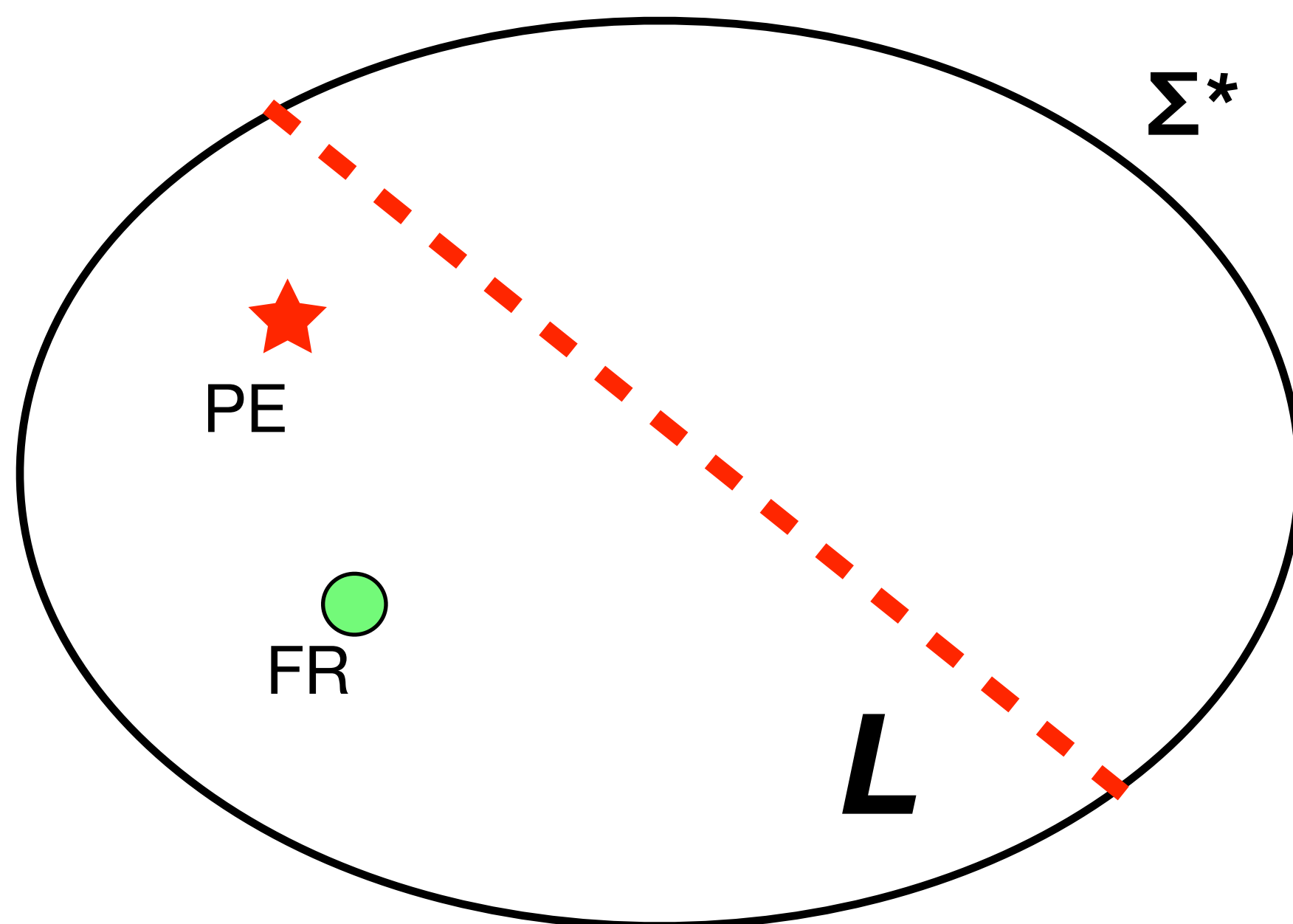


**L** - množina obsahující všechny existující formy daného jazyka

**FR**  $\in$  **L**

# Existující vs. neexistující formy v daném jazyce

- každý řetězec nad abecedou  $\Sigma$  odlišný od požadované formy FR z jazyka L.



$$PE \in L \wedge PE \neq FR$$

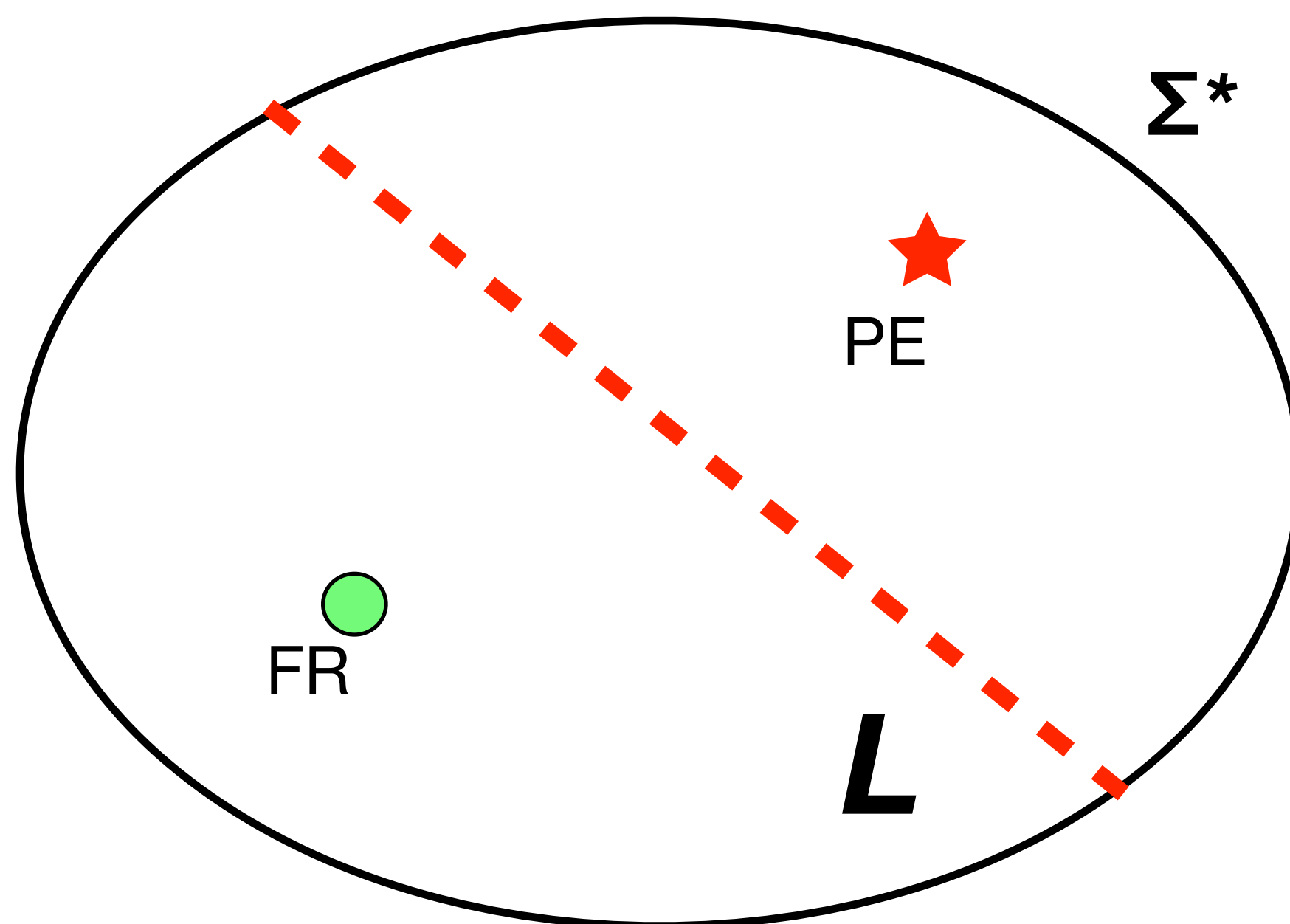
matce@matku

matce@mate

55 % ze všech PE (podle Aspell)

# Existující vs. neexistující formy v daném jazyce

- každý řetězec nad abecedou  $\Sigma$  odlišný od požadované formy  $FR$  z jazyka  $L$ .



$$PE \notin L \wedge PE \neq FR$$

matce@matke

matce@mzrke

45 % ze všech PE (podle Aspell)

# Morfologická interpretace

## ► Hypotéza předpokládající morfologicky motivované chyby

*podepřená pracemi v oboru osvojování cizího jazyka*

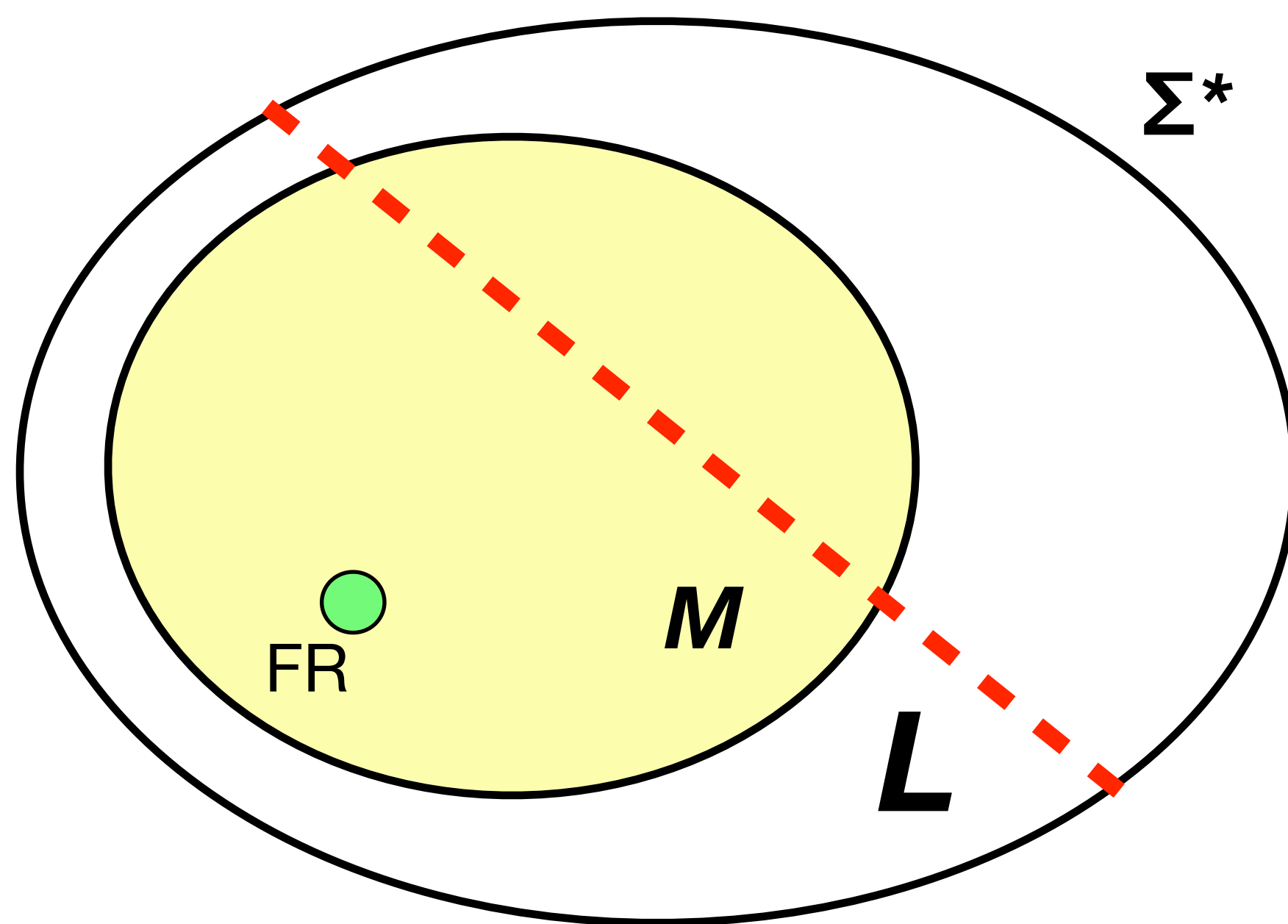
*Chybná produkce (**PE**) není nahodilá forma, ale lze ji vypočítat z požadované formy (**FR**)*

- nesprávná **kombinace** morfémů (porušení morfematických omezení)
- neadekvátní aplikace alternačních **pravidel**

### Motivace:

- rozlišování mezi chybami v **kompetenci** (error) a **performanci** (mistake)
- **interference**, vlastnosti **mezijazyka**, u-shaped learning curve

# Morfologická interpretace

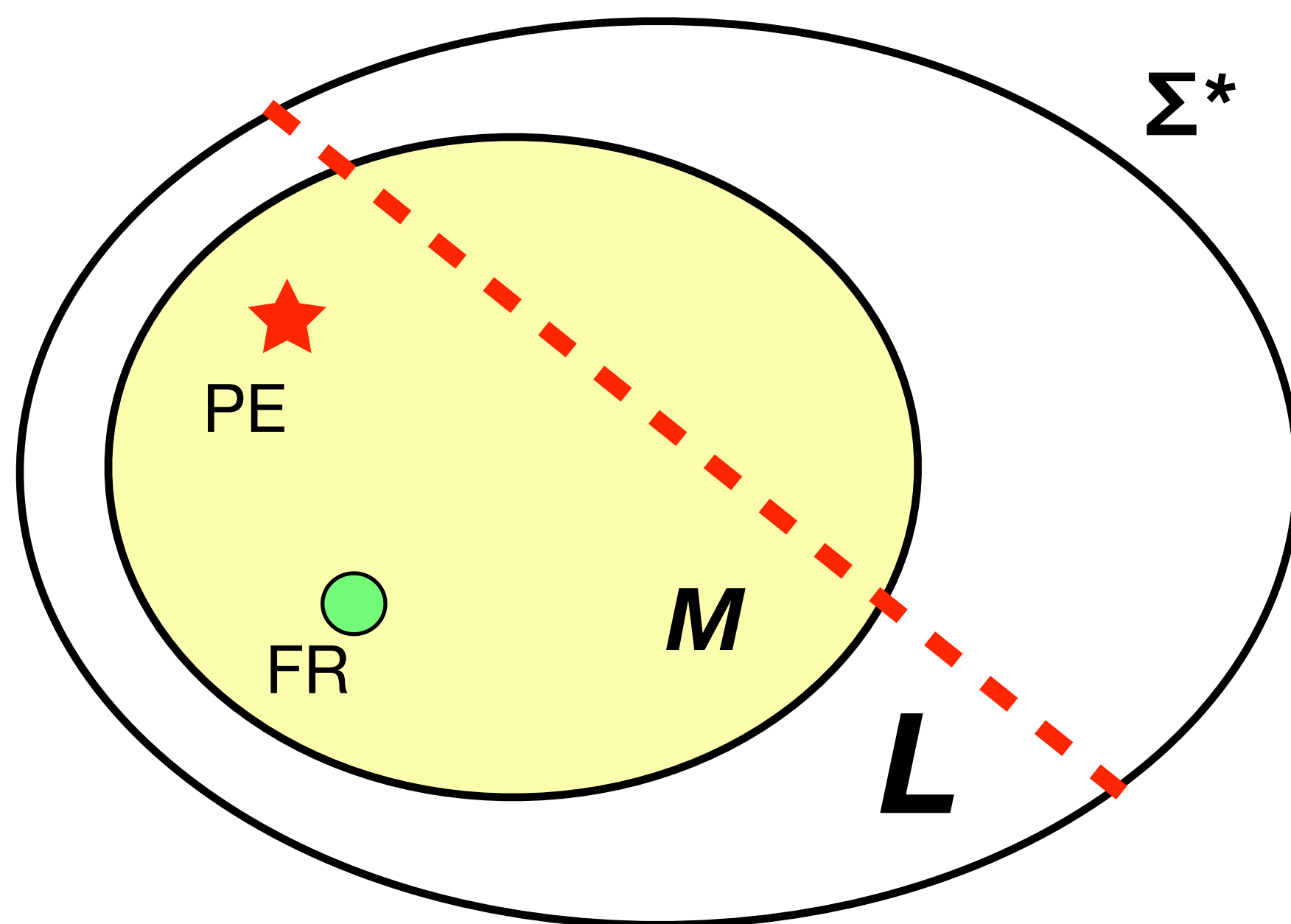


**$M$**  : množina obsahující všechny formy, které mají **morfologickou interpretaci** pro  $FR$ , tzv. **hypotetické formy**

## Hypotetické formy (FH) z $FR$

všechny formy vytvořené od **kmenu**  $FR$  pomocí **jakékoliv koncovky** a s eventuální chybnou realizací alternace, diakritiky a kapitalizace.

# Morfologická interpretace

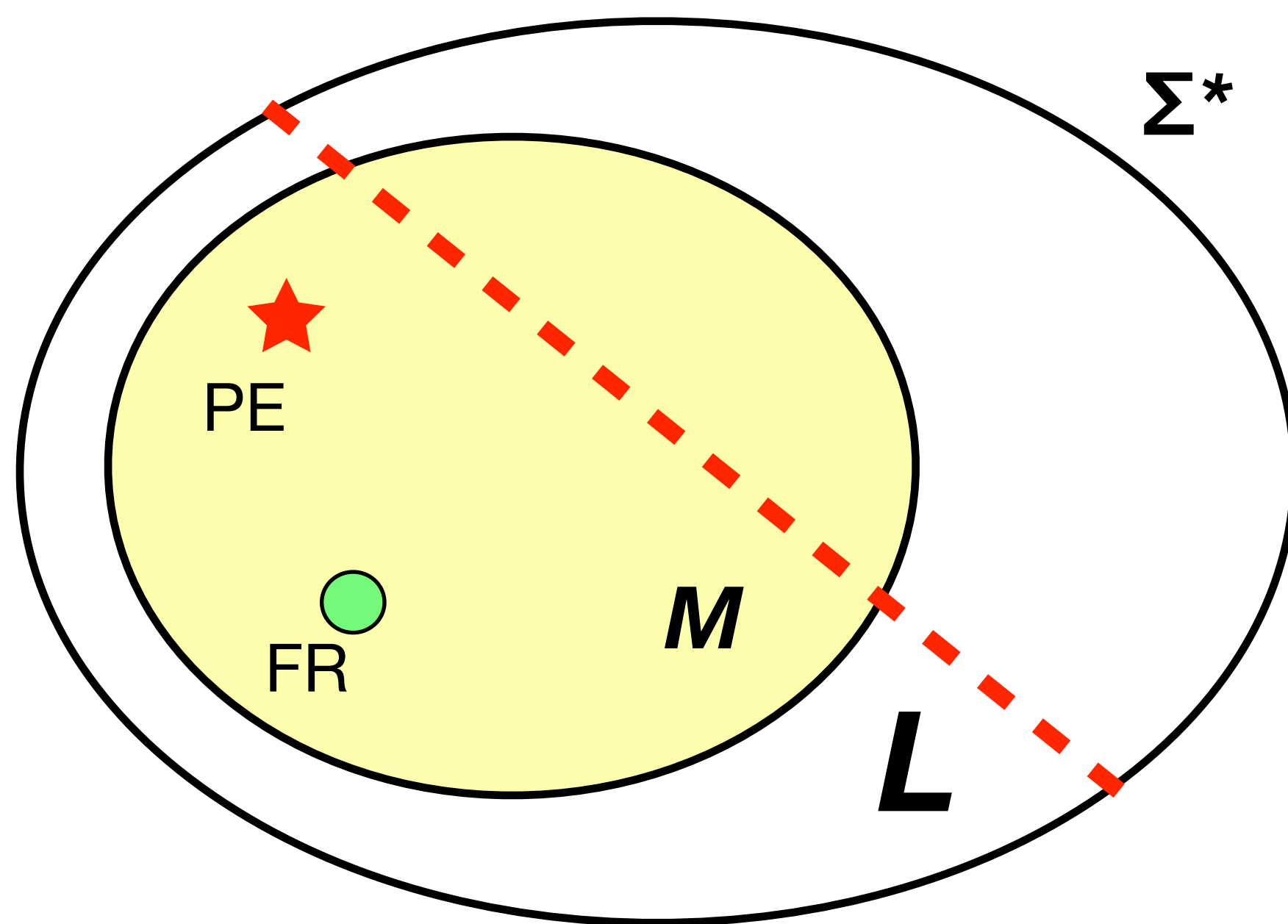


$$PE \in L \wedge PE \in M \wedge PE \neq FR$$

Morfologická interpretace PE 

$$PE = FH \Rightarrow PEM$$

# Morfologická interpretace

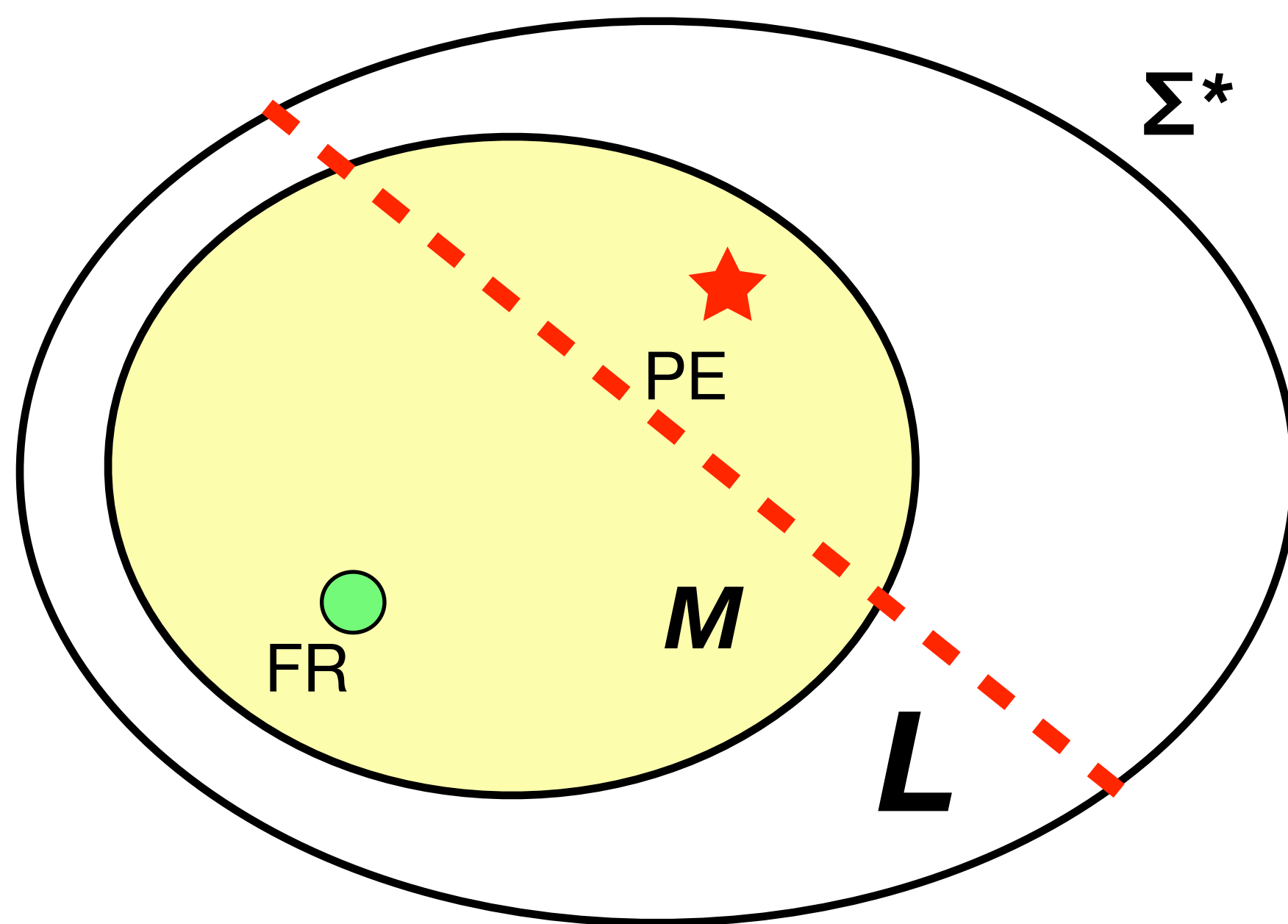


$$\mathbf{PE} \in \mathbf{L} \wedge \mathbf{PE} \in \mathbf{M} \wedge \mathbf{PE} \neq \mathbf{FR}$$

matce@matku

**57 %** ze všech PEM (*podle Aspell*)

# Morfologická interpretace

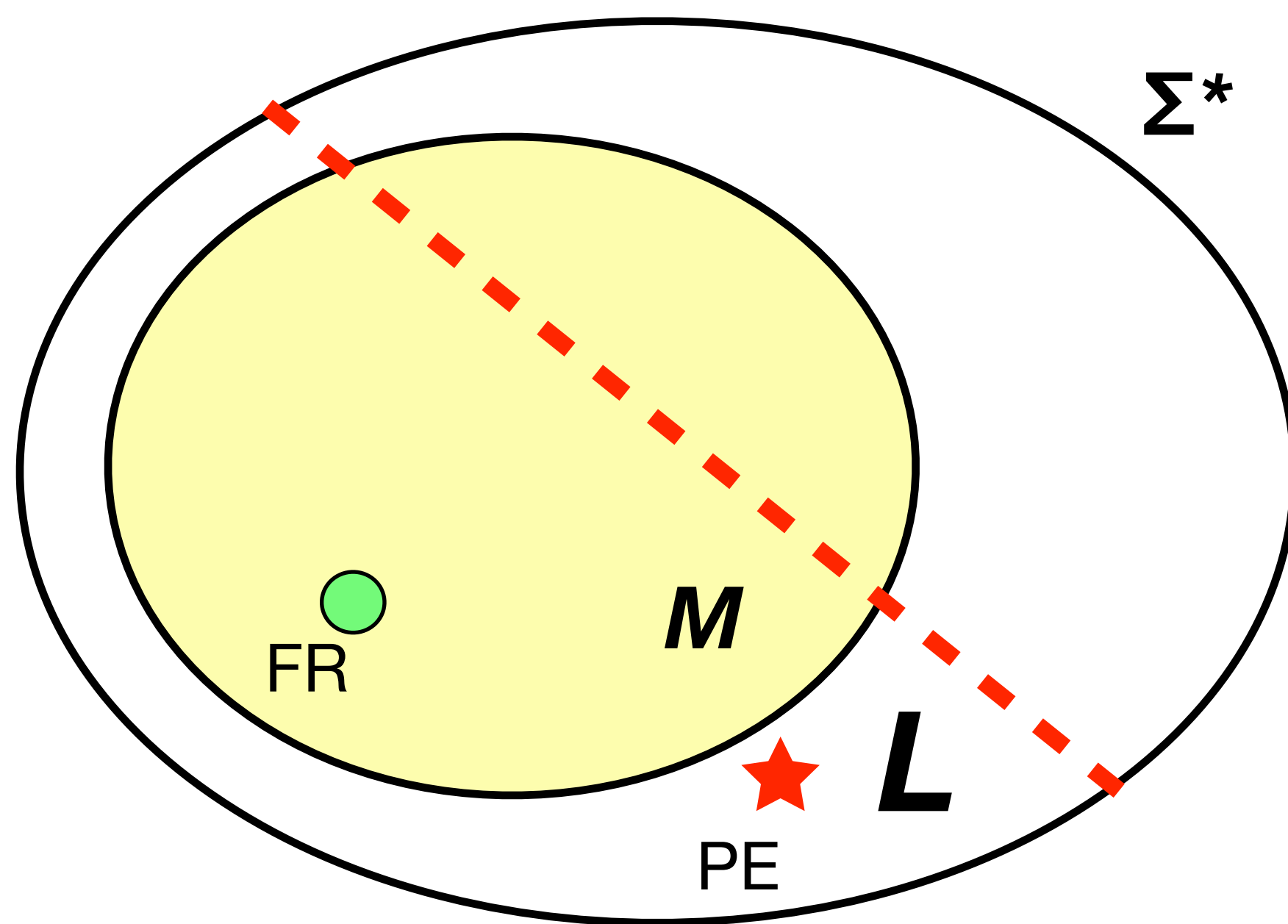


$$\mathbf{PE} \notin \mathbf{L} \wedge \mathbf{PE} \in \mathbf{M} \wedge \mathbf{PE} \neq \mathbf{FR}$$

matce@matke

**43 %** ze všech PEM (*podle Aspell*)

# Nemorfologická interpretace

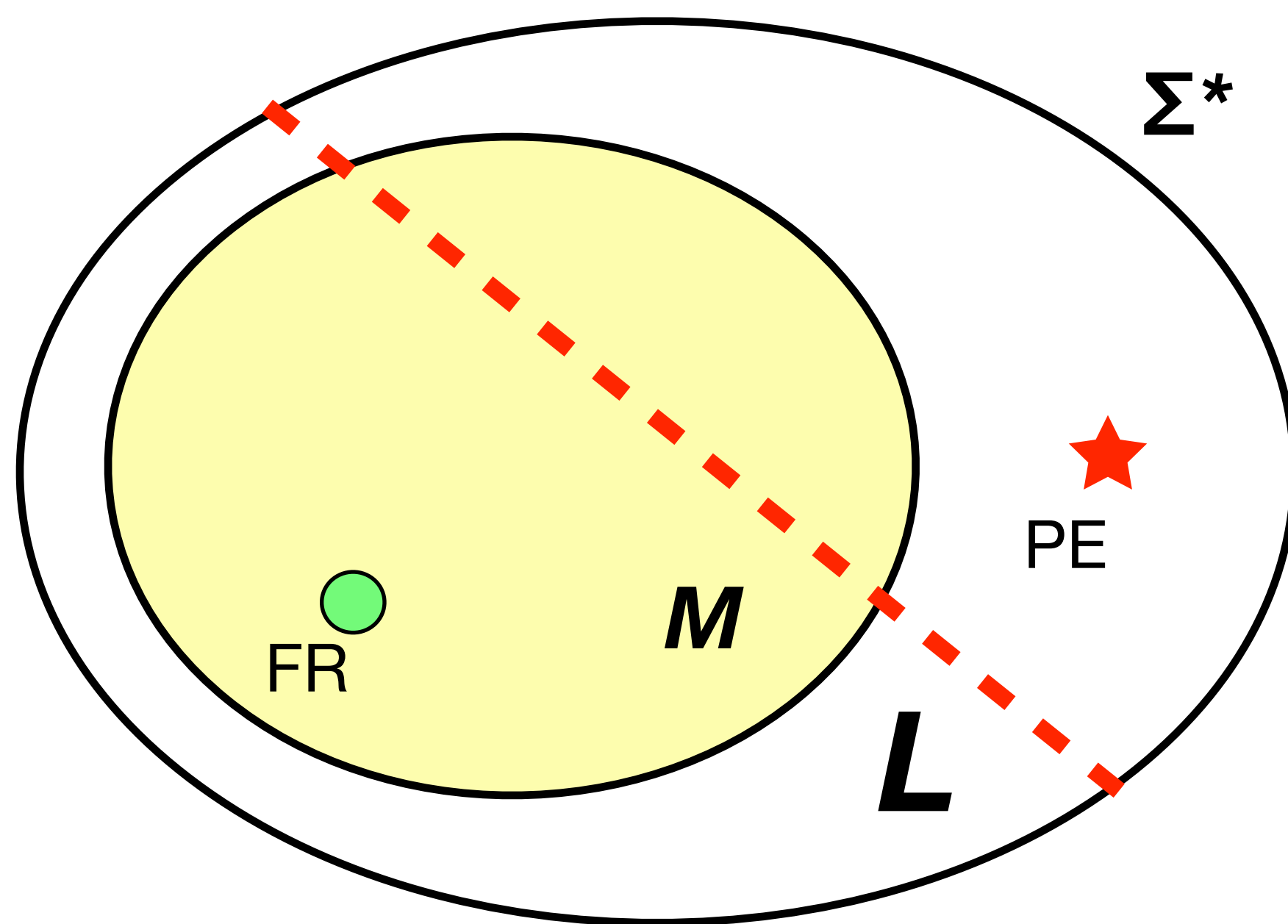


$$PE \in L \wedge PE \notin M \wedge PE \neq FR$$

matce@mate

43 % ze všech non-PEM (podle Aspell)

# Nemorfologická interpretace



$$PE \in L \wedge PE \notin M \wedge PE \neq FR$$

matce@mzrke

57 % ze všech non-PEM (podle Aspell)

# Nalezení interpretace pro danou PE

- základní předpoklady k nalezení interpretace PE :

## (1) morfologická anotace požadované formy FR

|        |       |        |          |     |     |     |       |       |
|--------|-------|--------|----------|-----|-----|-----|-------|-------|
| requis | lemme | catLex | catMorph | cas | num | gen | alt   | pdgm  |
| matce  | matka | SUBST  | N        | dat | sg  | f   | k > c | zn_Re |

## (2) model české morfologie umožňující generování hypotetických forem (FR)

- **deklarativní** část : seznam morfologicky anotovaných koncovek seřazených do skloňovacích paradigmat
- **procedurální** část : pravidla pro realizaci morfonologických alternací kmene v závislosti na určité koncovce

# Skloňovací vzory - paradigma pádových koncovek

- **koncovky** a hodnoty jejich **morfologických kategorií** uspořádané do skloňovacích vzorů (14 základních typů pro substantiva, 101 podvzorů, dále adjektiva, zájmena, číslovky)
- Vymezení paradigmatu nebere ohled na případné **alternace kmene**
- **kmen** substantiva : lemma bez koncové samohlásky (*matk + a*)

## pražská morfologie

R|**zn6e**lkočlkočka|0|0|N|0|0|0

R|**0**hůl|hůl|NFS1@|NNFS1-----@----|N|0|0|(klacek)

R|**0**hole|hůl|NFS2@|NNFS2-----@----|N|0|0|(klacek)

strukturováno v **XML** (použití pro vizualizace na studentské platformě)

# Pravidla pro realizaci alternací

- Implementováno jako ad hoc funkce **Alternuj(Kmen, Koncovka)** obsahující algoritmy pro cca 32 možných alternací
- Ověření realizace alternací a jejich výjimek na 50 000 nejčastějších českých slov (*Frekvenční slovník češtiny*, 2004)

## Vokalické

- 1) kvantitativní
- 2) kvalitativní
- 3) zánik epentetického e
- 4) vložení epentetického e

*jméno* > *jmen*, *houba* > *hub*, *sníh* > *sněhu*, ...  
*vejce* > *vajec*, *přítel* > *přátelé*, *Jan* > *Jene*  
*stromek* > *stromku*, ...  
*cihla* > *cihel*, ...

## Konsonantické

- 5) měkčení r
- 6) palatalizace A
- 7) palatalizace B
- 8) palatalizace souhláskové skupiny
- 9) grafická alternace
- 10) depalatalizace

*doktor* > *doktoři*, *starý* > *staří*, ...  
*kočka* > *kočce*, *moucha* > *mouše*, ....  
*člověk* > *člověče* ....  
*pražský* > *pražští*, ....  
*choť* > *choťem*, *žákyně* > *žákyně*, ...  
*zed'* > *zdmí*

# Jiné (lepší?) řešení pro morfologickou analýzu PE

Náročnost vytváření a údržby morfologického modelu **ad hoc**... 

## Komputačně elegantnější a univerzálnější řešení

- implementovat morfologii pomocí **lexikálního převodníku** složeného z kaskády jednotlivých dílčích převodníků kombinovaných pomocí regulární operace **kompozice** : *morfotaktický filtr, slovník morfémů a kontextová pravidla pro alternace* (Karttunen & Beesley 2003)
- 
- použít k analýze sérii **neúplných převodníků** s uvolněnými omezeními v morfotaktickém filtru (pro nesprávné kombinace kmene a koncovky) a bez zapojení alternačních pravidel (pro formy s nesprávně provedenou alternací)

# Morfologická interpretace PE a typ chyby

- v případě, že se určitá **hypotetická forma** shoduje s **PE**, pak jsou její morfologické atributy považovány za morfologickou **interpretaci PE**
- **typ** chyby je určen podle **rozdílu** existujícího mezi anotací požadované formy (FR) a morfologickou interpretací **PE**
- pouze **morfologie** (popis chybné formy), nikoliv **syntax** (interpretace funkce chybné formy)

*Aničce bolí hlava.                    (La tête fait mal à Anička)*

*Aničkou ...*

*Aničky ...*

*...*

# Morfologická interpretace PE a typ chyby

- určení typu chyby podle **zasaženého atributu**:

| <i>matku</i>        | ≠ | <i>matky</i>        |
|---------------------|---|---------------------|
| <i>cas : acc</i>    |   | <i>cas : gen</i>    |
| <i>num : sg</i>     |   | <i>num : sg</i>     |
| <i>gen : f</i>      |   | <i>gen : f</i>      |
| <i>pdgm : zn_Re</i> |   | <i>pdgm : zn_Re</i> |
| <i>alt : sans</i>   |   | <i>alt : sans</i>   |
| <i>dia : 0</i>      |   | <i>dia : 0</i>      |

| atribut | typ chyby                  |
|---------|----------------------------|
| cas     | chyba v pádě               |
| num     | chyba v čísle              |
| gen     | chyba v rodě               |
| alt     | chyba v alternaci          |
| pdgm    | chyba ve skloňovacím vzoru |
| dia     | chyba v diakritice         |

Interpretace

1

# Morfologická interpretace PE a typ chyby

- určení typu chyby podle **zasaženého atributu**:

|                     |   |                     |
|---------------------|---|---------------------|
| <i>matku</i>        | ≠ | <i>matky</i>        |
| <i>cas : acc</i>    |   | <i>cas : nom</i>    |
| <i>num : sg</i>     |   | <i>num : pl</i>     |
| <i>gen : f</i>      |   | <i>gen : f</i>      |
| <i>pdgm : zn_Re</i> |   | <i>pdgm : zn_Re</i> |
| <i>alt : sans</i>   |   | <i>alt : sans</i>   |
| <i>dia : 0</i>      |   | <i>dia : 0</i>      |

| atribut | typ chyby                  |
|---------|----------------------------|
| cas     | chyba v pádě               |
| num     | chyba v čísle              |
| gen     | chyba v rodě               |
| alt     | chyba v alternaci          |
| pdgm    | chyba ve skloňovacím vzoru |
| dia     | chyba v diakritice         |

Interpretace

2



ambiguita koncovek, několik možných interpretací

# Generace hypotetických forem

- 1 **korektně provedená alternace:** *matce, hole, jmen*
- 2 **pouhá konkatence kmene a koncovky:** *matke, hůle, jmén*
- 3 **nekorektně provedená alternace:** *kolu (kůlu), Isu (lesu), překladatlovi*

Hypotetická (HF) a chybná forma (PE) jsou před srovnáním zbaveny **diakritiky** a **kapitalizace**. Případné rozdíly jsou zachyceny v interpretaci.

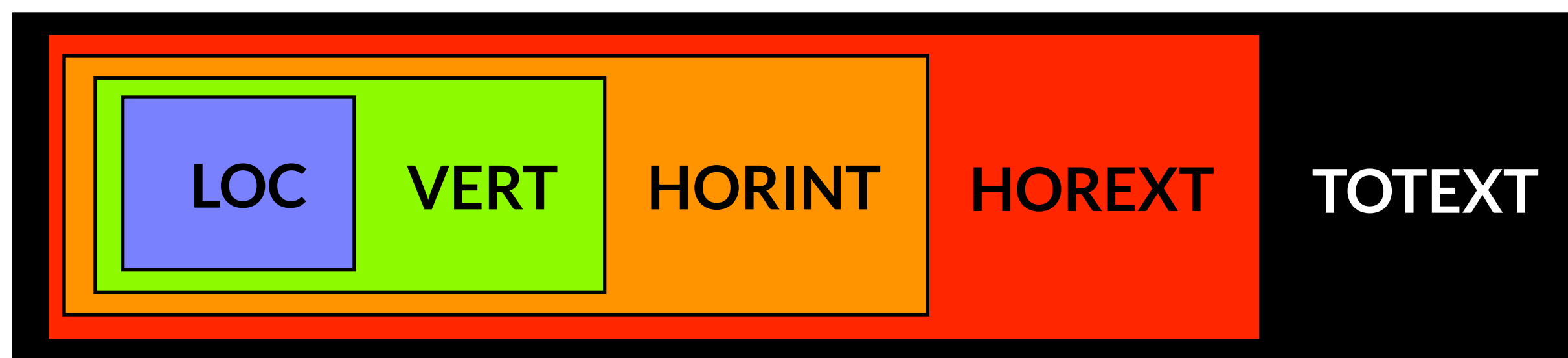
Např. *růži@růží* bude mít dvě interpretace: chyba v pádě a chyba v diakritice

# Výběr koncovek pro generaci hypotetických forem (FH)

- omezit počet koncovek pro generování hypotetických forem
  - ▶ Bez omezení by pro generaci FH bylo nutné vytvořit minimálně **1750 forem** pro každou FR (7 pádů x 2 čísla x 125 deklinačních vzorů + alternace a diakritika)

## Dva typy omezení:

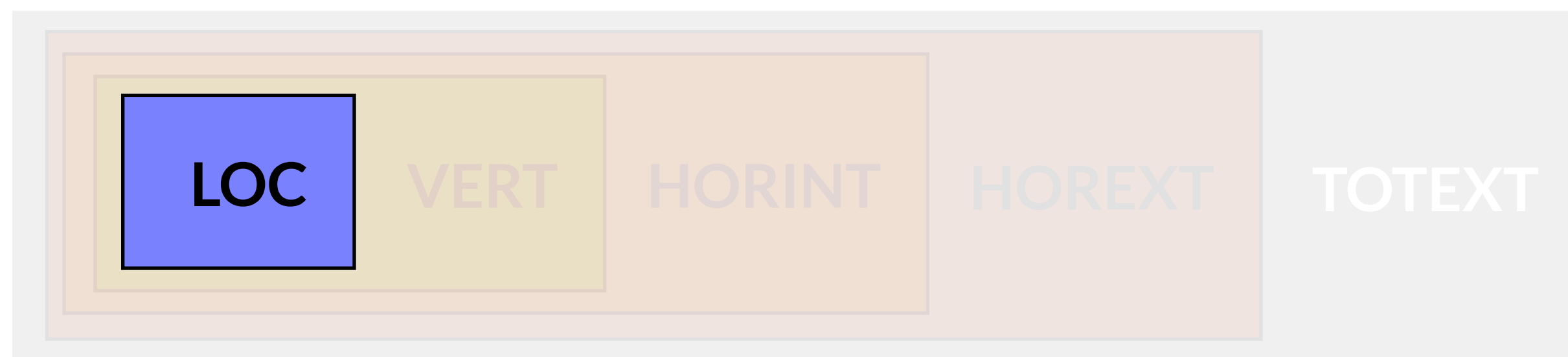
- (1) použít **pouze koncovky**, které lze izolovat na pravé periférii PE
- (2) kontrolovat **rozdíl** hodnot gramatických kategorií FH vůči FR



- ▶ lokální
- ▶ vertikální
- ▶ horizontální interní
- ▶ horizontální externí
- ▶ externí úplné

# Výběr koncovek pro generaci hypotetických forem (FH)

## • LOKÁLNÍ INTERPRETACE



domu@důmu[ů>o@sans]

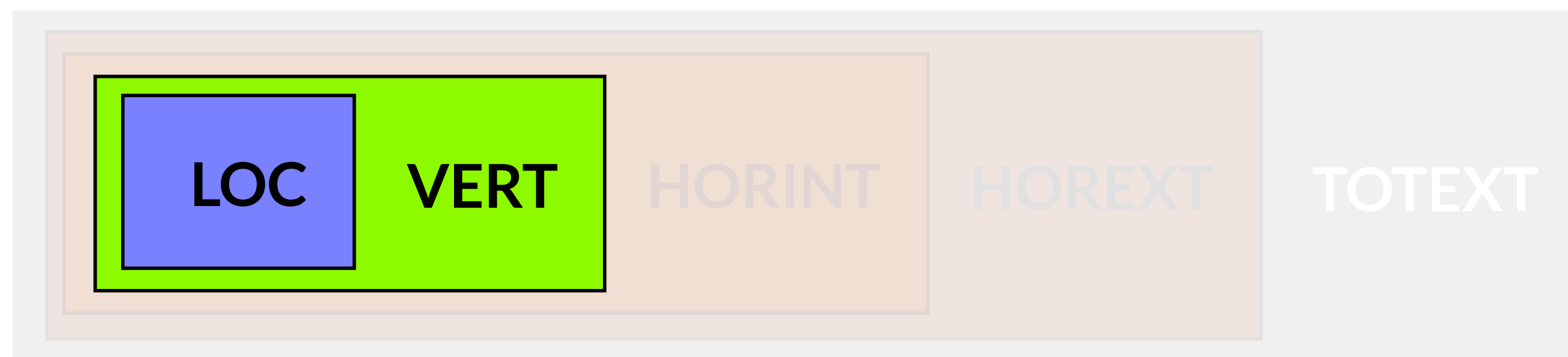
Atributy zasažené v nově rozpoznaných formách pro daný rozsah generace FH\*:

| cas                                              | num   | gen   | pdgm  | alt      | dia      |
|--------------------------------------------------|-------|-------|-------|----------|----------|
| shoda<br>možný rozdíl ve<br>variantě<br>koncovky | shoda | shoda | shoda | korektní | korektní |
|                                                  |       |       |       | chybná   | chybná   |

*\*) Rozdíl v kapitalizaci je z přehledu vynechán, protože se nejedná o morfologický problém. Diagnostika se ním ovšem pracuje.*

# Výběr koncovek pro generaci hypotetických forem (FH)

## • VERTIKÁLNÍ INTERPRETACE



turistu@turista[acc@nom]

Atributy zasažené v nově rozpoznaných formách pro daný rozsah generace FH:

| cas    | num    | gen                        | pdgm  | alt      | dia      |
|--------|--------|----------------------------|-------|----------|----------|
| shoda  | shoda  | shoda                      | shoda | korektní | korektní |
| rozdíl | rozdíl | rozdíl<br>pouze adj / pron |       | chybná   | chybná   |

# Výběr koncovek pro generaci hypotetických forem (FH)

- HORIZONTALNÍ INTERNÍ INTERPRETACE**



jablcích@jablkech[mt\_s@mt]

Atributy zasažené v nově rozpoznaných formách pro daný rozsah generace FH:

| cas   | num   | gen   | pdgm                                                           | alt      | dia      |
|-------|-------|-------|----------------------------------------------------------------|----------|----------|
| shoda | shoda | shoda | rozdí<br>pouze v rámci<br>modelového<br>vzoru a<br>podvzorů FR | korektní | korektní |
|       |       |       |                                                                | chybná   | chybná   |

# Výběr koncovek pro generaci hypotetických forem (FH)

- HORIZONTALNÍ INTERNÍ INTERPRETACE - bez omezení**



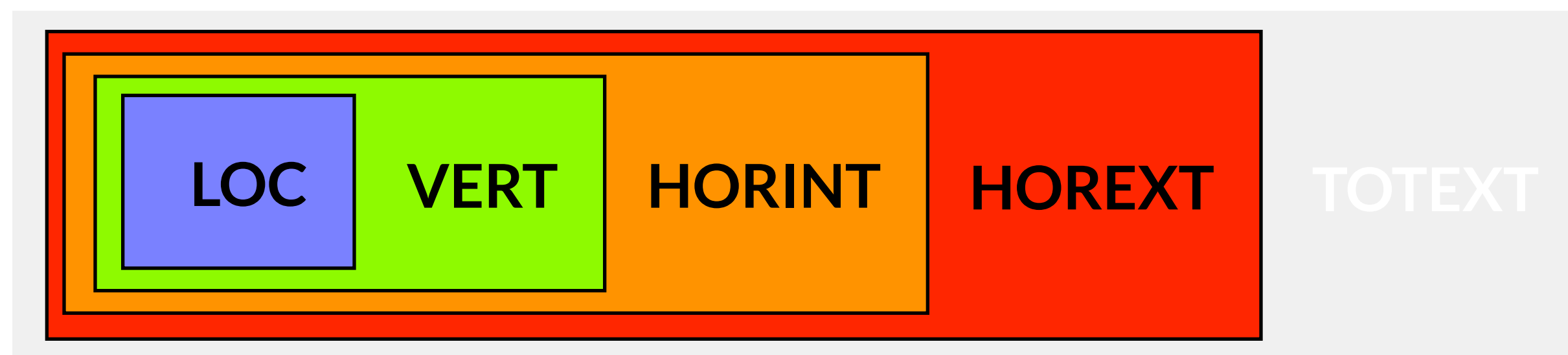
tatínka@tatínce[pn\_h@pn l  
acc@voc]

Atributy zasažené v nově rozpoznaných formách pro daný rozsah generace FH:

| cas    | num    | gen                        | pdgm                                                            | alt      | dia      |
|--------|--------|----------------------------|-----------------------------------------------------------------|----------|----------|
| shoda  | shoda  | shoda                      | rozdíl<br>pouze v rámci<br>modelového<br>vzoru a<br>podvzorů FR | korektní | korektní |
| rozdíl | rozdíl | rozdíl<br>pouze adj / pron |                                                                 | chybná   | chybná   |

# Výběr koncovek pro generaci hypotetických forem (FH)

- HORIZONTALNÍ EXTERNÍ INTERPRETACE**



zvířat@zvířů[kr@scIn@m]

Atributy zasažené v nově rozpoznaných formách pro daný rozsah generace FH:

| cas   | num   | gen    | pdgm                                              | alt      | dia      |
|-------|-------|--------|---------------------------------------------------|----------|----------|
| shoda | shoda | shoda  | rozdíl                                            | korektní | korektní |
|       |       | rozdíl | pouze vzory ze<br>stejným<br>zkončením<br>jako FR | chybná   | chybná   |

# Výběr koncovek pro generaci hypotetických forem (FH)

- HORIZONTALNÍ EXTERNÍ INTERPRETACE - bez omezení**



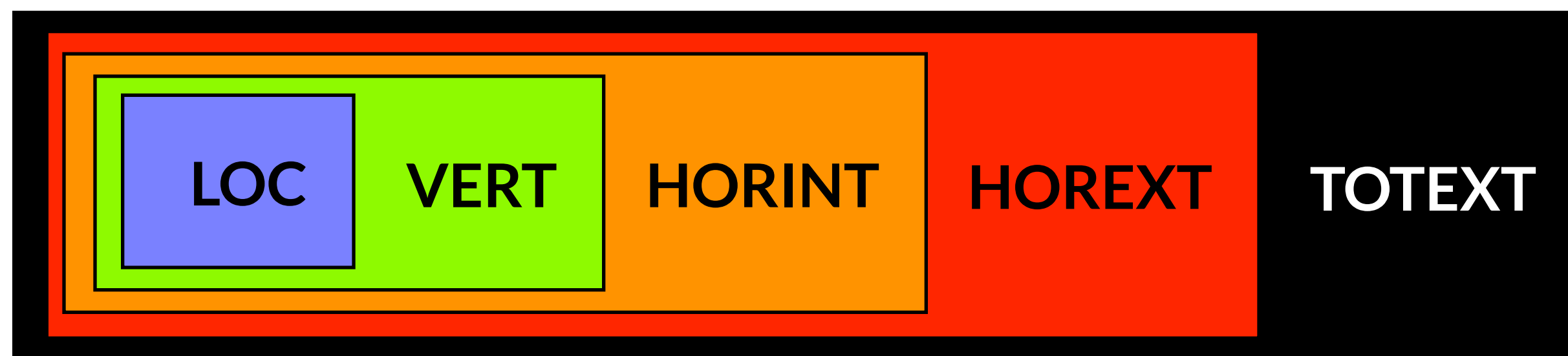
růži@růžu[rz@sc|acc@gen|  
sg@pl|f@ml|dia]

Atributy zasažené v nově rozpoznaných formách pro daný rozsah generace FH:

| cas    | num    | gen    | pdgm                                                        | alt      | dia      |
|--------|--------|--------|-------------------------------------------------------------|----------|----------|
| shoda  | shoda  | shoda  | rozdíl<br>pouze vzory ze<br>stejným<br>zkončením<br>jako FR | korektní | korektní |
| rozdíl | rozdíl | rozdíl |                                                             | chybná   | chybná   |

# Výběr koncovek pro generaci hypotetických forem (FH)

- HORIZONTALNÍ EXTERNÍ INTERPRETACE**



dopis@dopisa[hd\_RE@pnli@m]

Atributy zasažené v nově rozpoznaných formách pro daný rozsah generace FH:

| cas   | num   | gen    | pdgm                                                         | alt      | dia      |
|-------|-------|--------|--------------------------------------------------------------|----------|----------|
| shoda | shoda | shoda  | rozdíl<br>všechny vzory<br>v rámci<br>morfologického<br>typu | korektní | korektní |
|       |       | rozdíl |                                                              | chybná   | chybná   |

# Výběr koncovek pro generaci hypotetických forem (FH)

- HORIZONTÁLNÍ EXTERNÍ INTERPRETACE - bez omezení

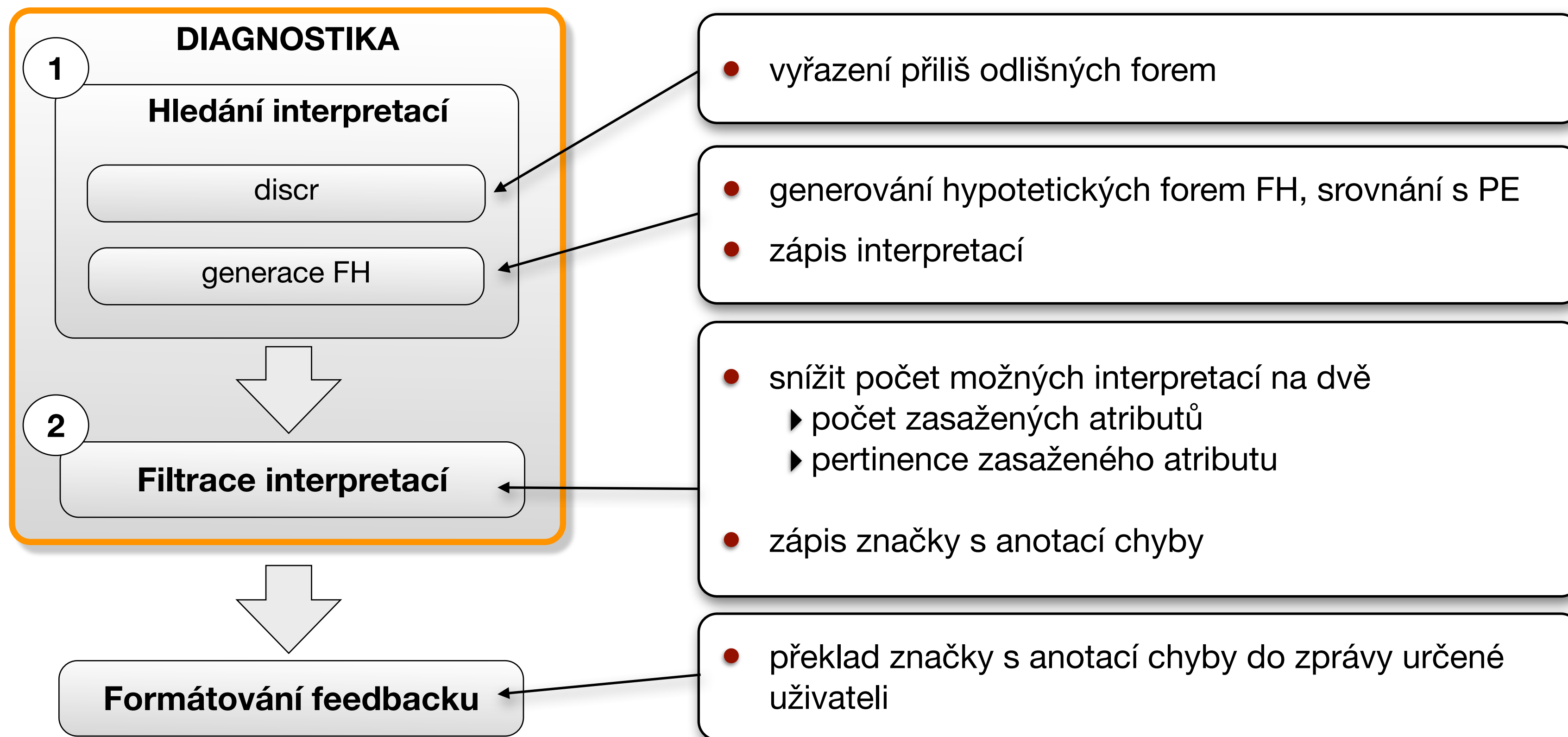


mrkve@mrkevové

Atributy zasažené v nově rozpoznaných formách pro daný rozsah generace FH:

| cas    | num    | gen    | pdgm                                                         | alt      | dia      |
|--------|--------|--------|--------------------------------------------------------------|----------|----------|
| shoda  | shoda  | shoda  | rozdíl<br>všechny vzory<br>v rámci<br>morfologického<br>typu | korektní | korektní |
| rozdíl | rozdíl | rozdíl |                                                              | chybná   | chybná   |

# Průběh diagnostiky chyb



# Vyřazené PE - DISCR

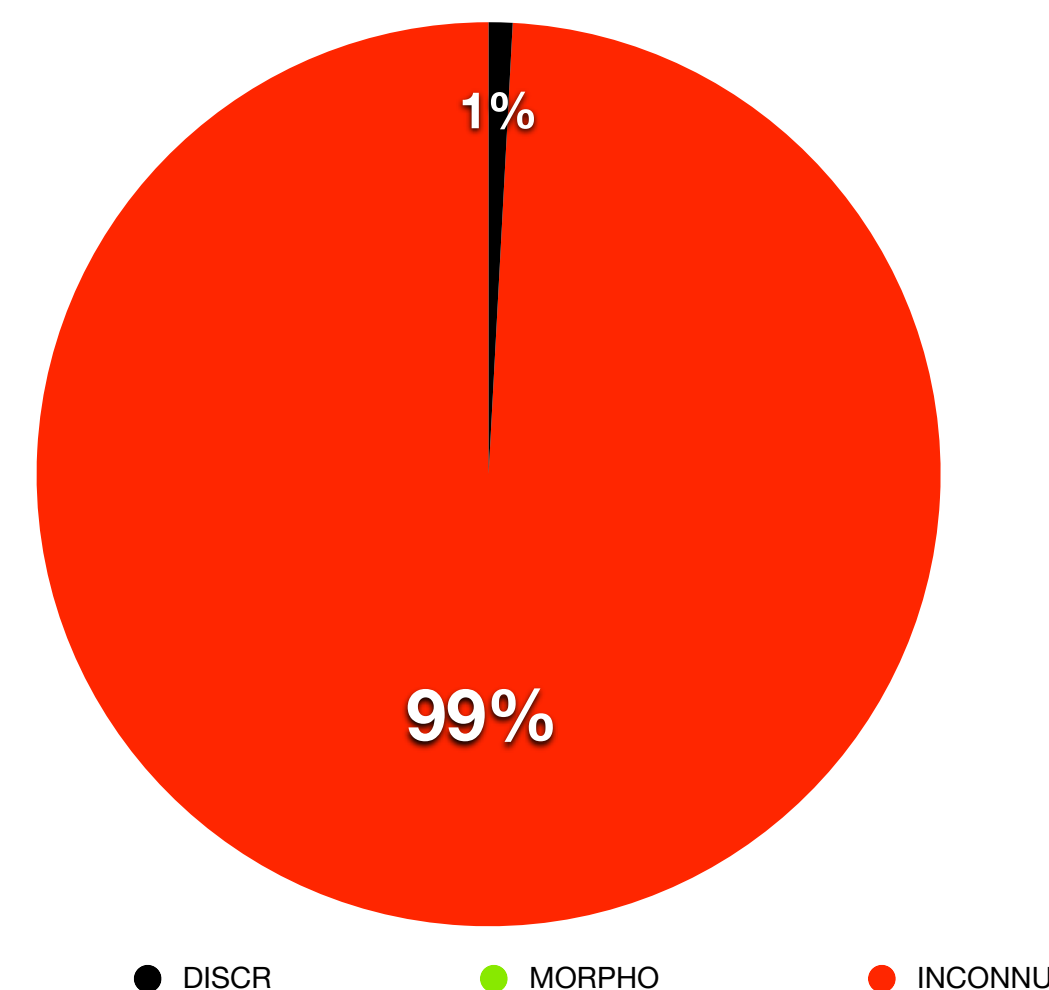
## zpracování před zahájením morfologické analýzy

- Levenshtein > 5: `psa@pesmrkev` (LEV)  
`hranolkami@ihoiug` (LEV)
- příliš krátké: `krav@kr` (COURT)
- příliš dlouhé: `turistů@turistsedadel` (LONG)
- jediný znak: `stromy@h` (LONG)

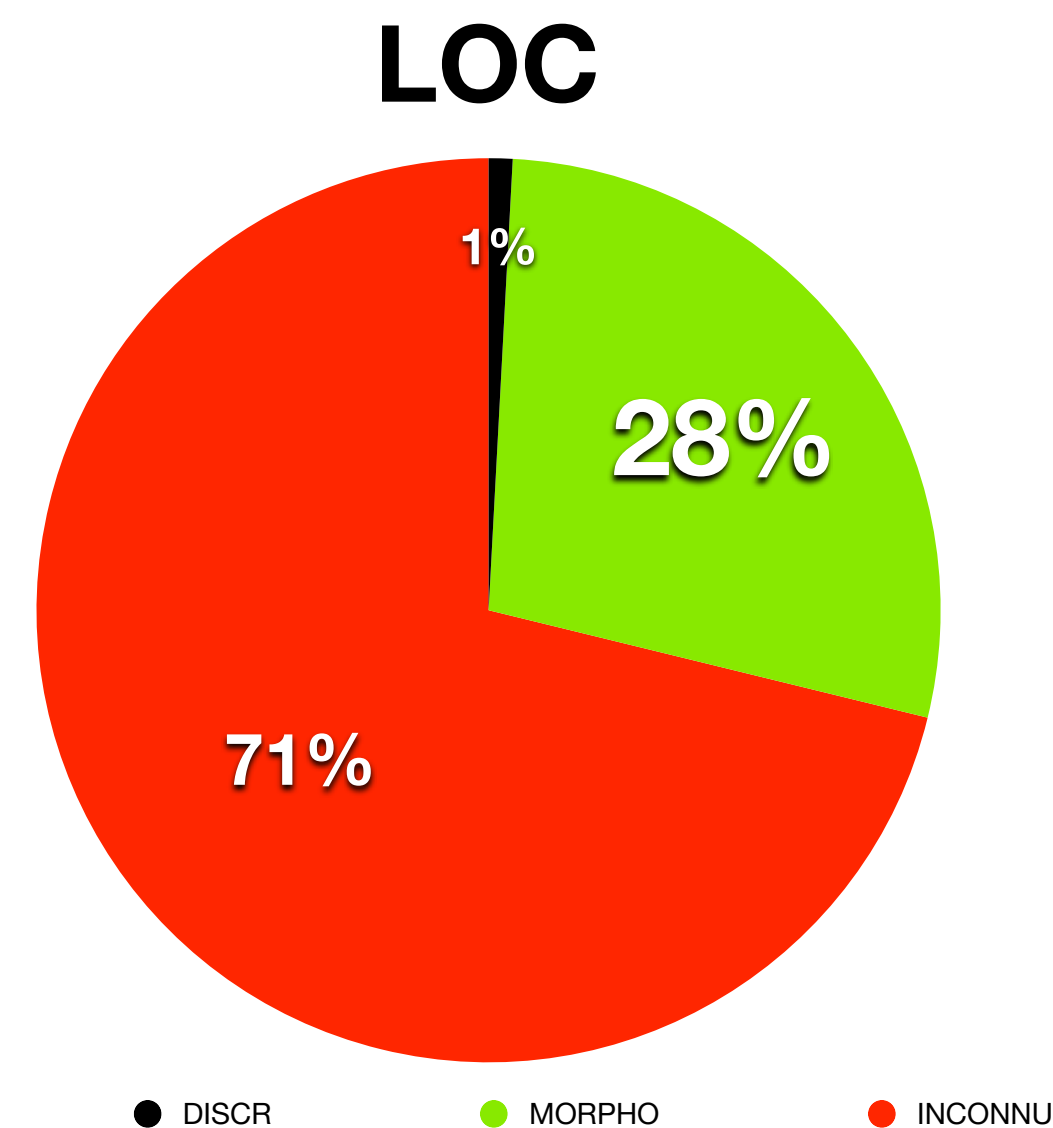
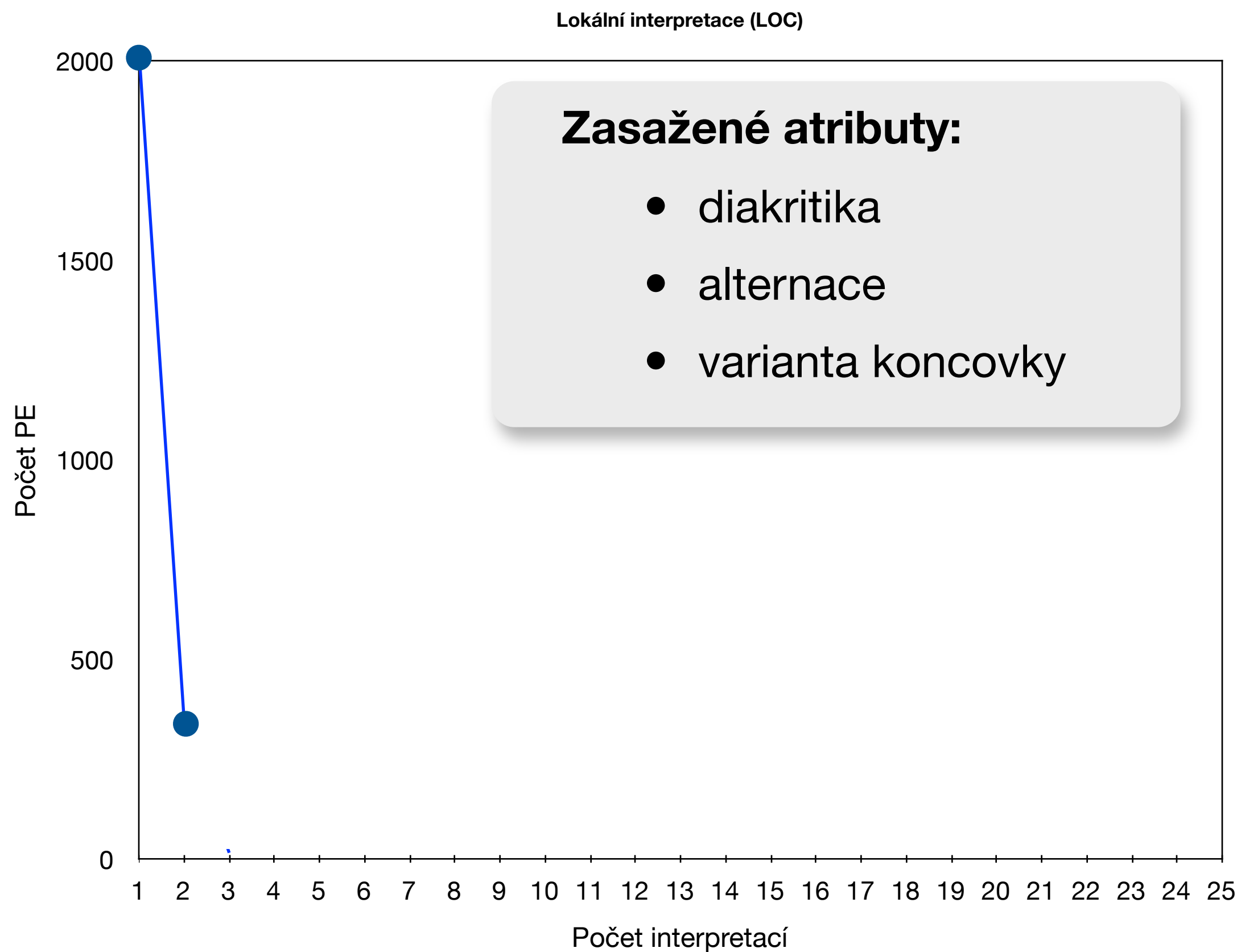


pouze **73** forem (0,83 %)

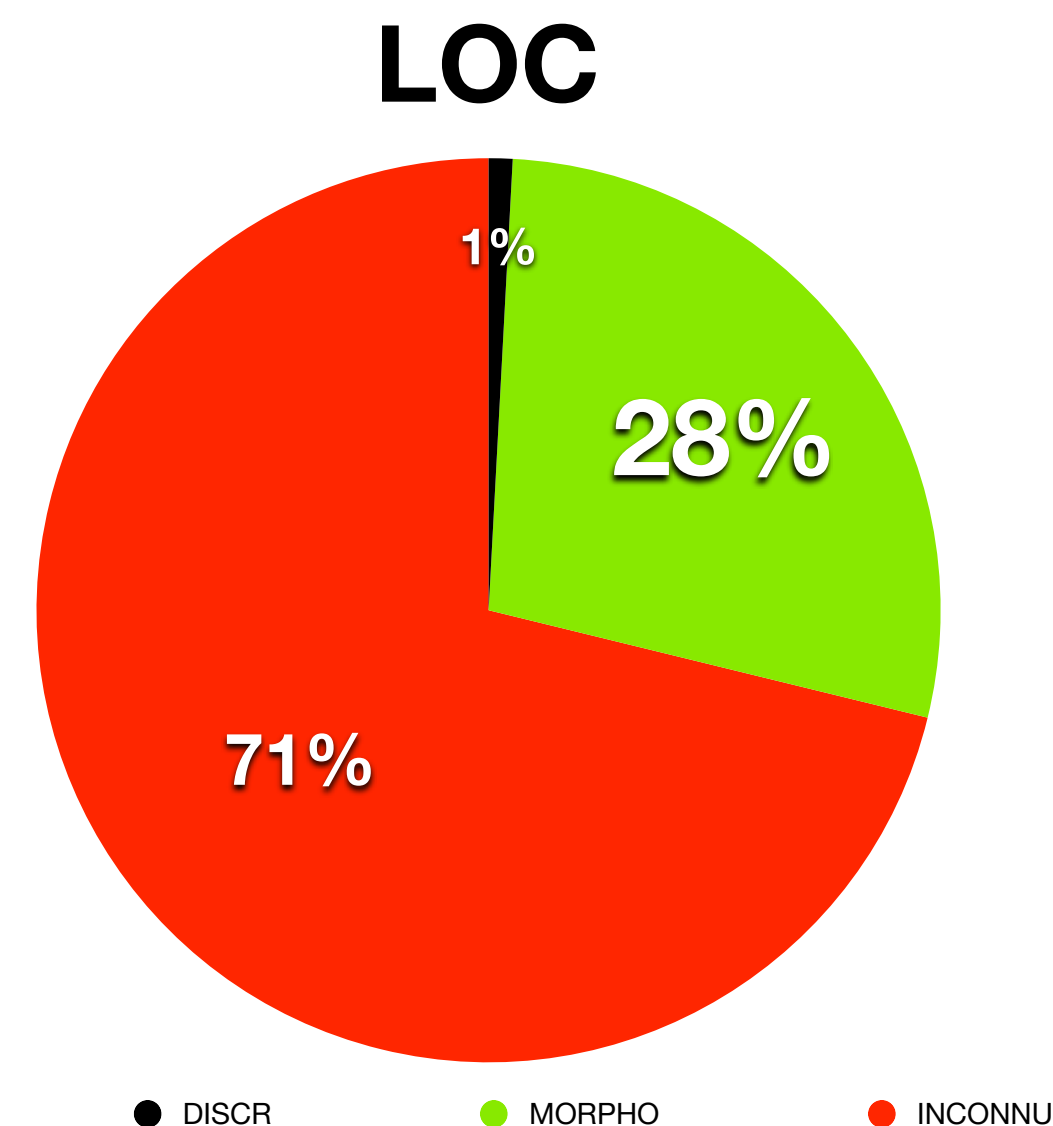
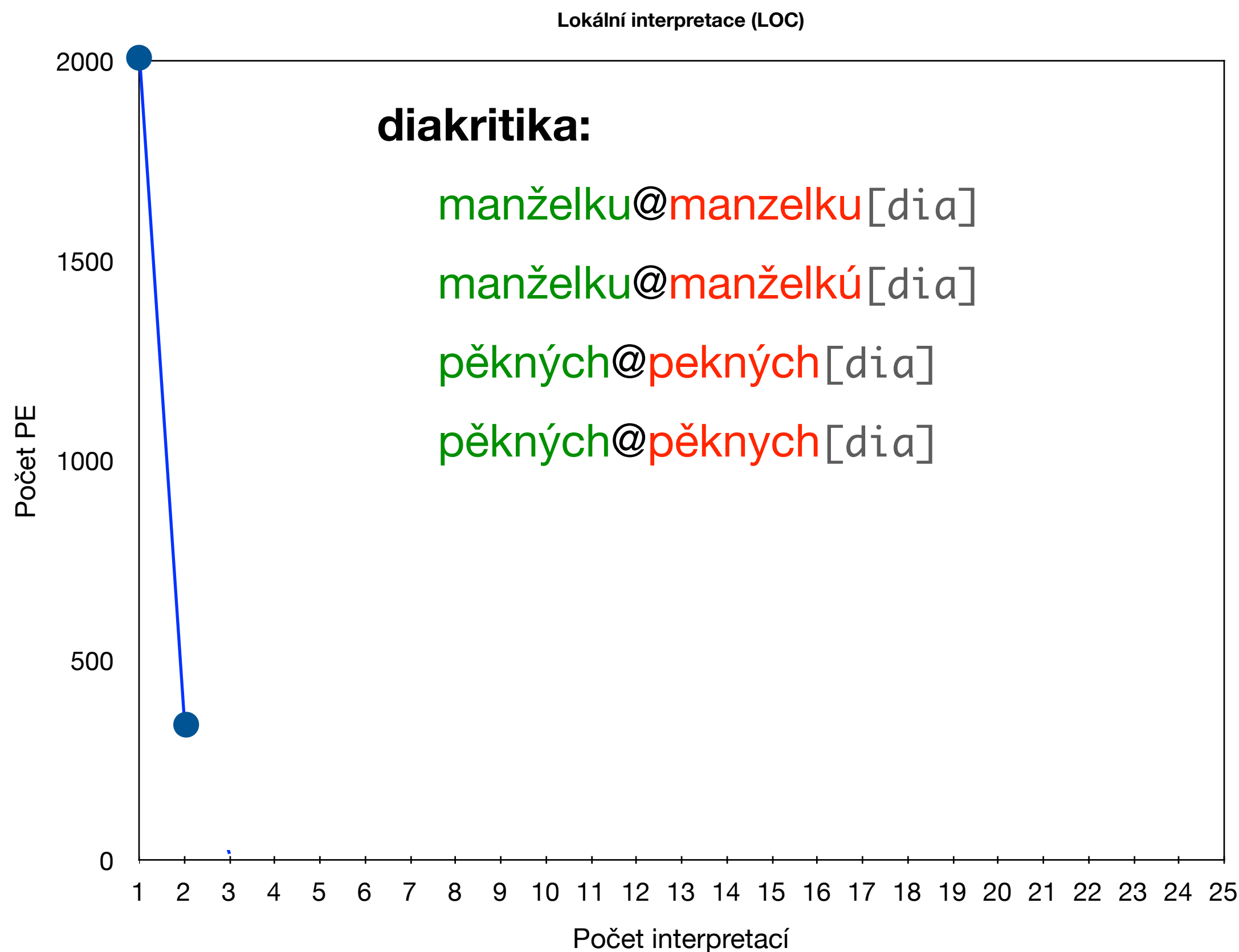
## DISCR



# Lokální interpretace



# Lokální interpretace



# Lokální interpretace

Lokální interpretace (LOC)

**alternace:**

domu@důmu[ů>o@sans]

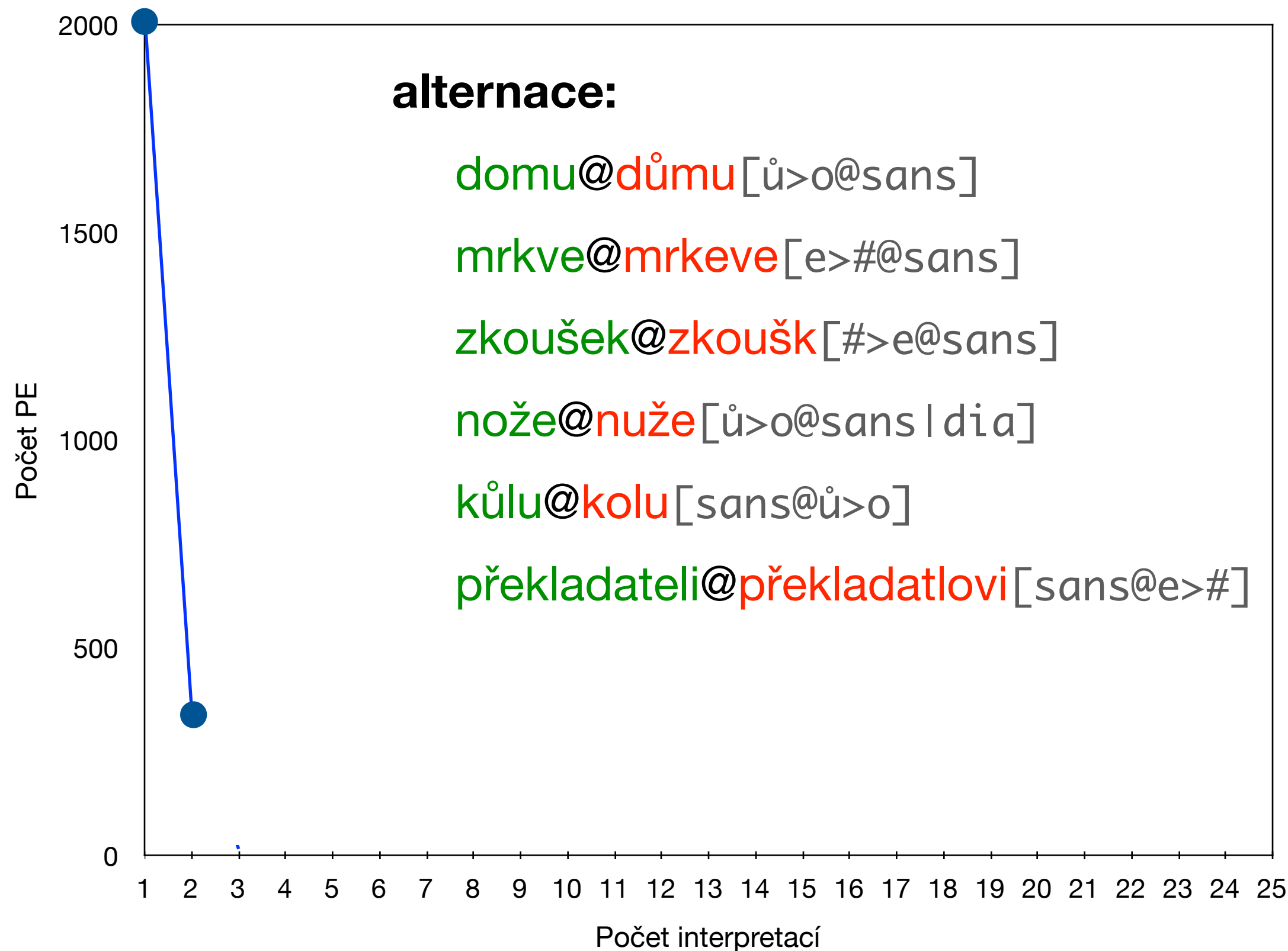
mrkve@mrkeve[e>#@sans]

zkoušek@zkoušk[#>e@sans]

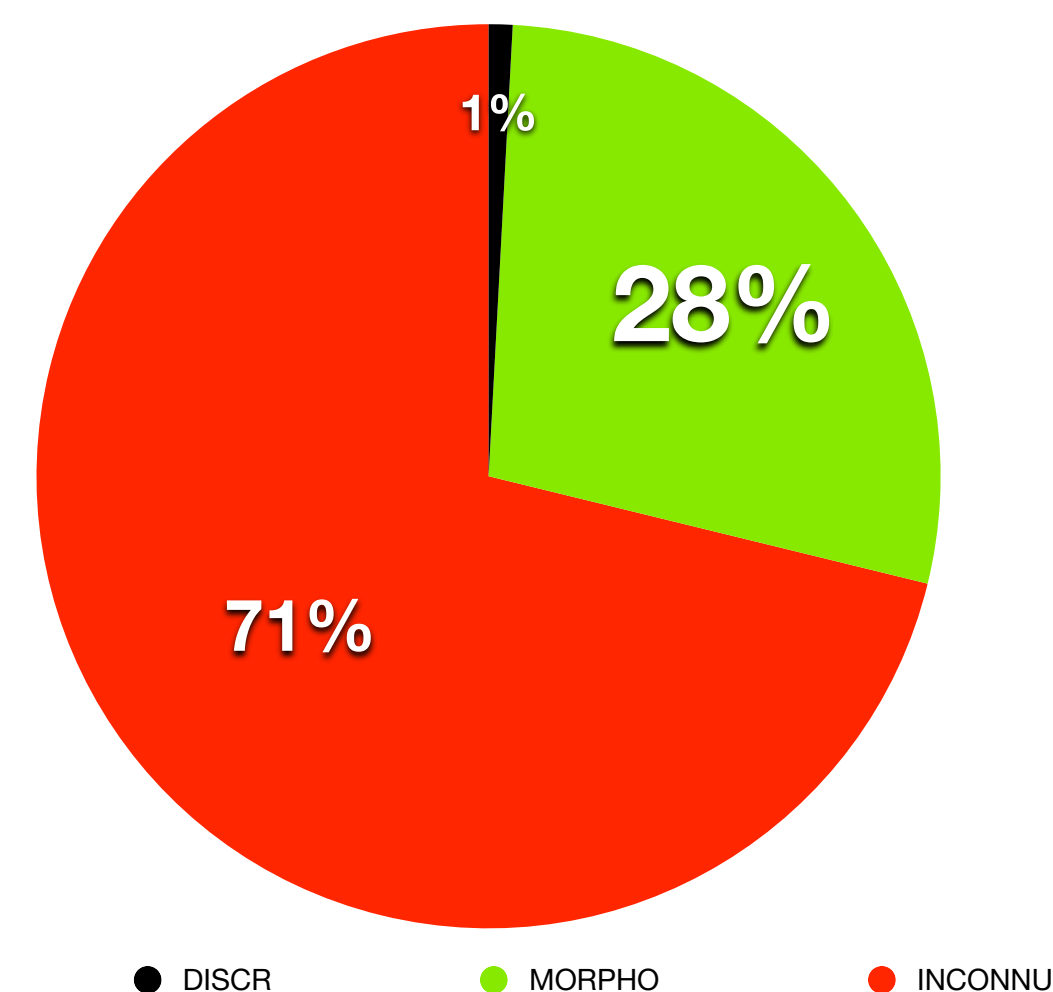
nože@nuže[ů>o@sans|dia]

kůlu@kolu[sans@ů>o]

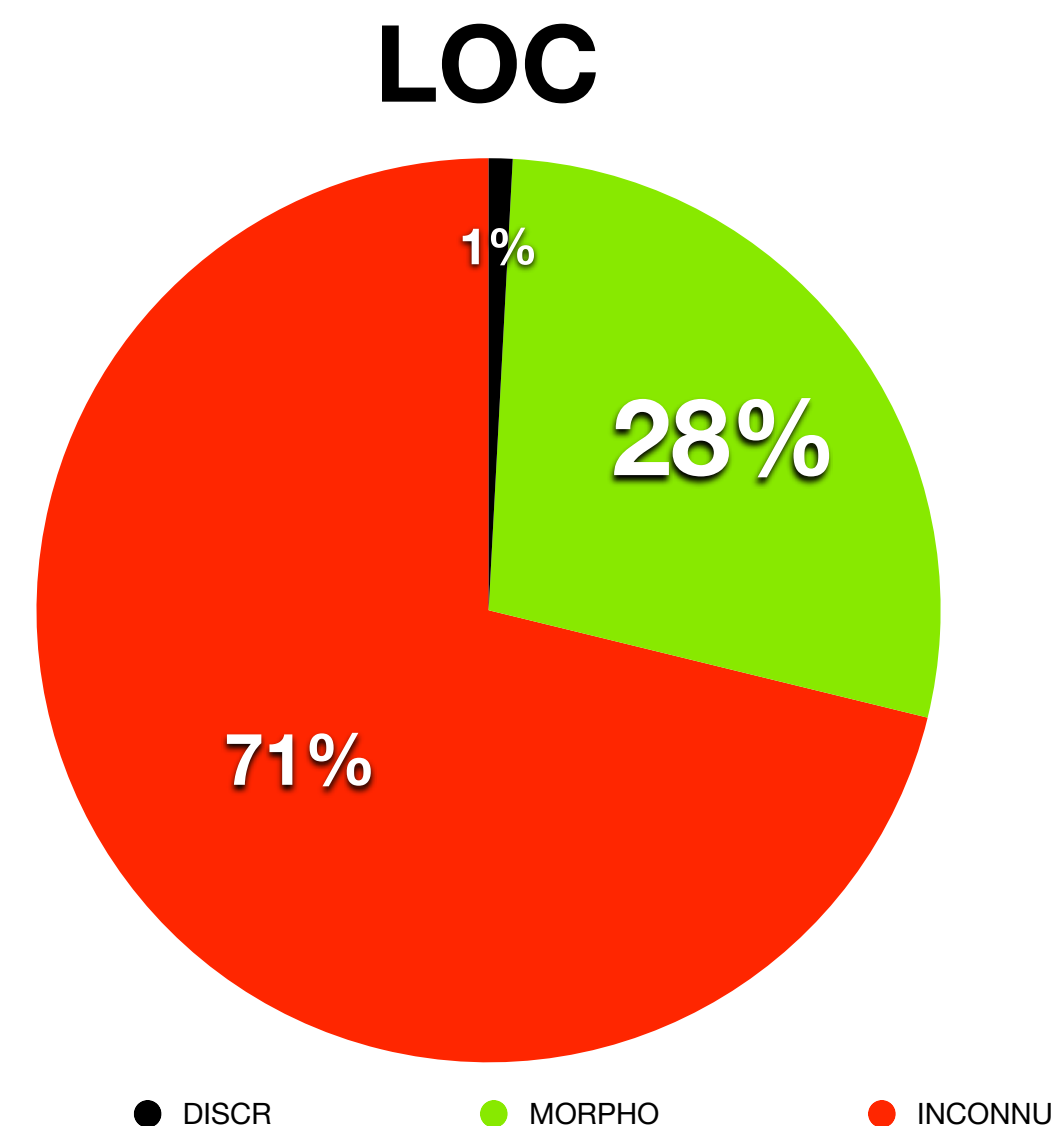
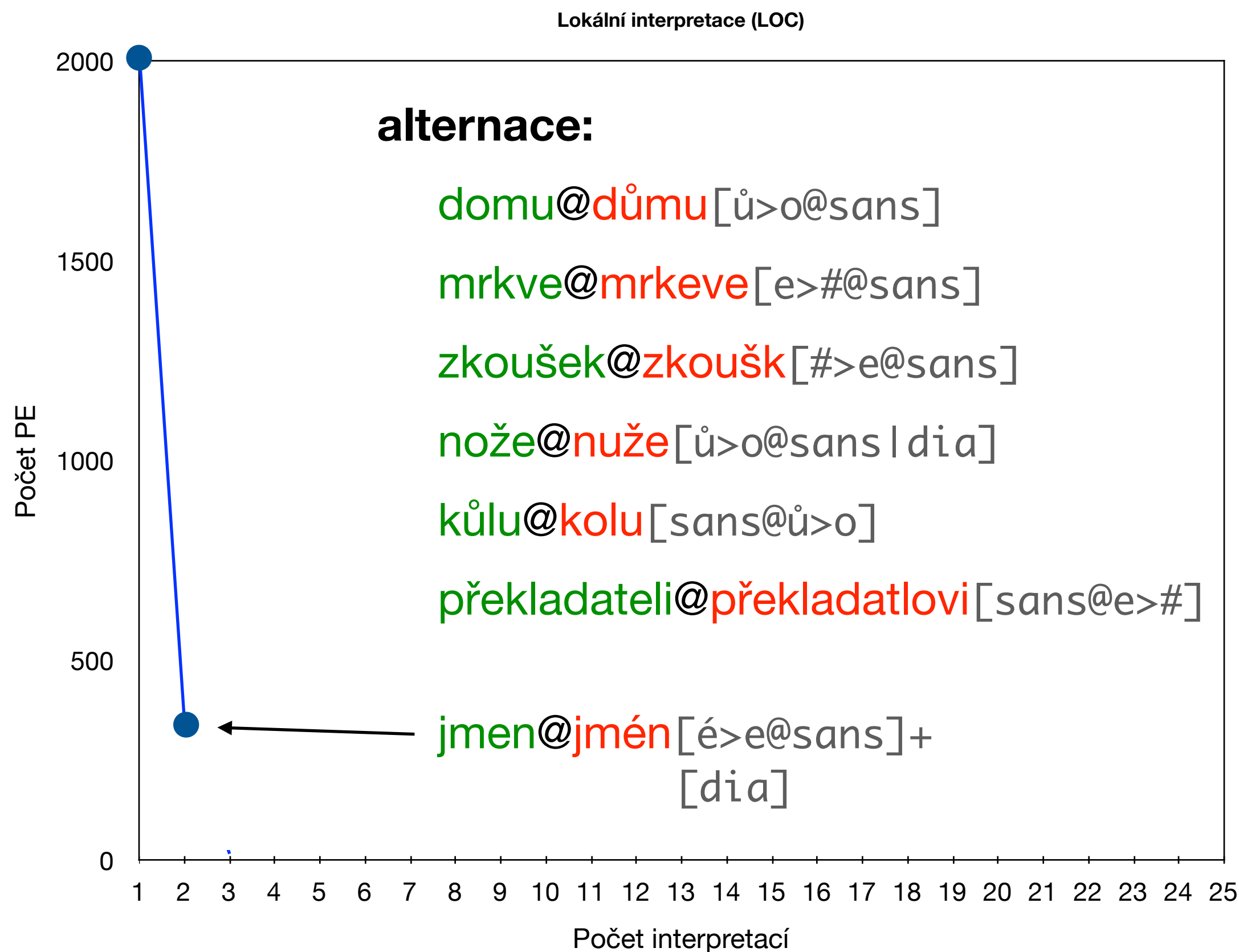
překladařeli@překladařlovi[sans@e>#]



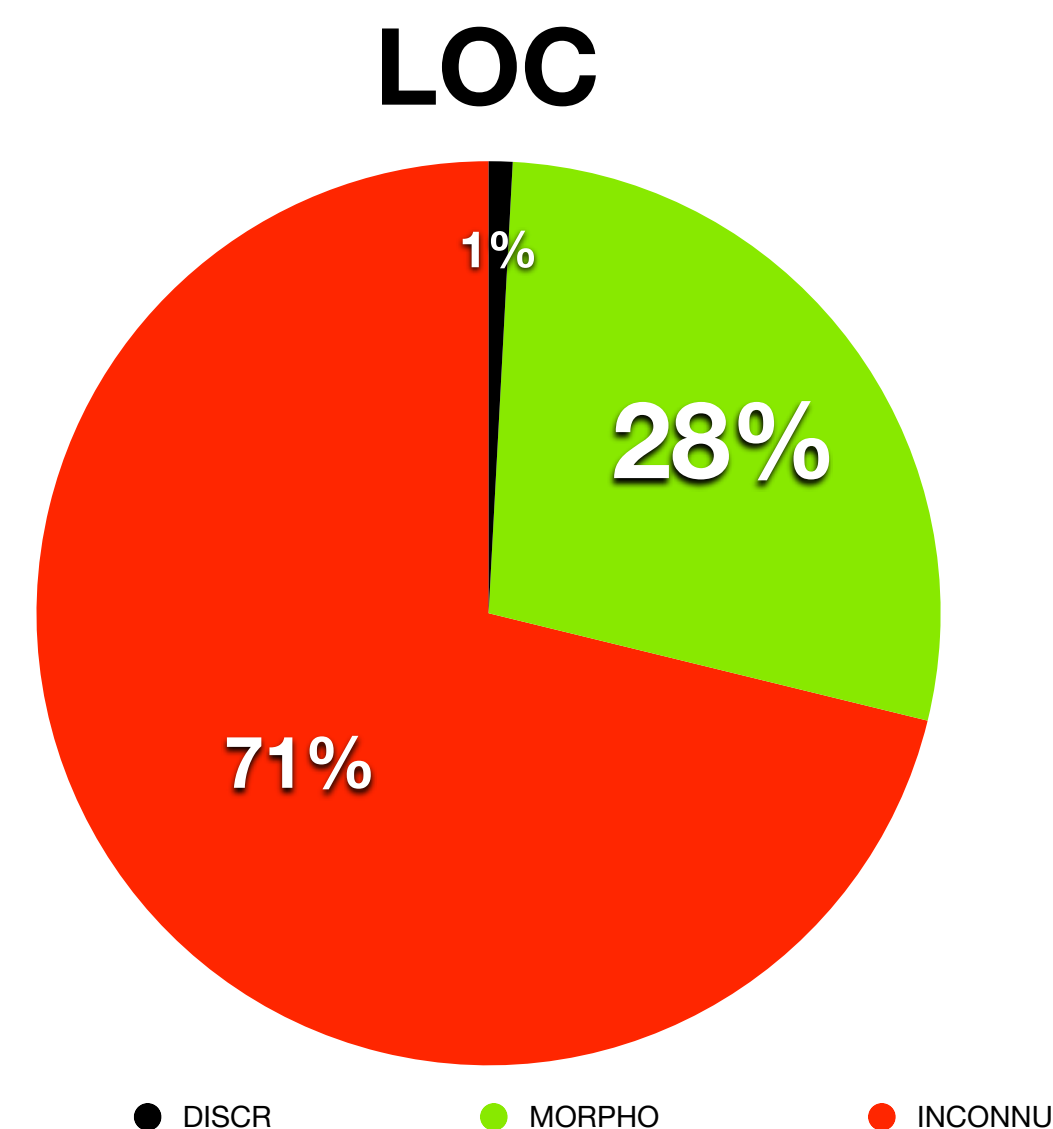
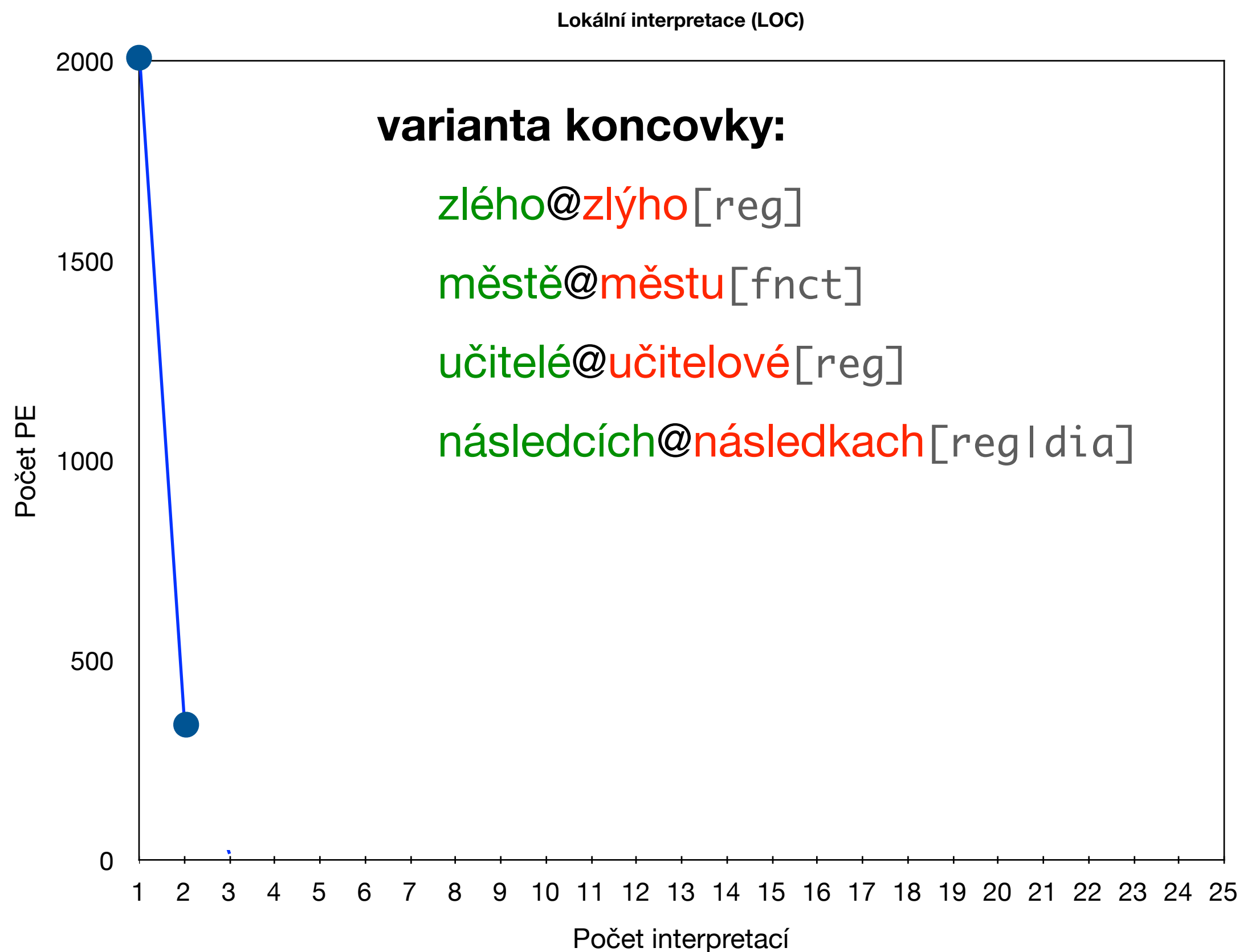
## LOC



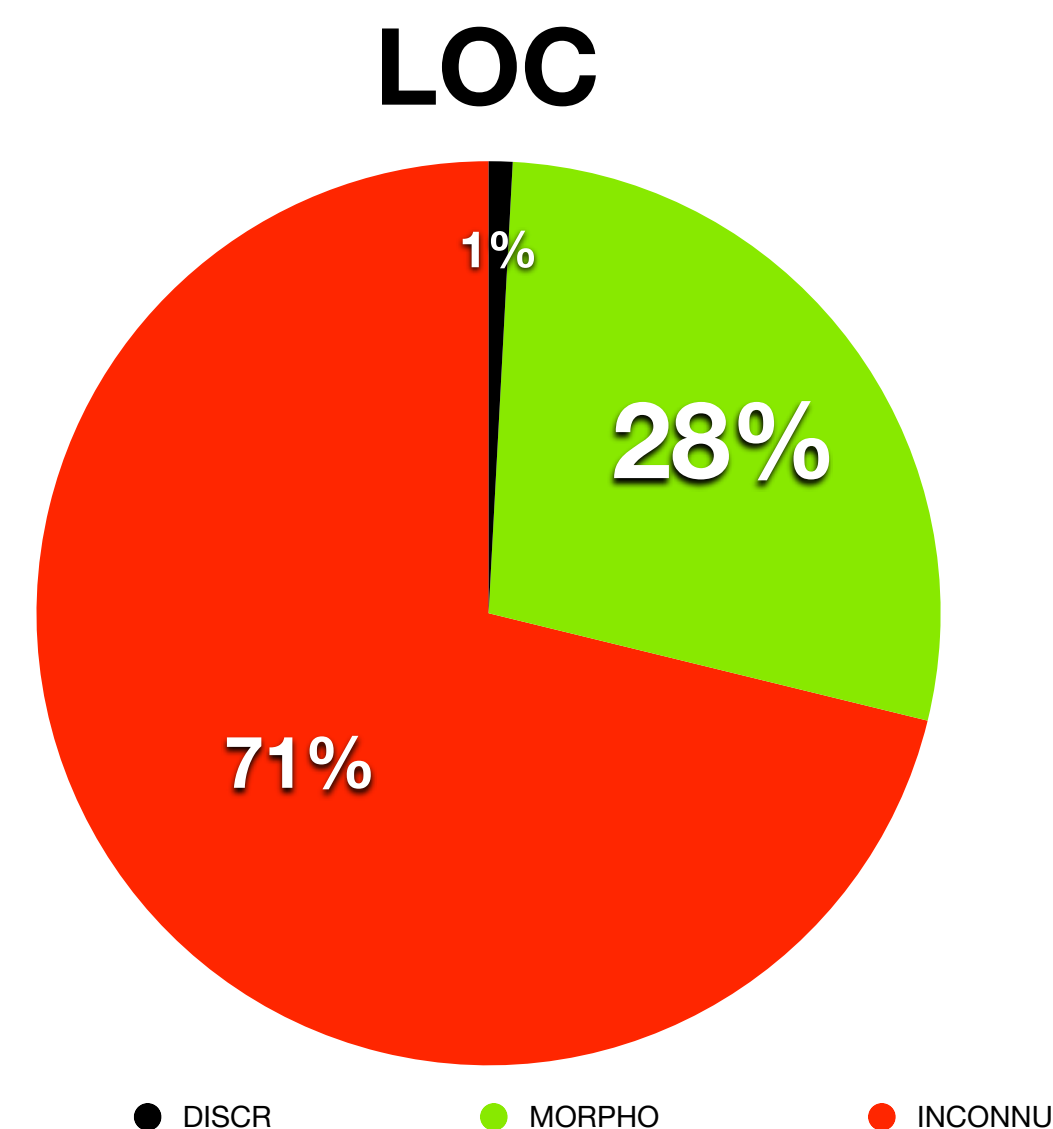
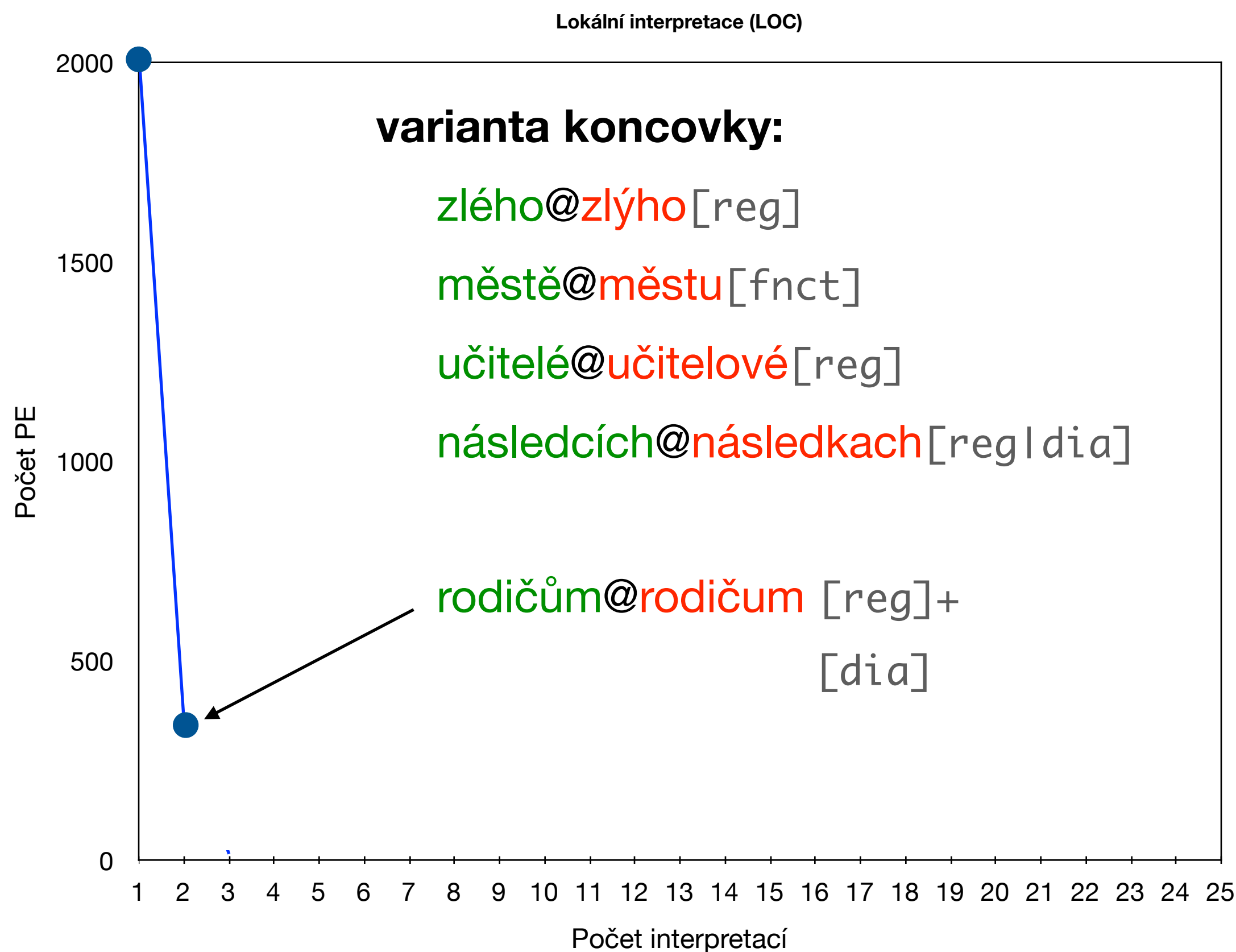
# Lokální interpretace



# Lokální interpretace

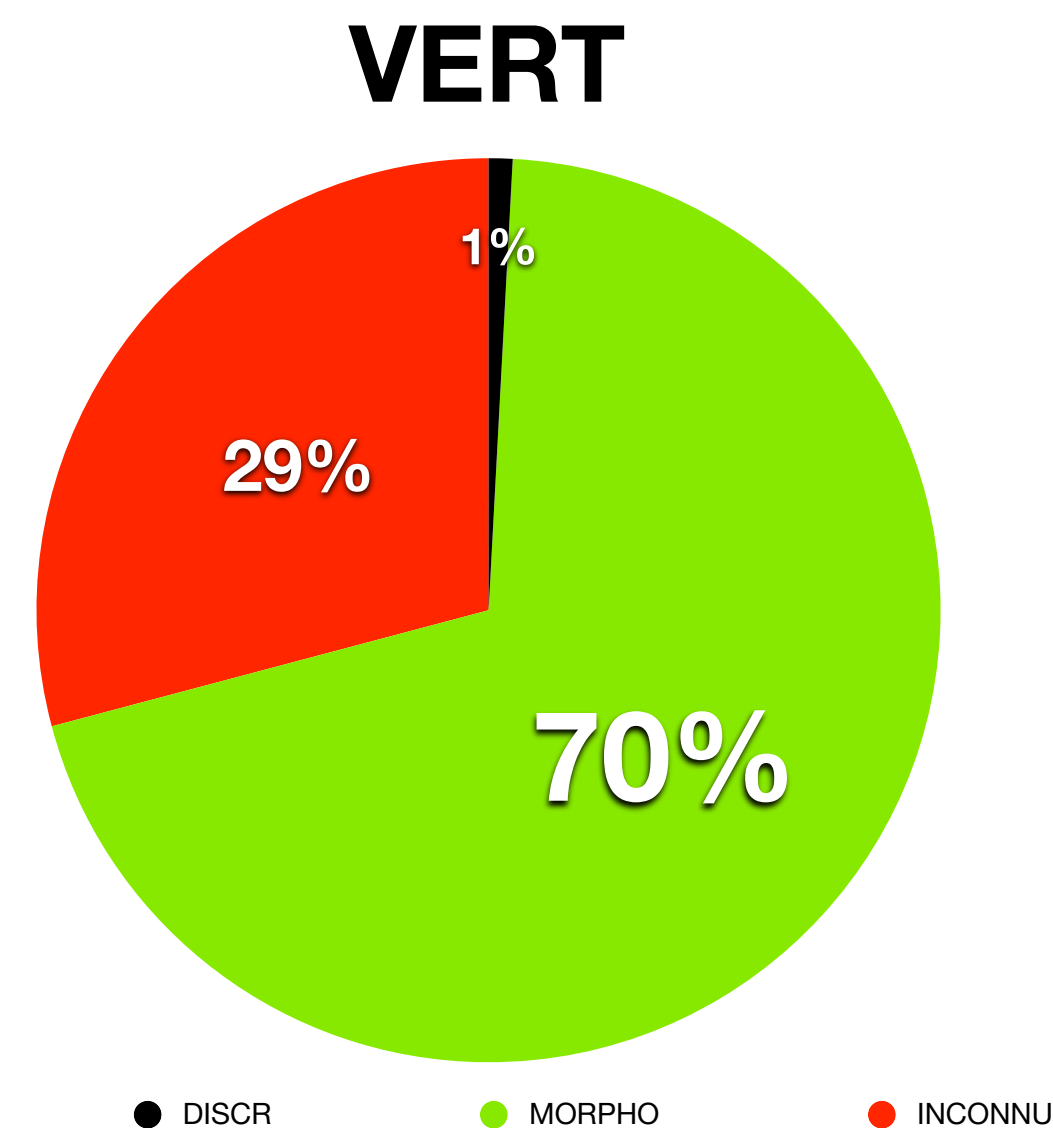
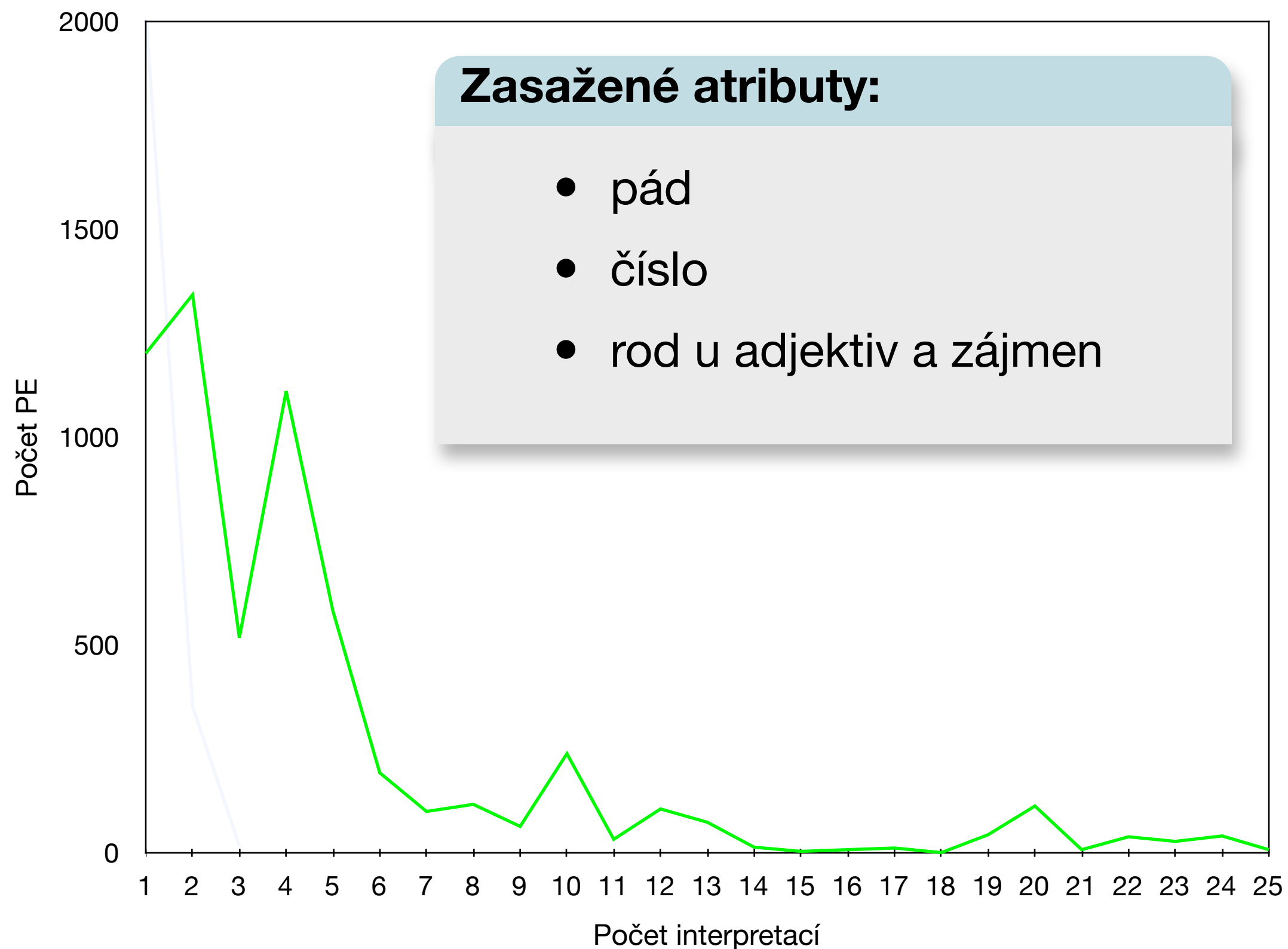


# Lokální interpretace



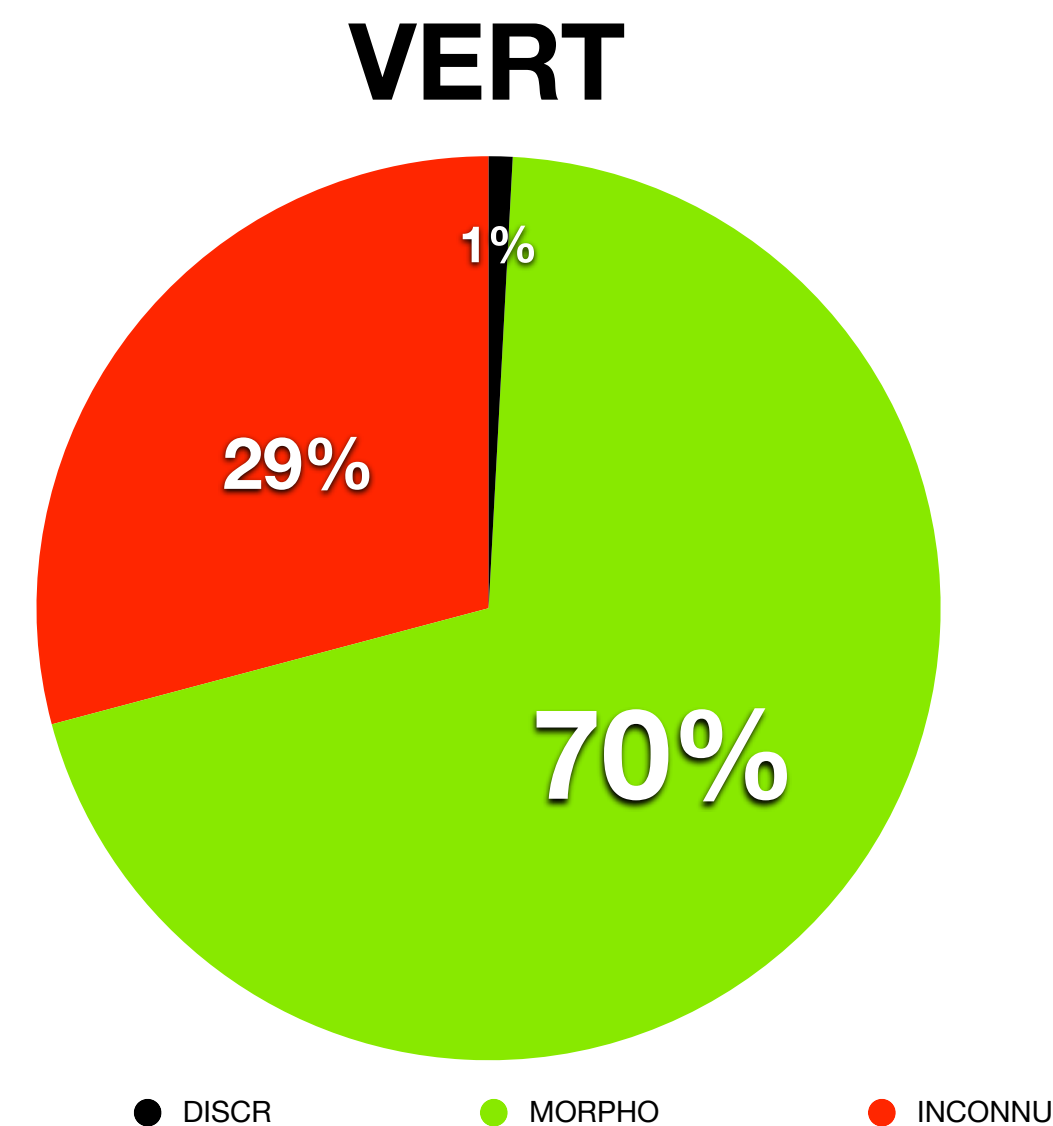
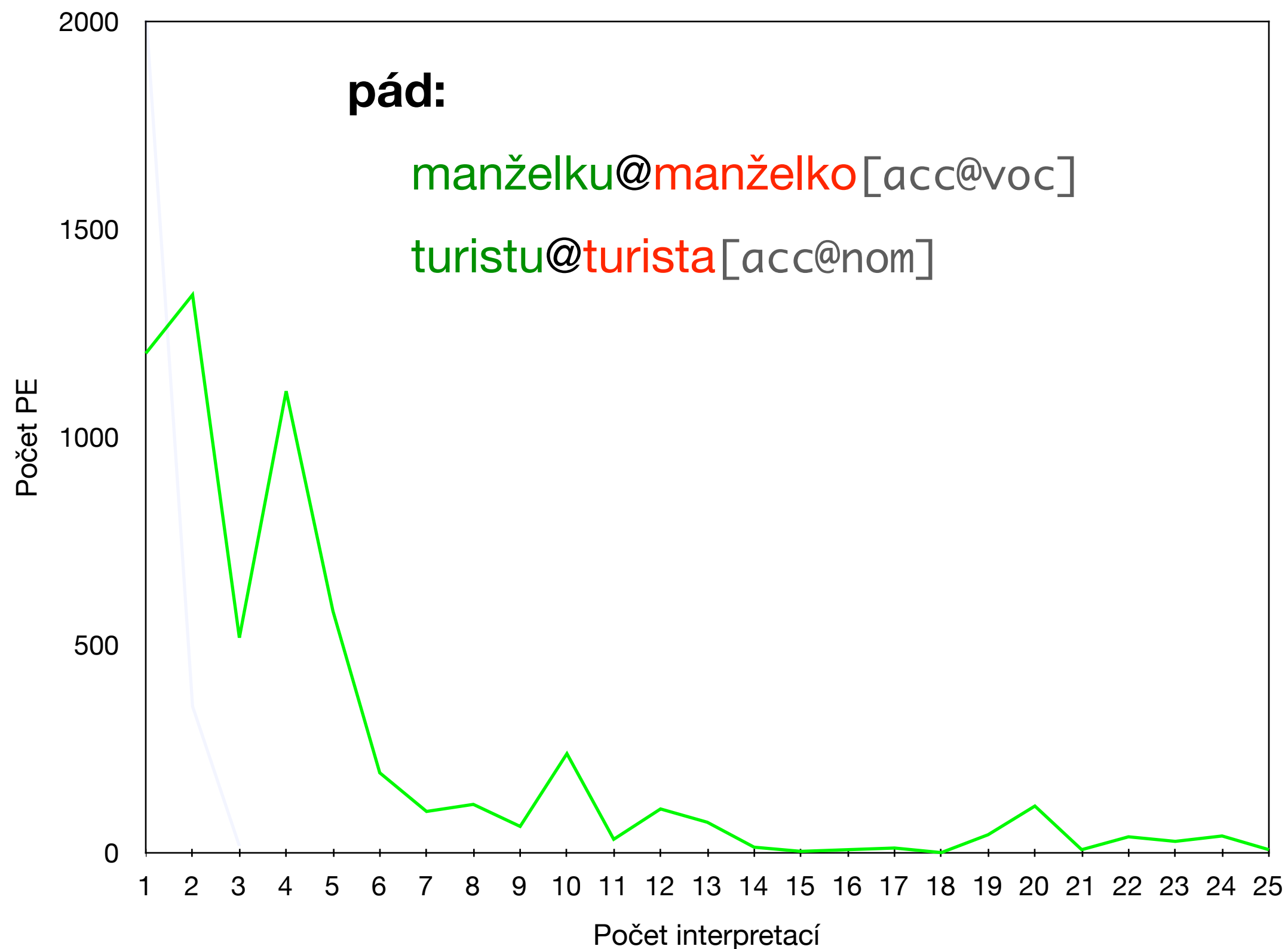
# Vertikální interpretace

Vertikální interpretace (VERT)



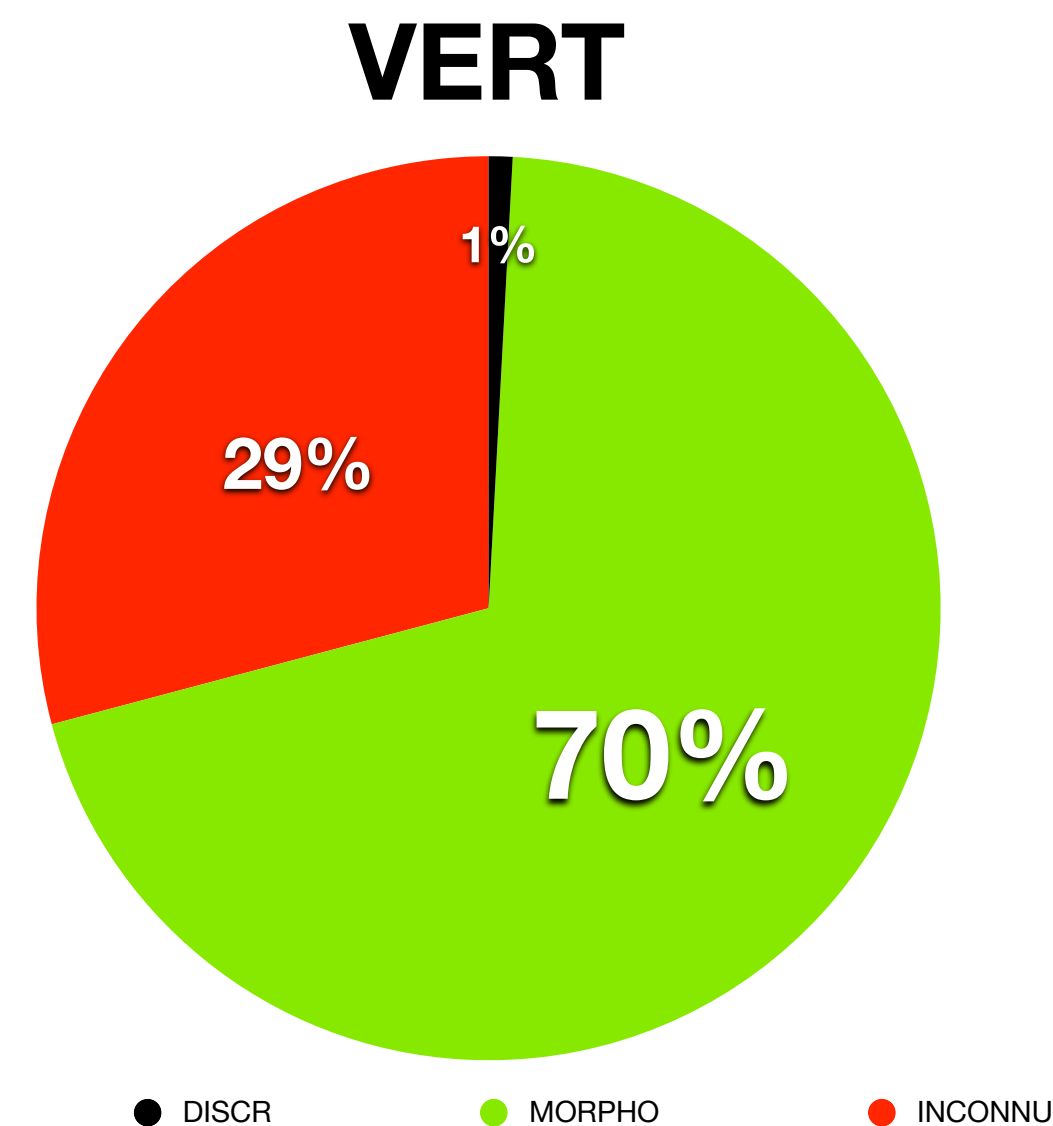
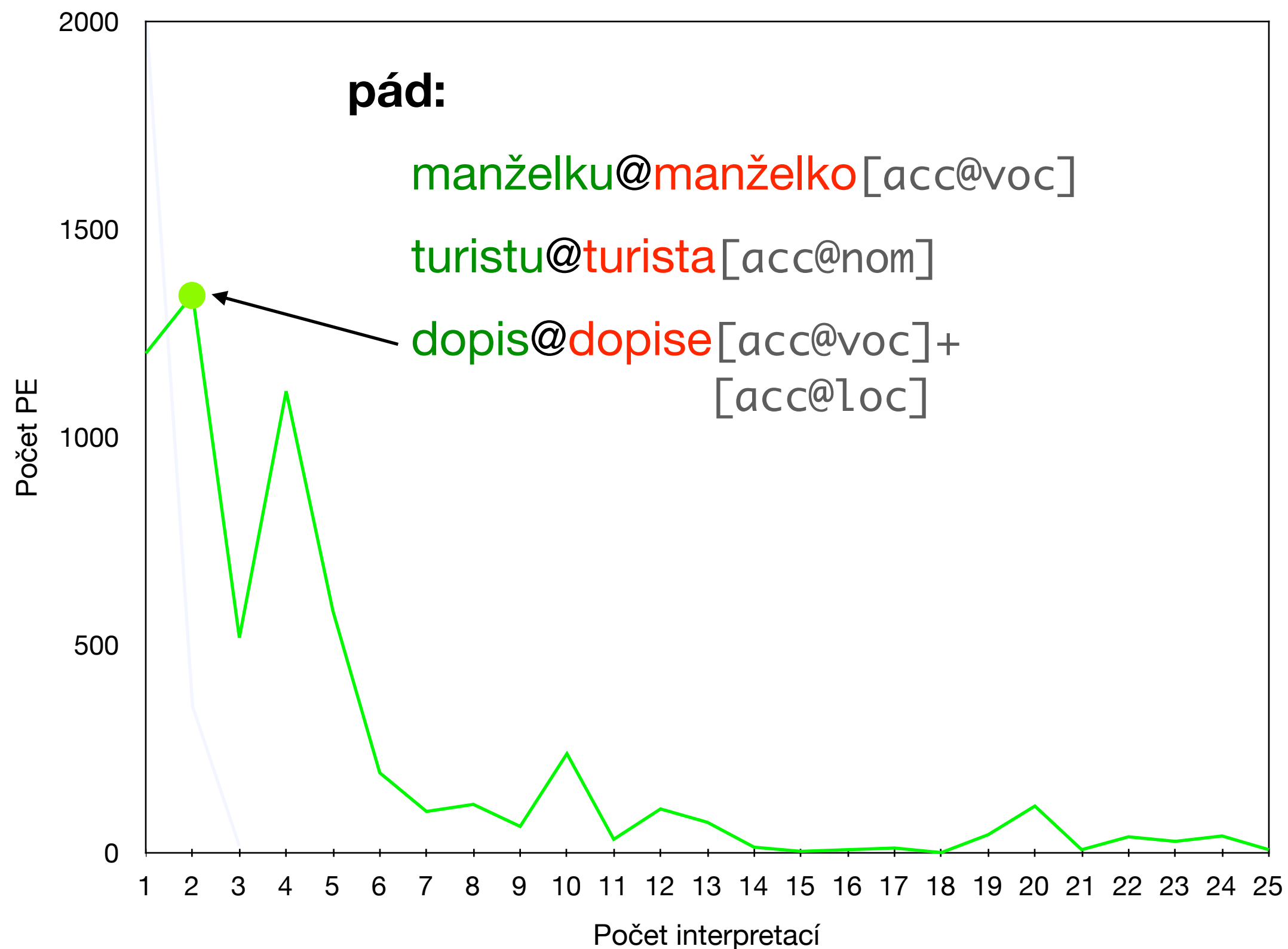
# Vertikální interpretace

Vertikální interpretace (VERT)



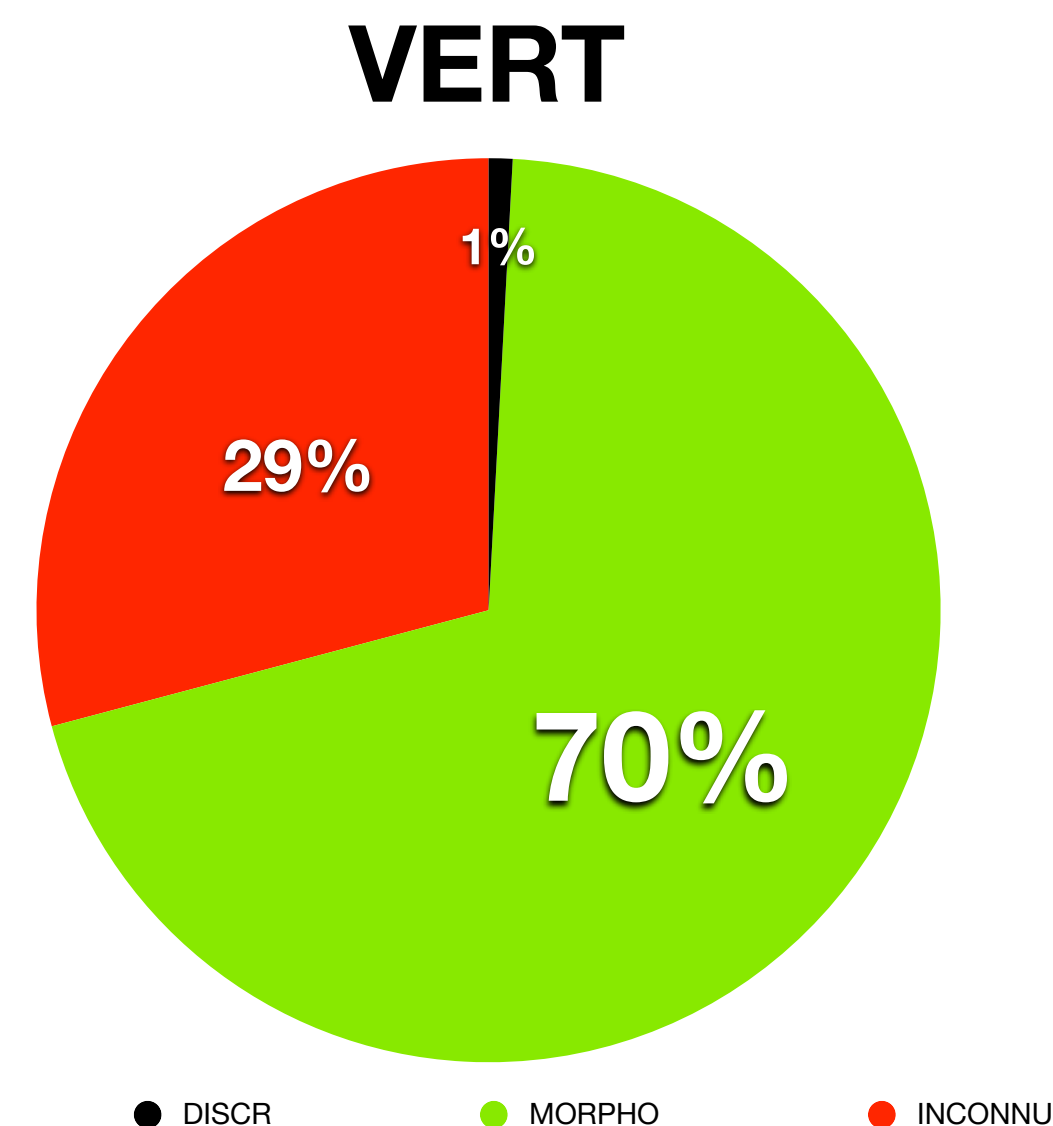
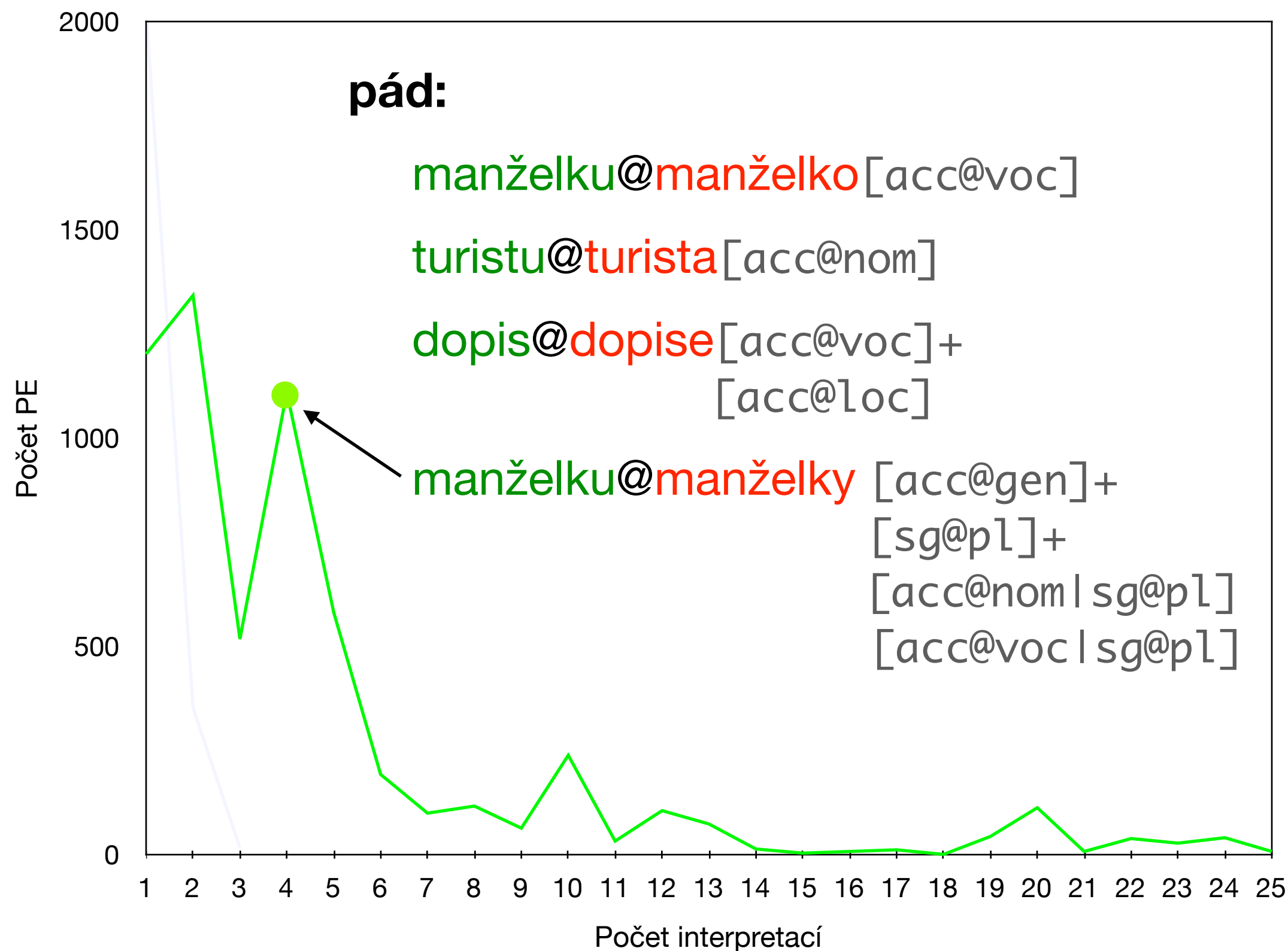
# Vertikální interpretace

Vertikální interpretace (VERT)



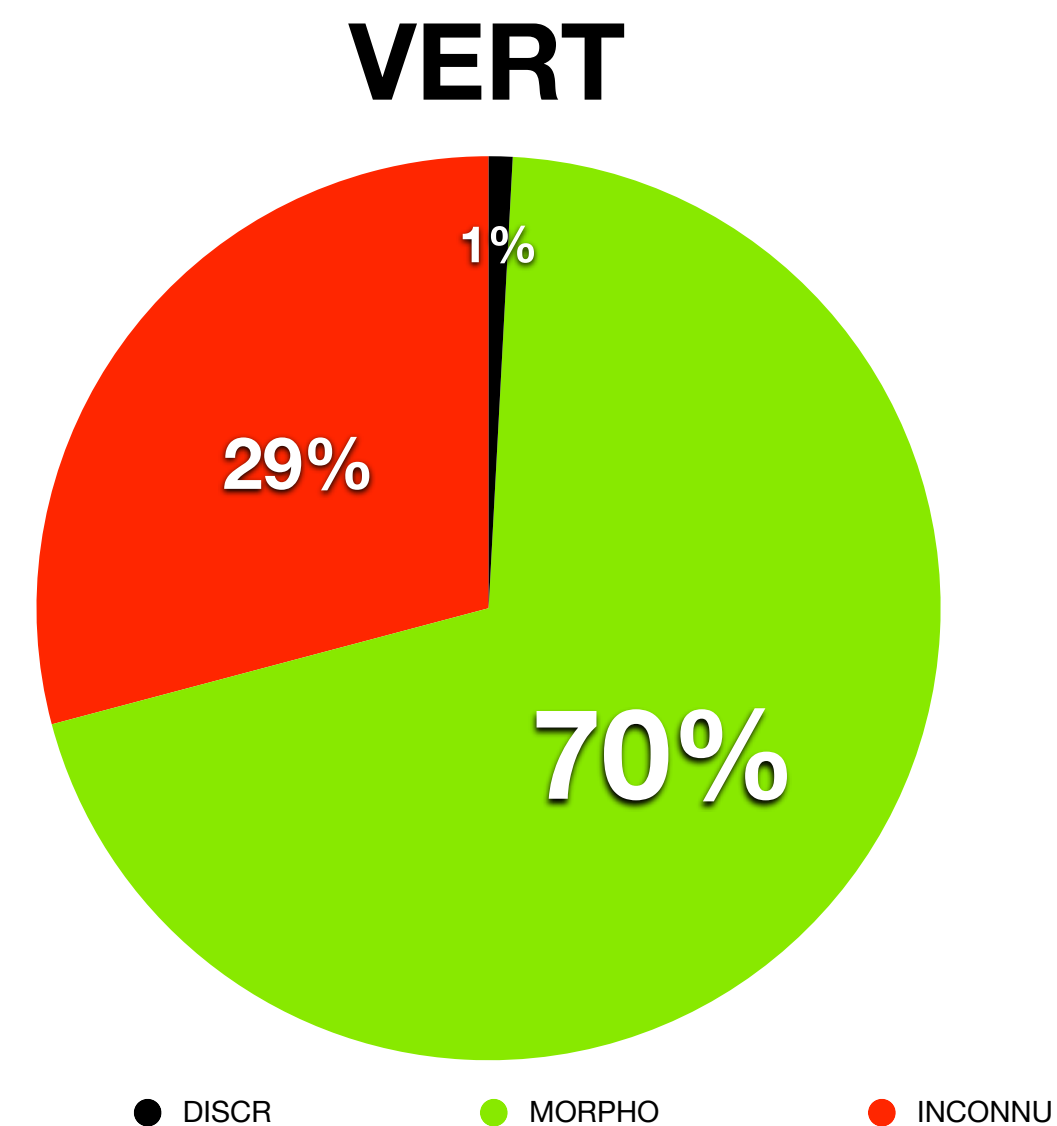
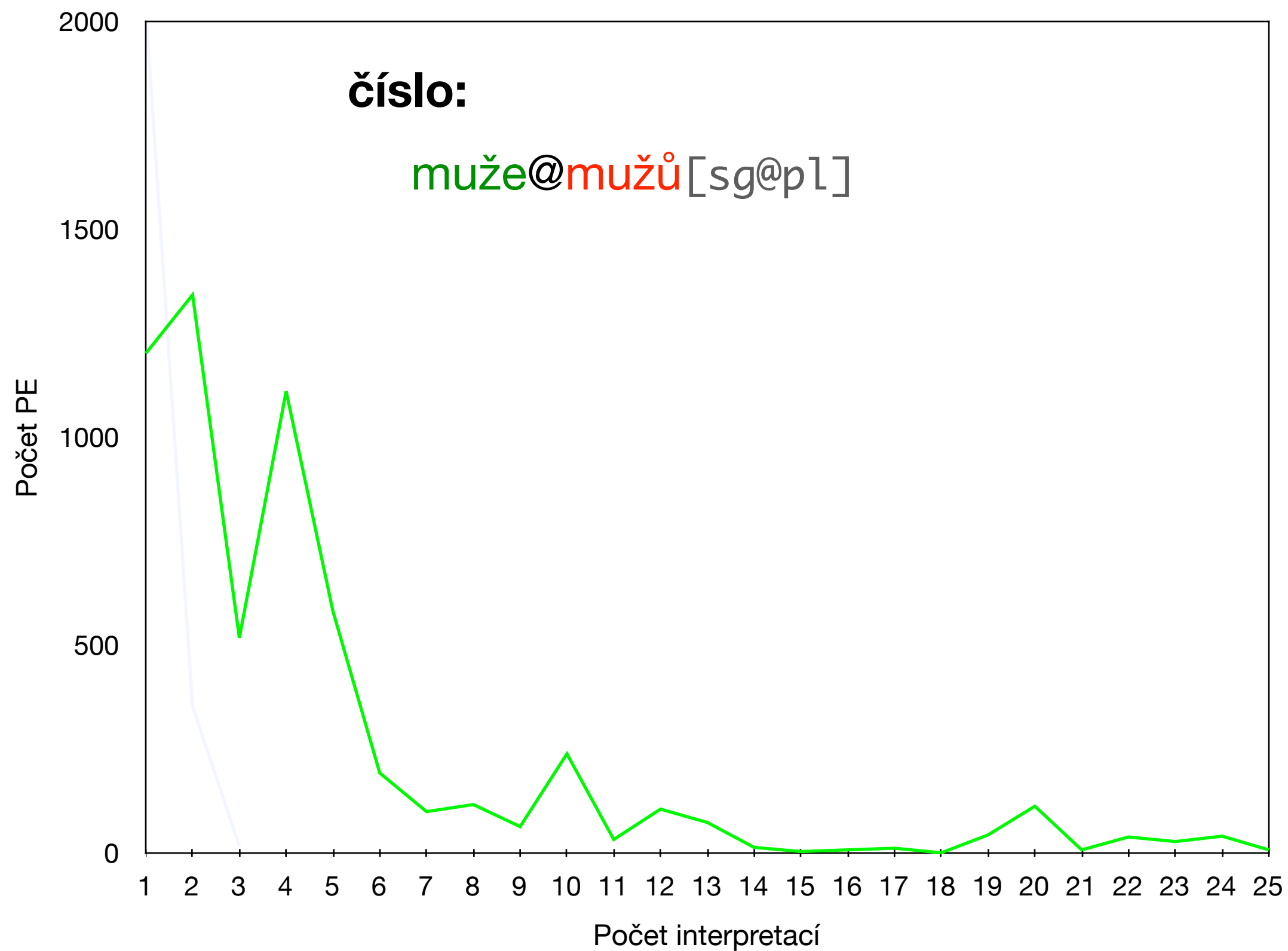
# Vertikální interpretace

Vertikální interpretace (VERT)



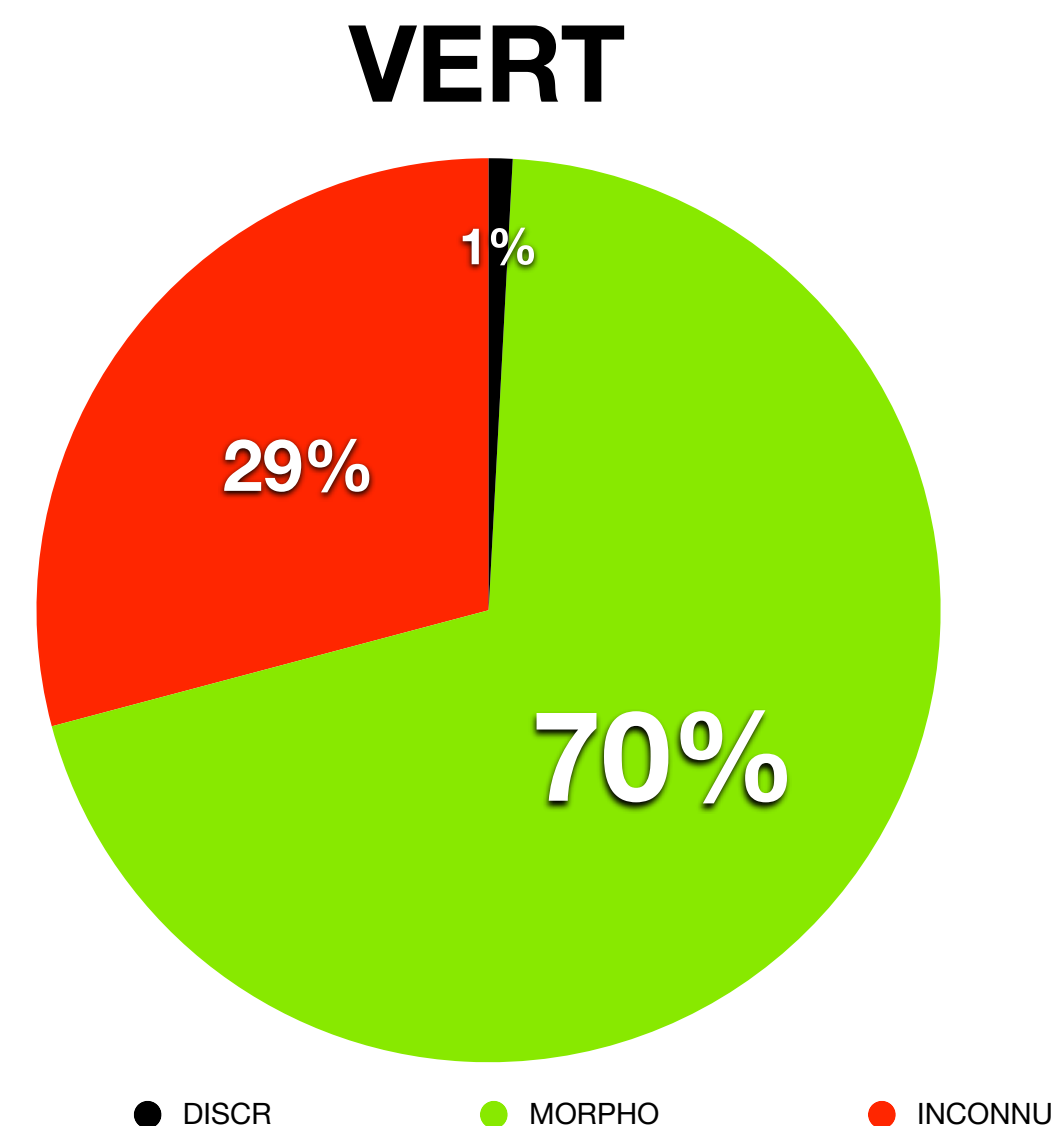
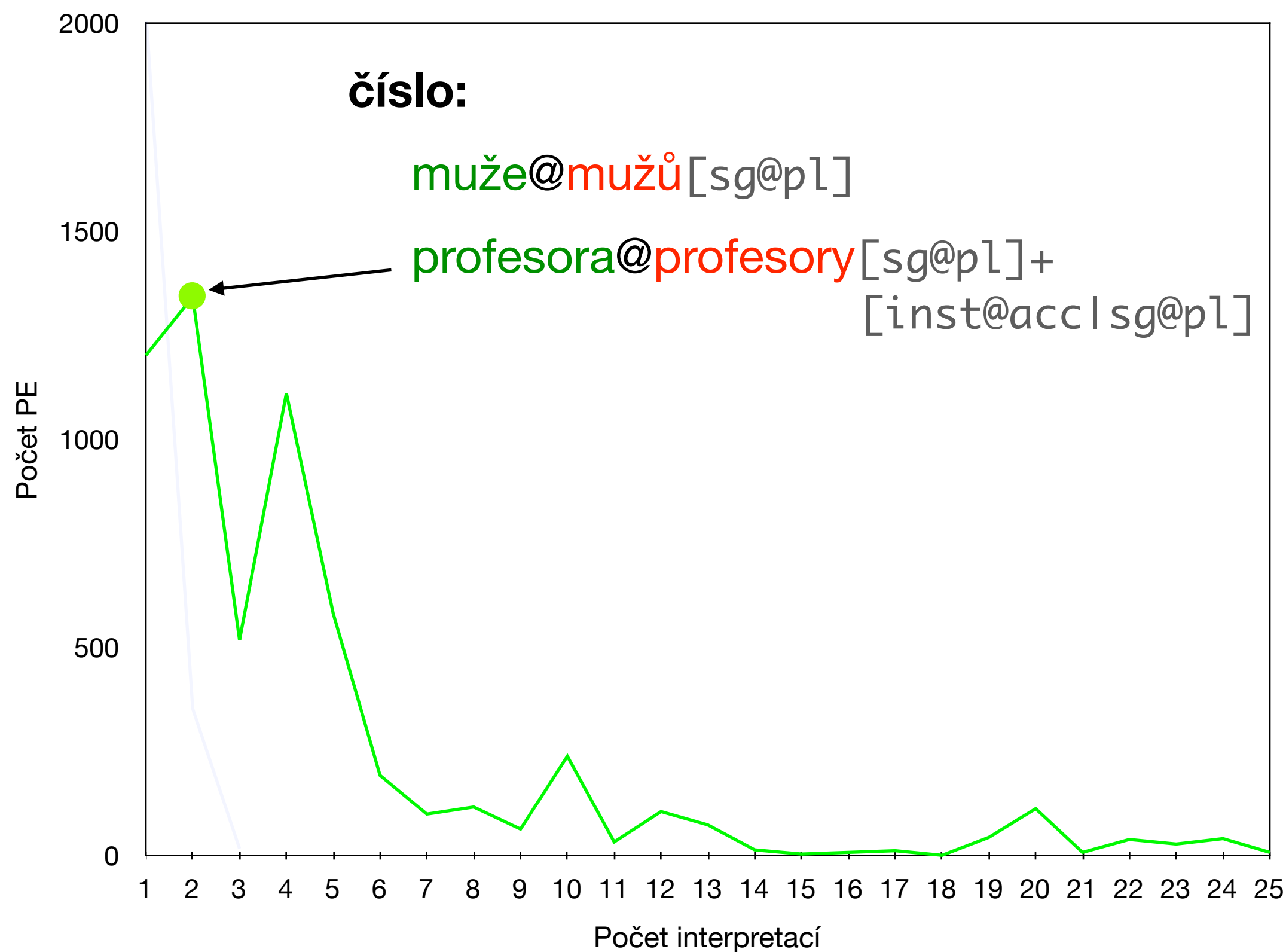
# Vertikální interpretace

Vertikální interpretace (VERT)



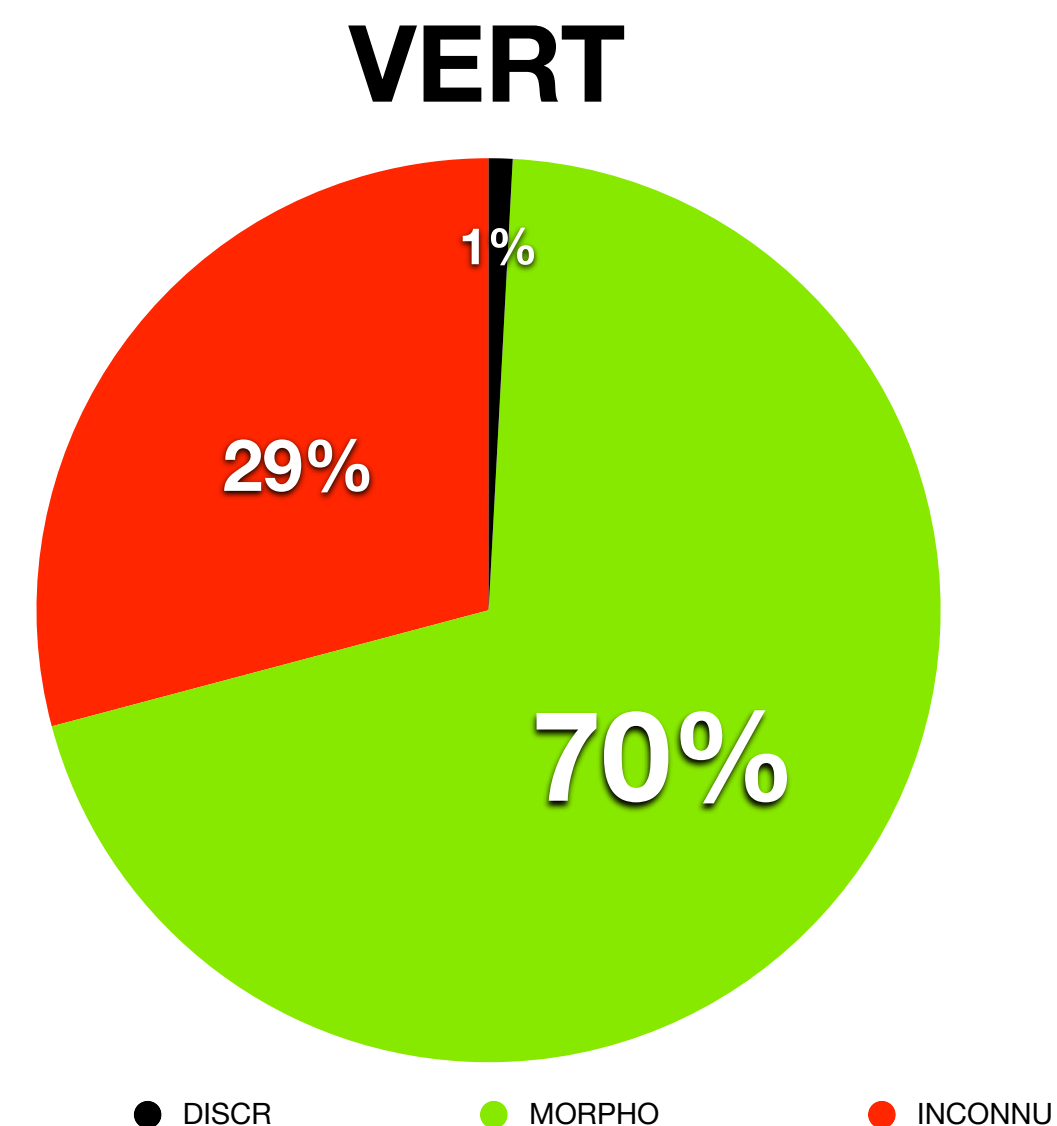
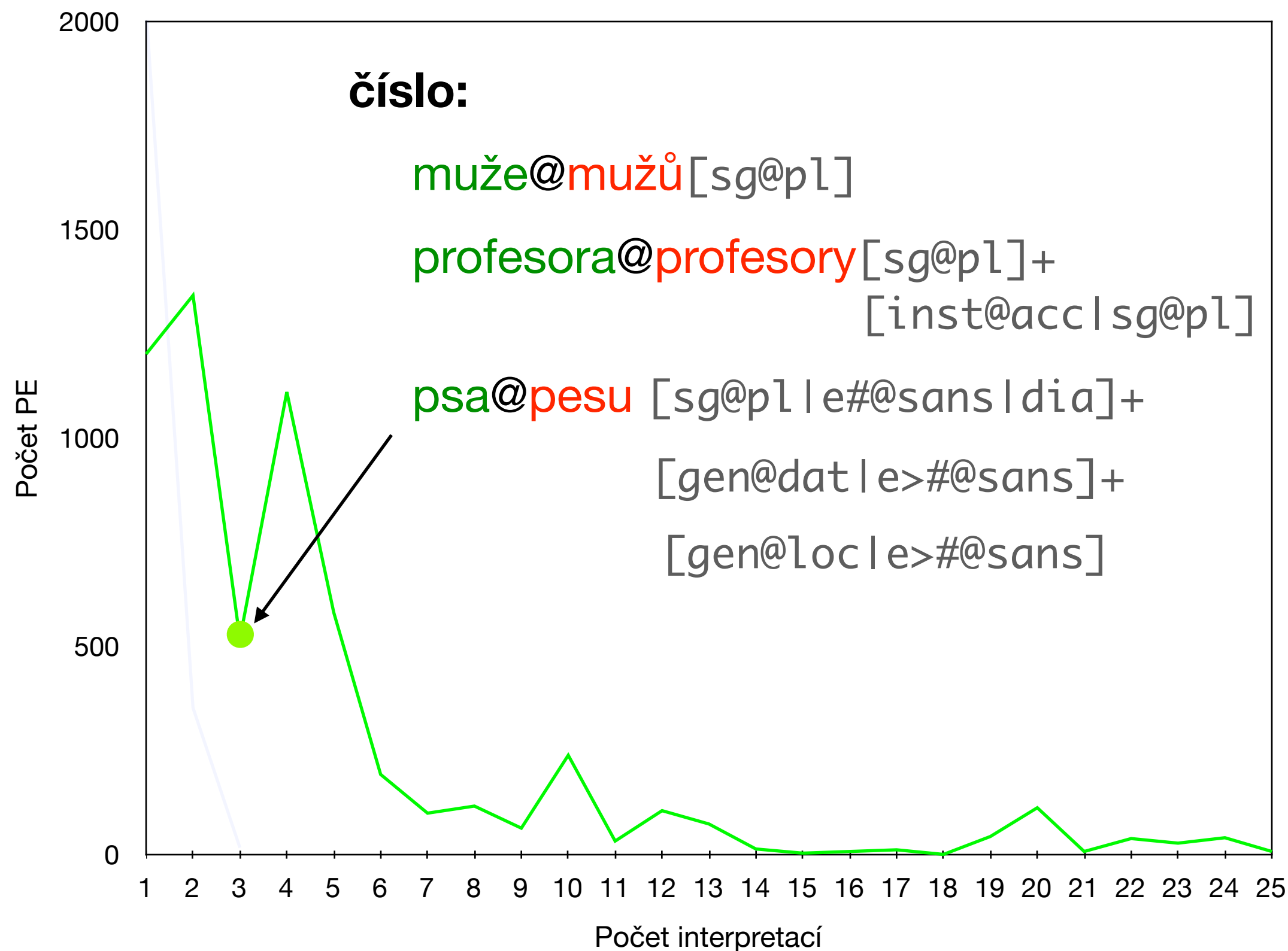
# Vertikální interpretace

Vertikální interpretace (VERT)



# Vertikální interpretace

Vertikální interpretace (VERT)



# Vertikální interpretace

Vertikální interpretace (VERT)

rod u adjektiv a zájmen:

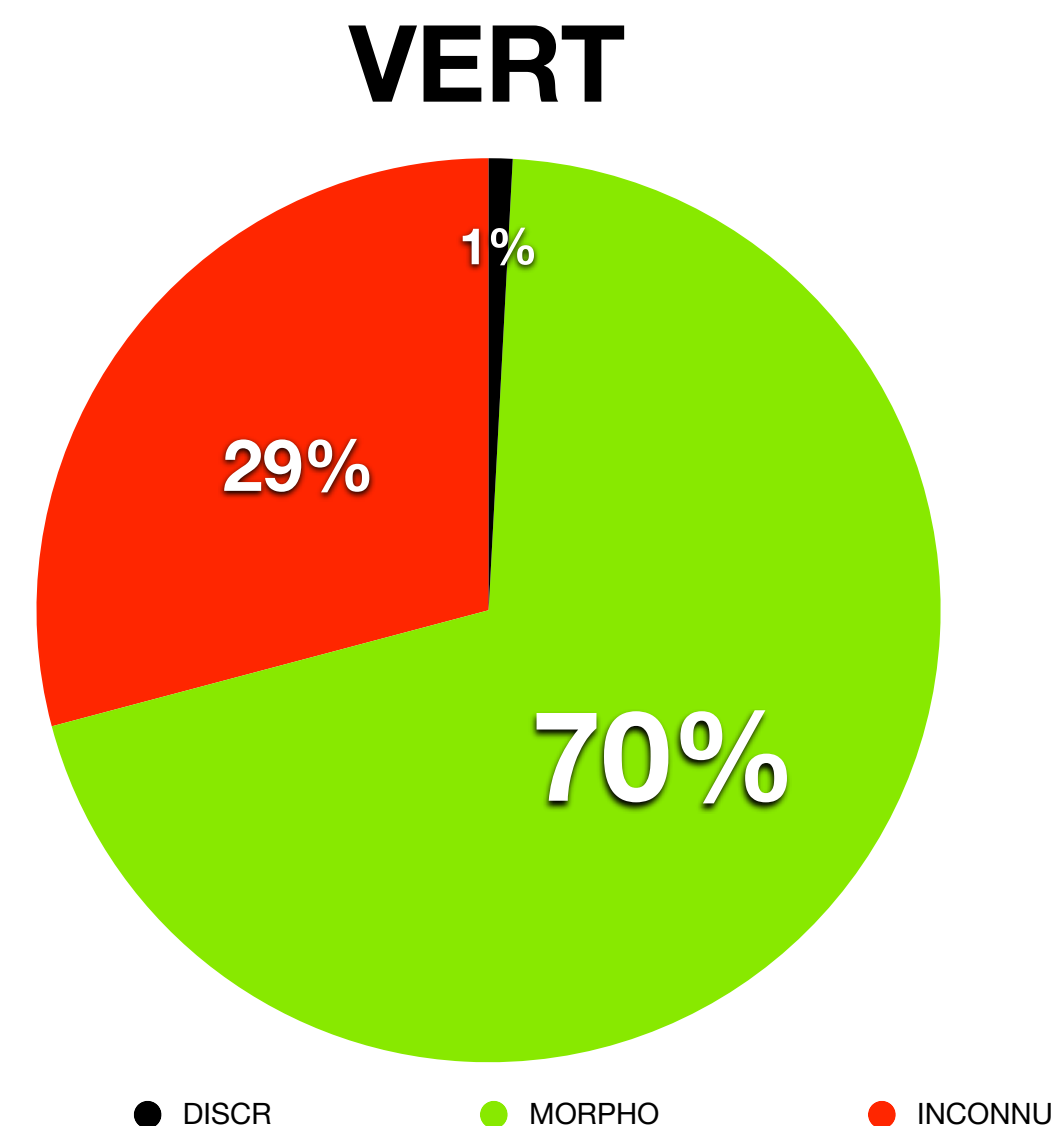
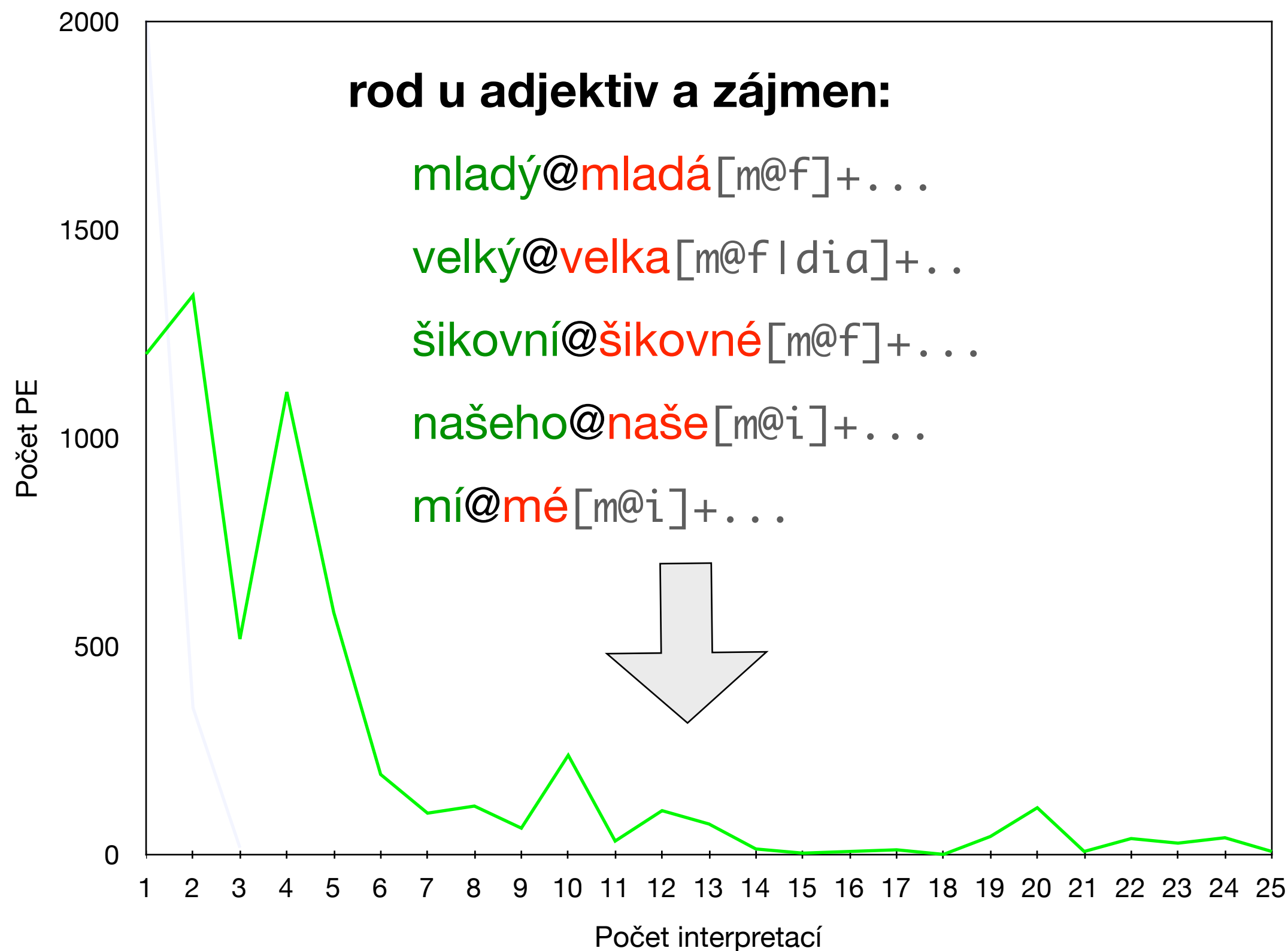
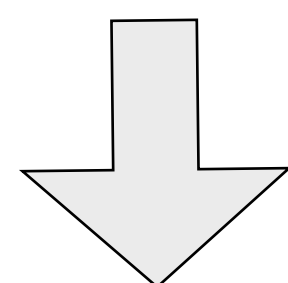
mladý@mladá[m@f]+...

velký@velka[m@f|dia]+...

šikovní@šikovné[m@f]+...

našeho@naše[m@i]+...

mí@mé[m@i]+...

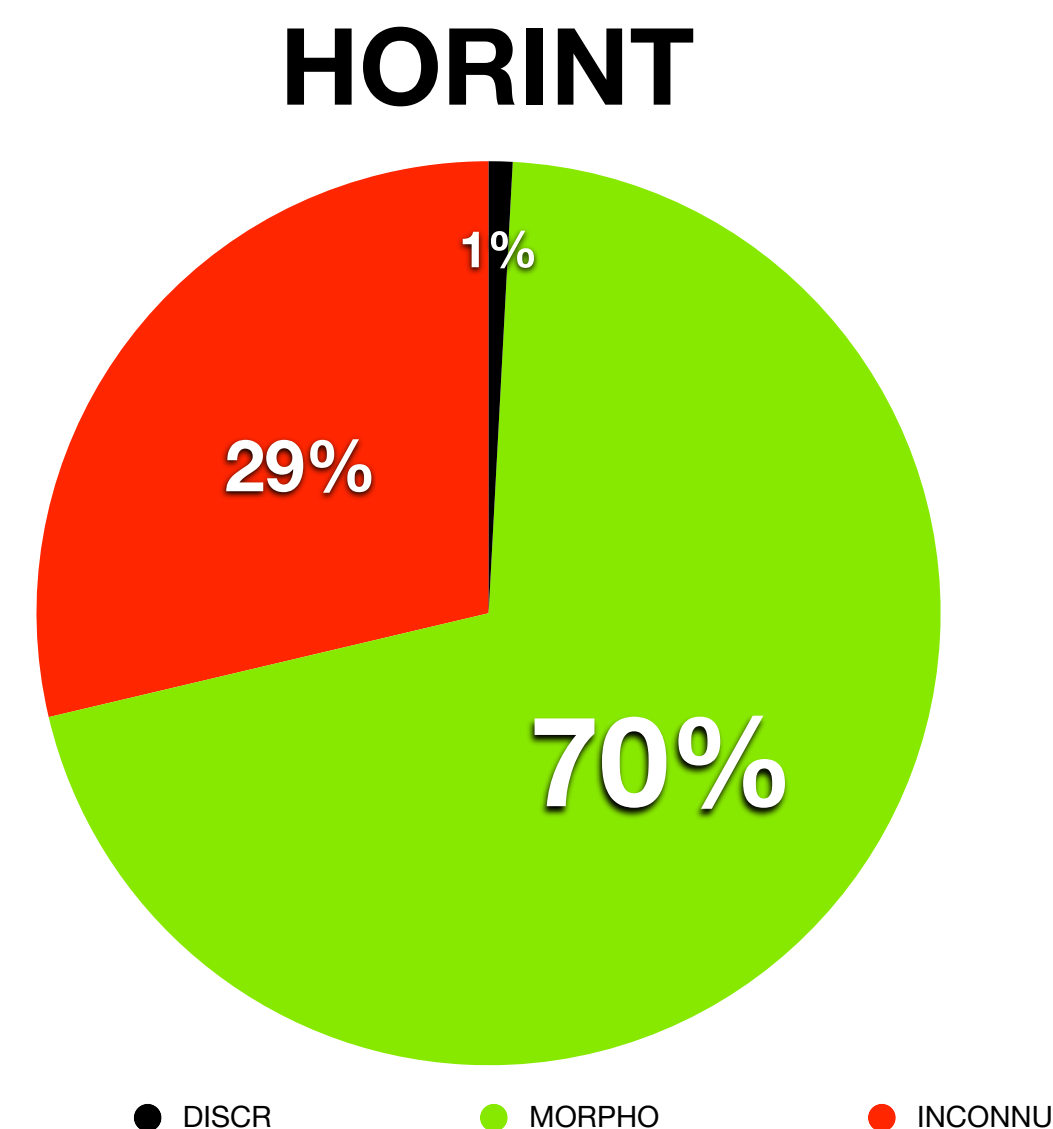
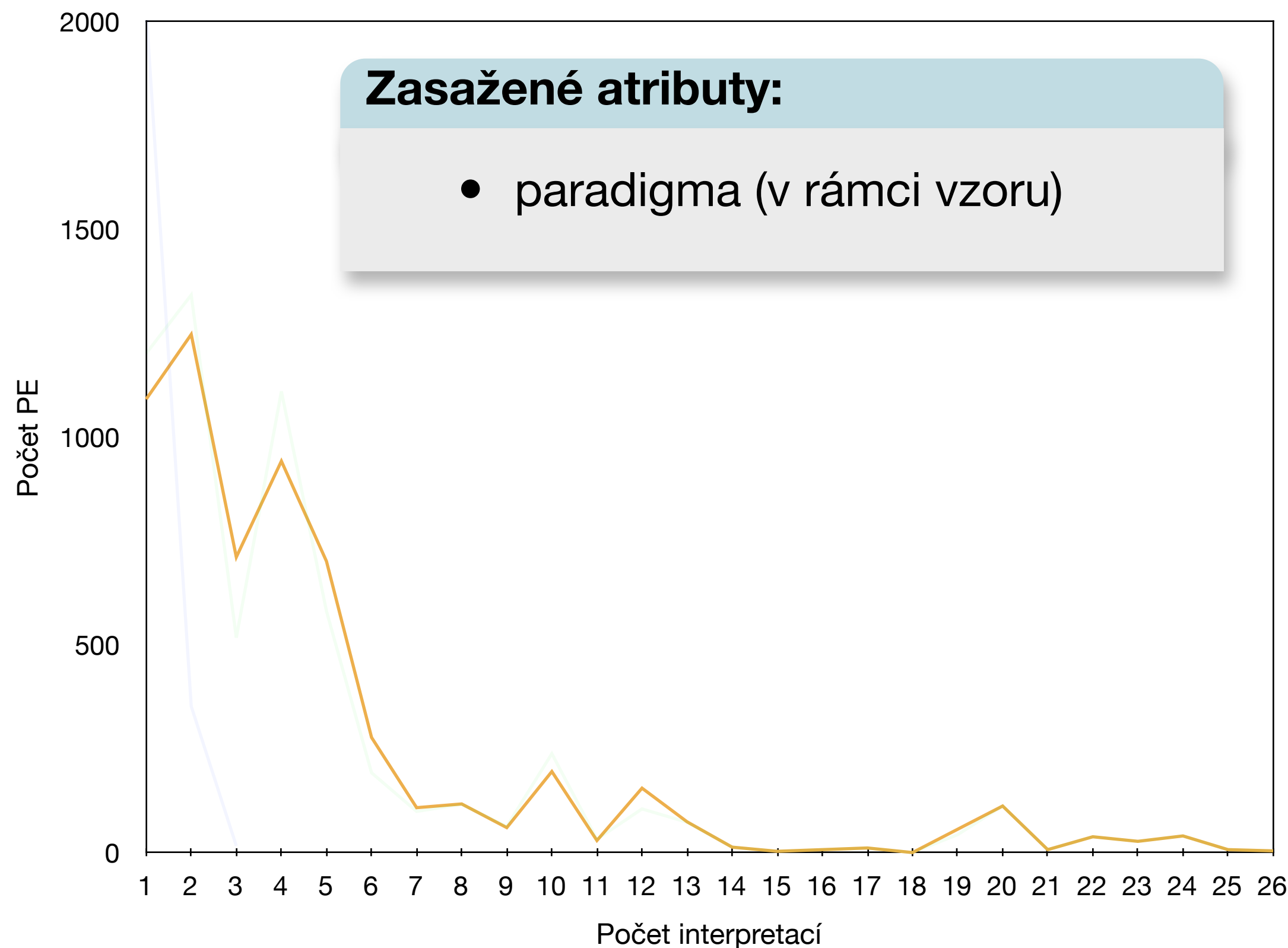


# Horizontální interpretace interní

Horizontální interpretace interní (HORINT a HORINT\_norestr)

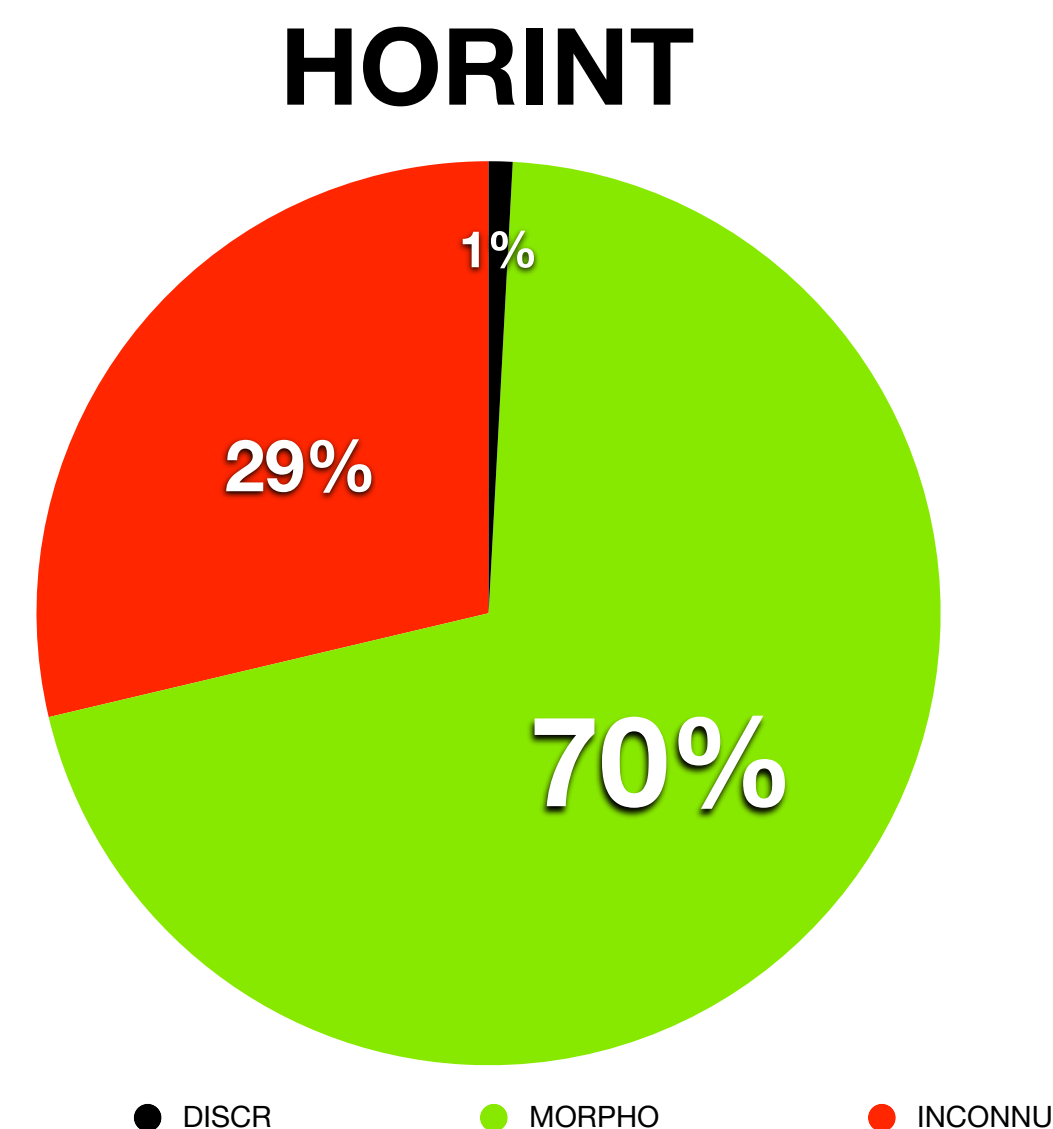
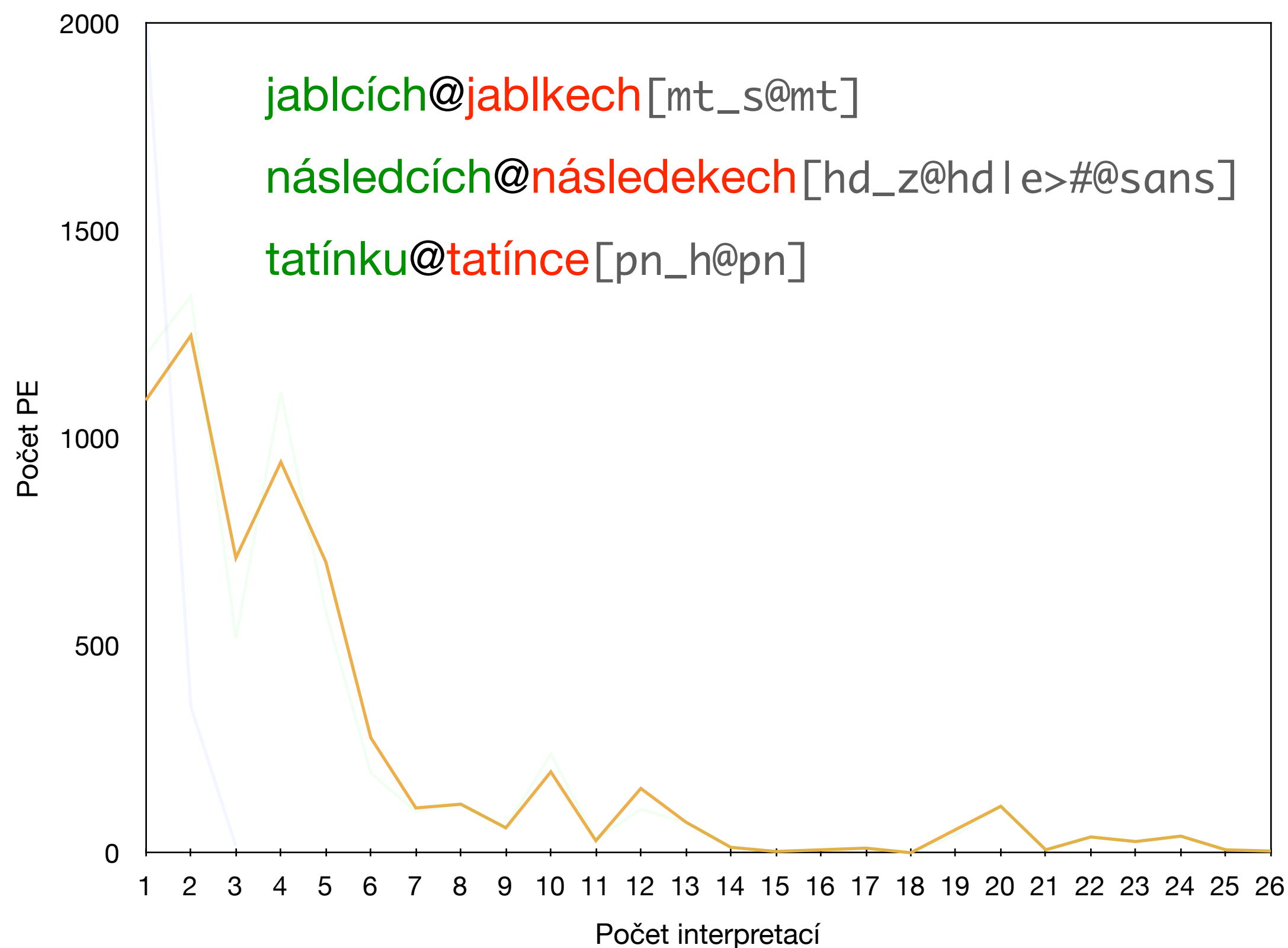
## Zasažené atributy:

- paradigma (v rámci vzoru)



# Horizontální interpretace interní

Horizontální interpretace interní (HORINT a HORINT\_norestr)

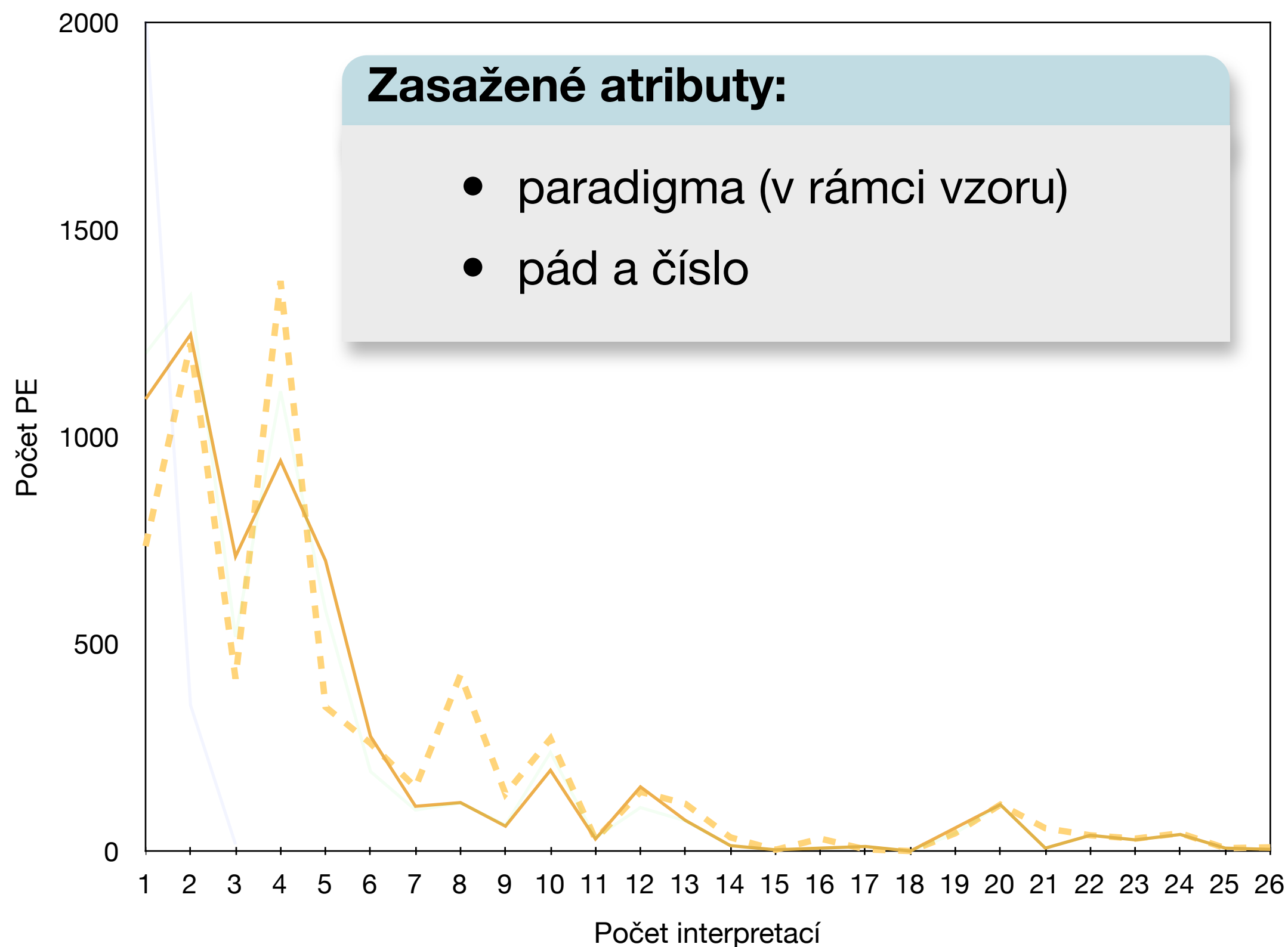


# Horizontální interpretace interní (bez omezení)

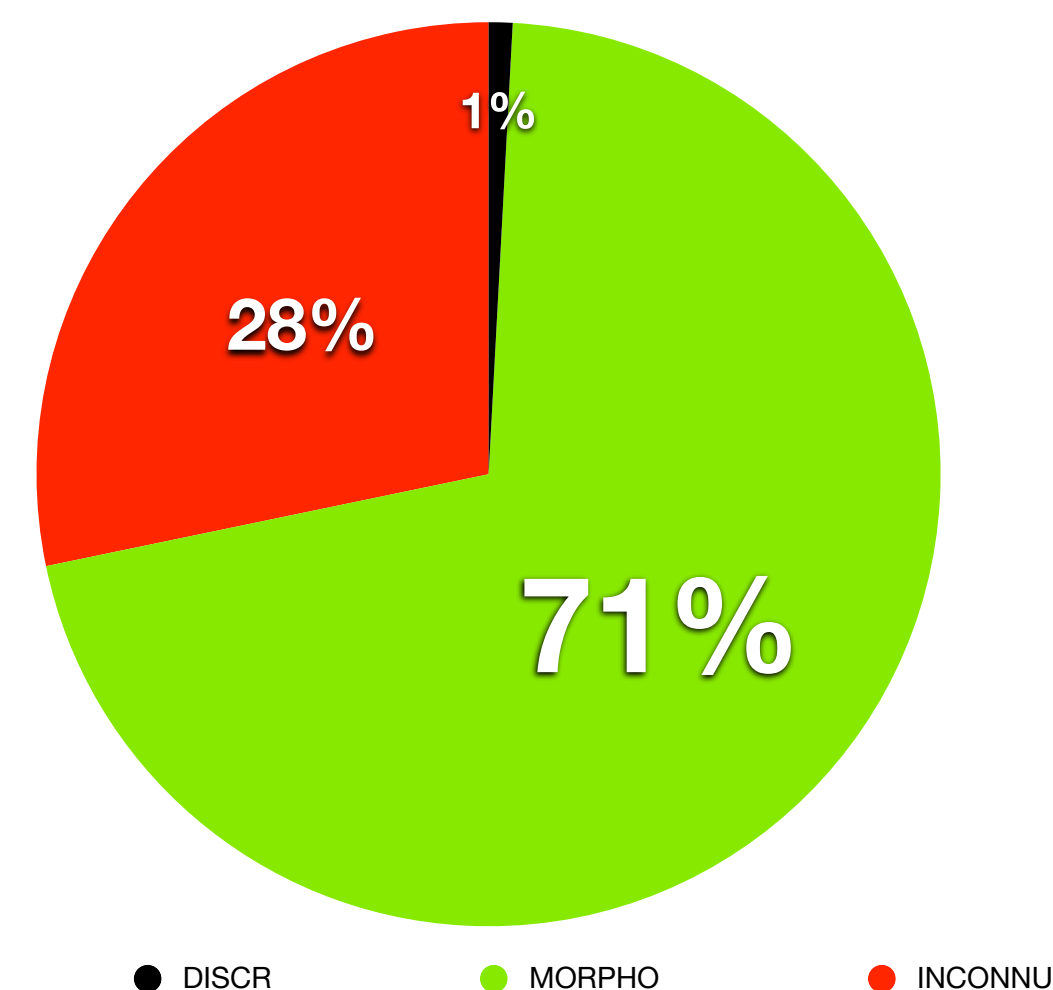
Horizontální interpretace interní (HORINT a HORINT\_norestr)

## Zasažené atributy:

- paradigma (v rámci vzoru)
- pád a číslo

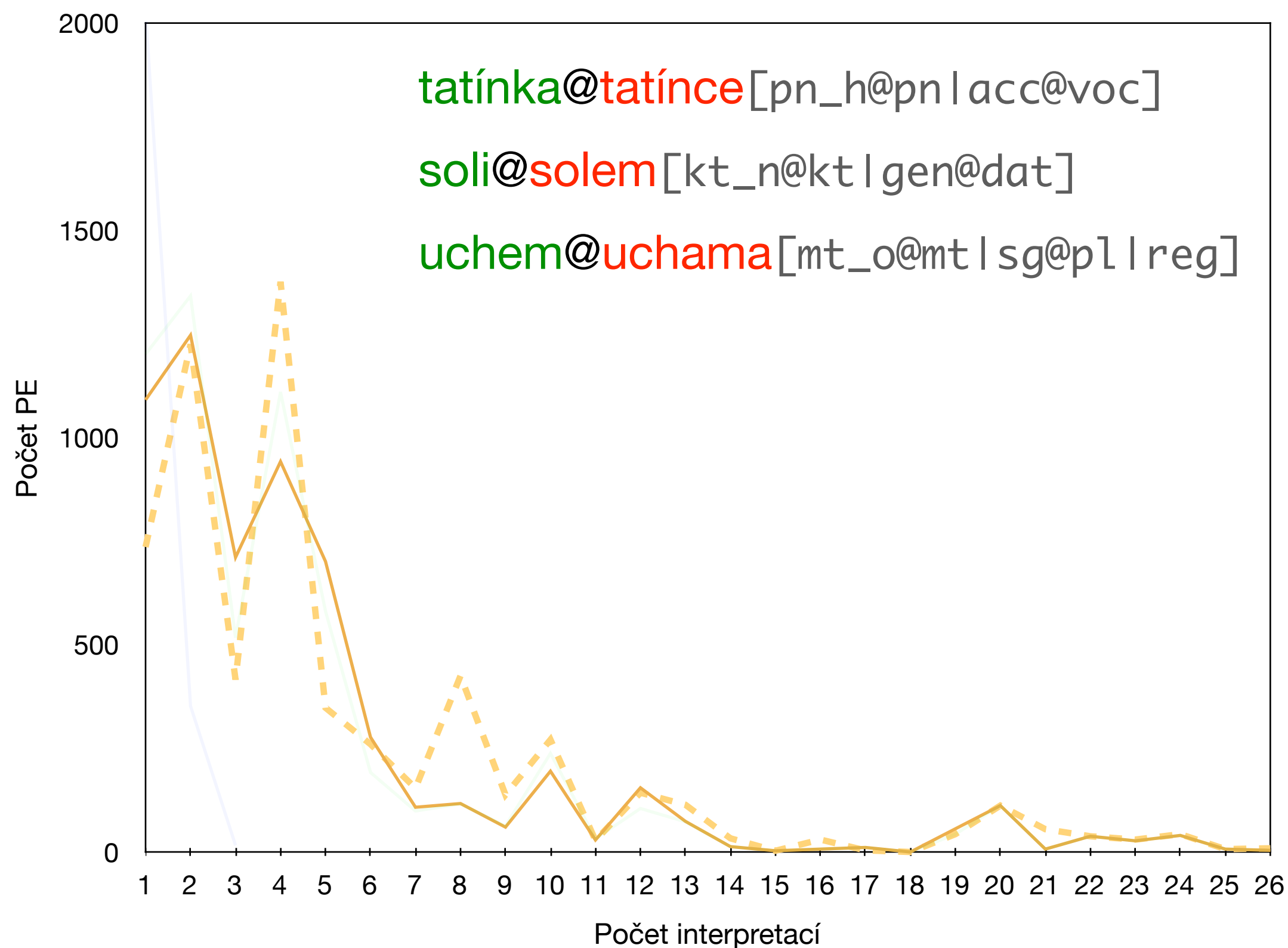


## HORINT\_norestr

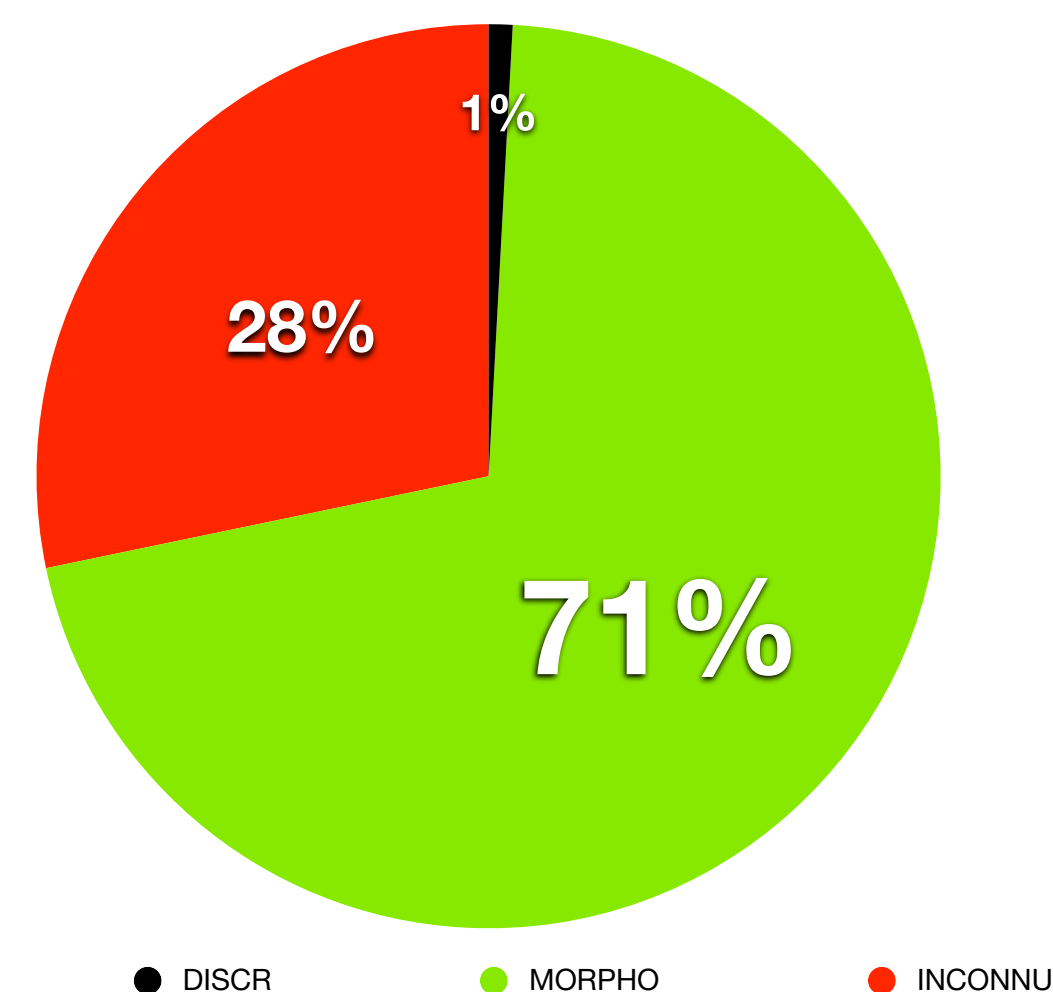


# Horizontální interpretace interní (bez omezení)

Horizontální interpretace interní (HORINT a HORINT\_norestr)

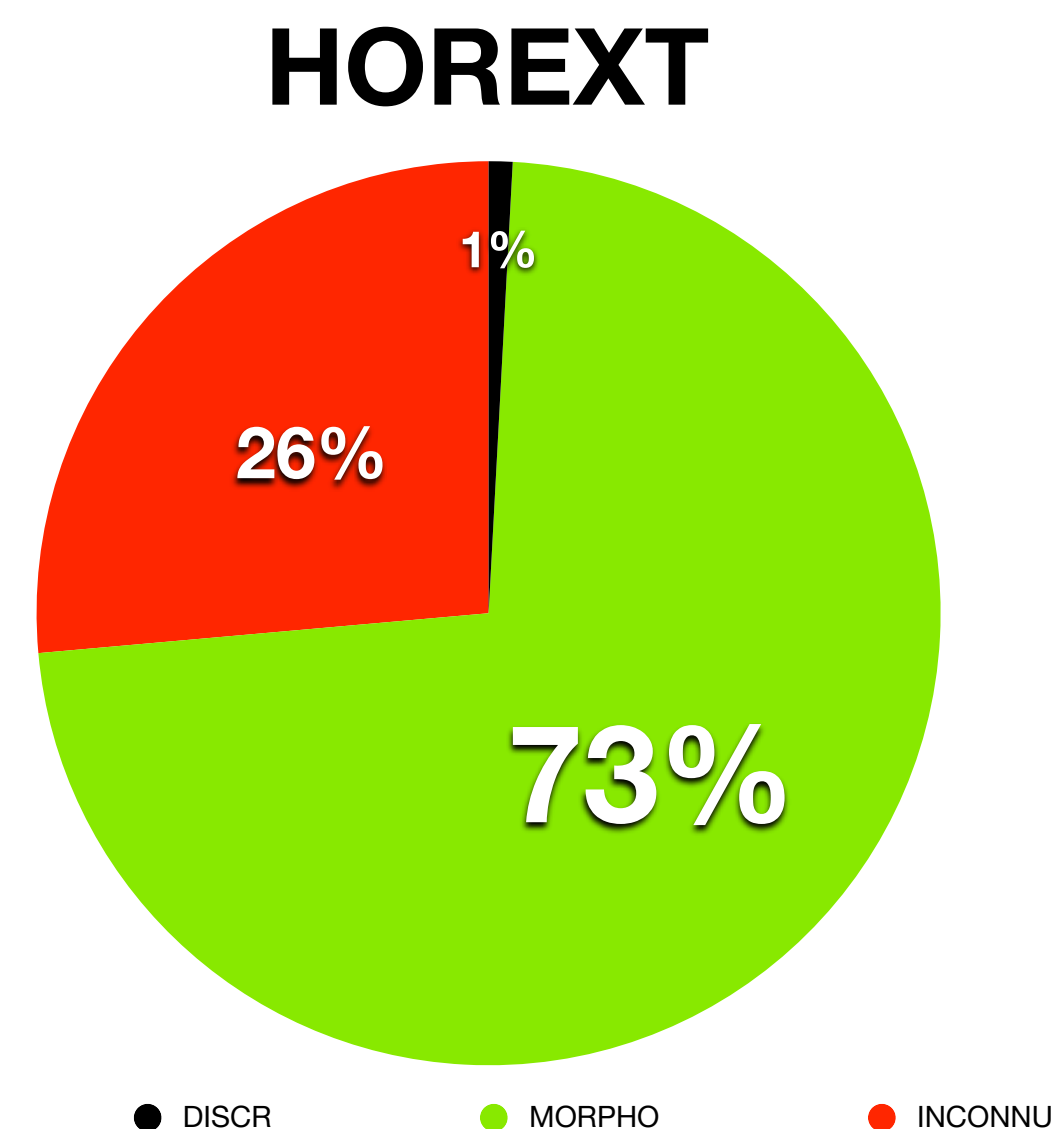
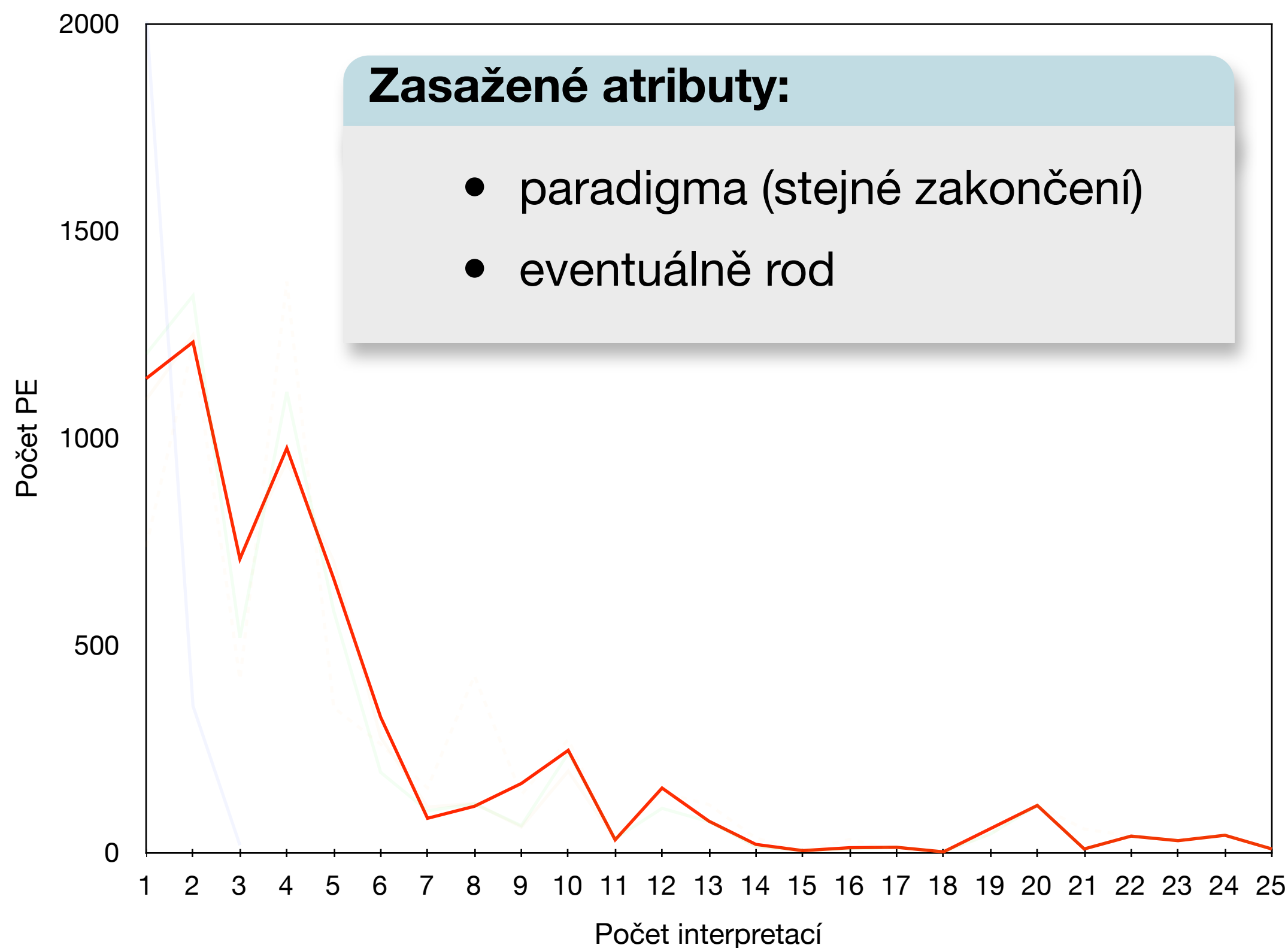


## HORINT\_norestr



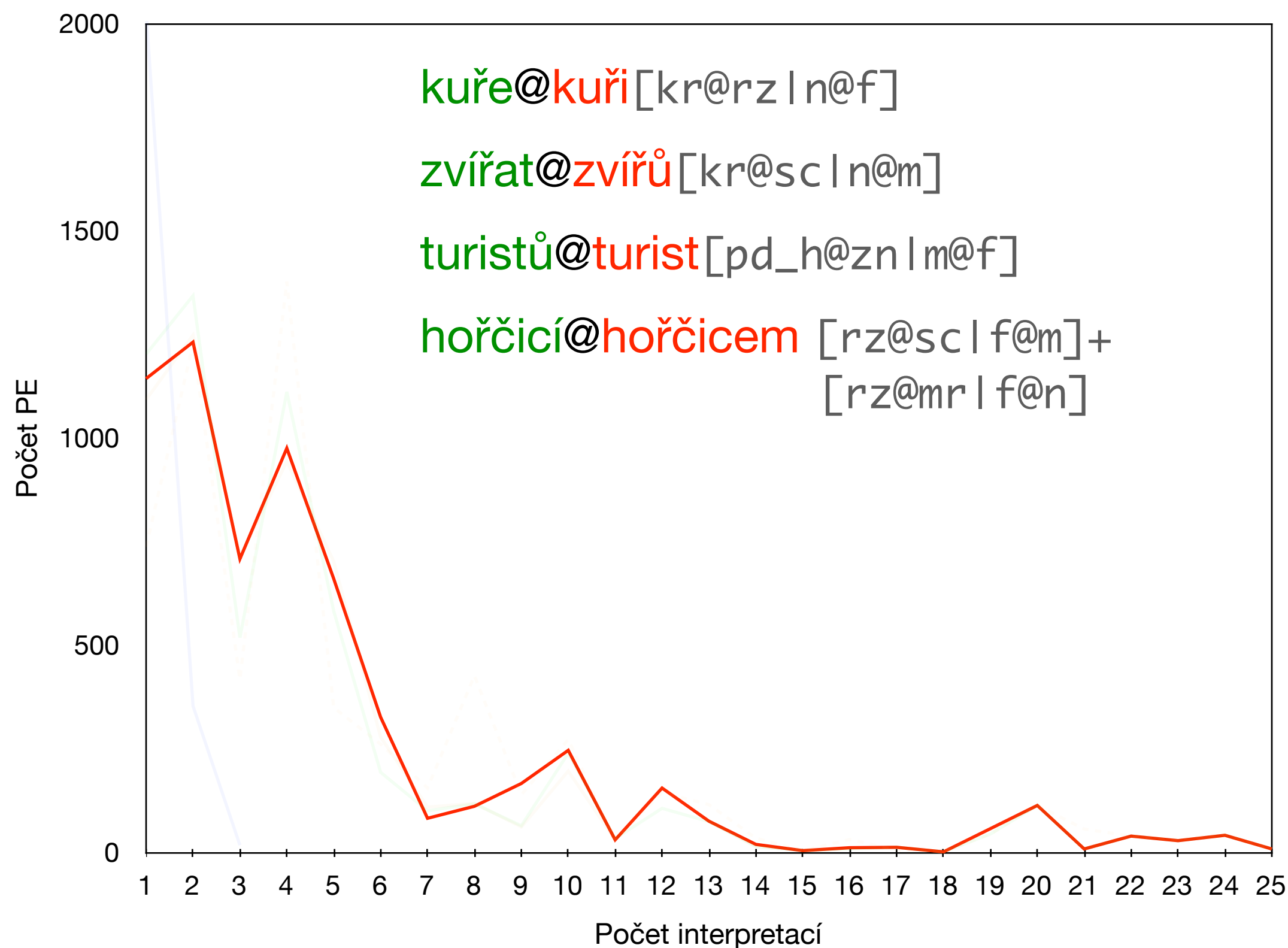
# Horizontální interpretace externí

Horizontální interpretace externí

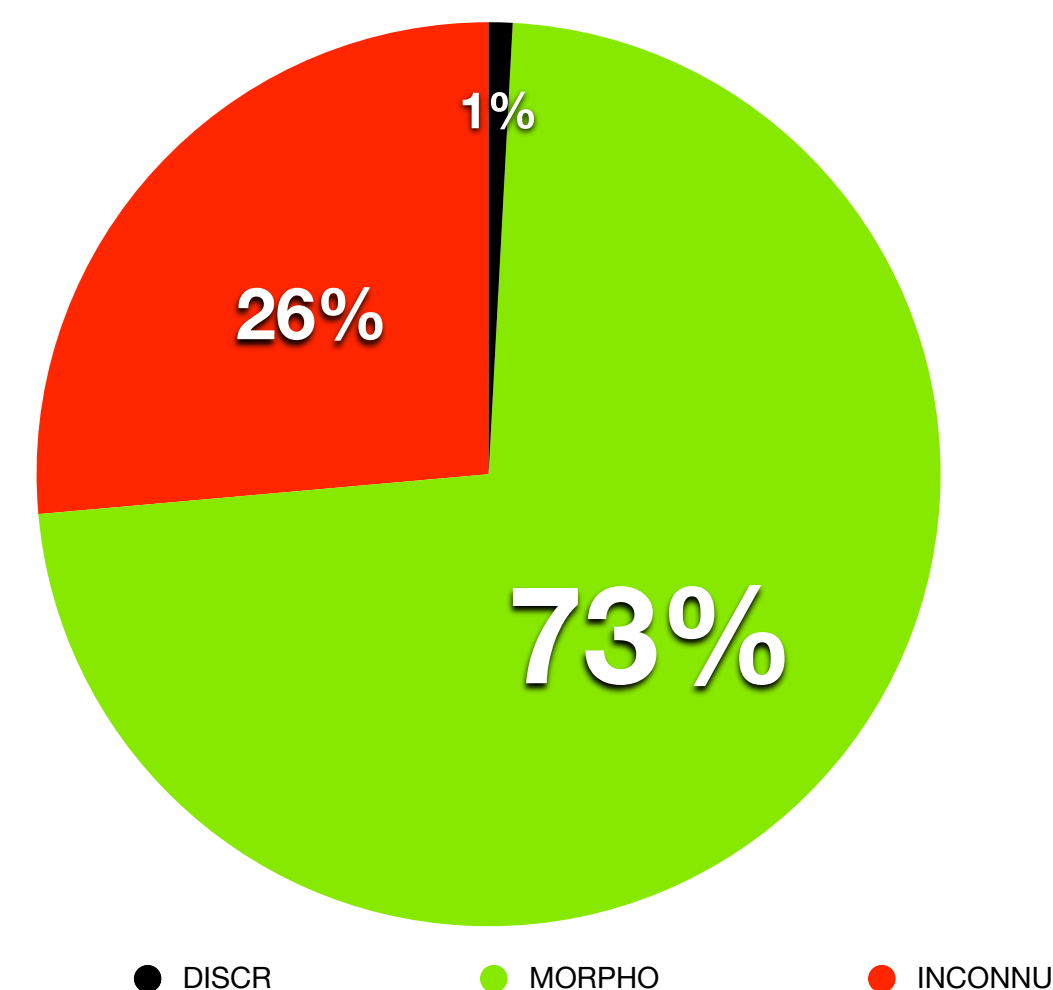


# Horizontální interpretace externí

Horizontální interpretace externí

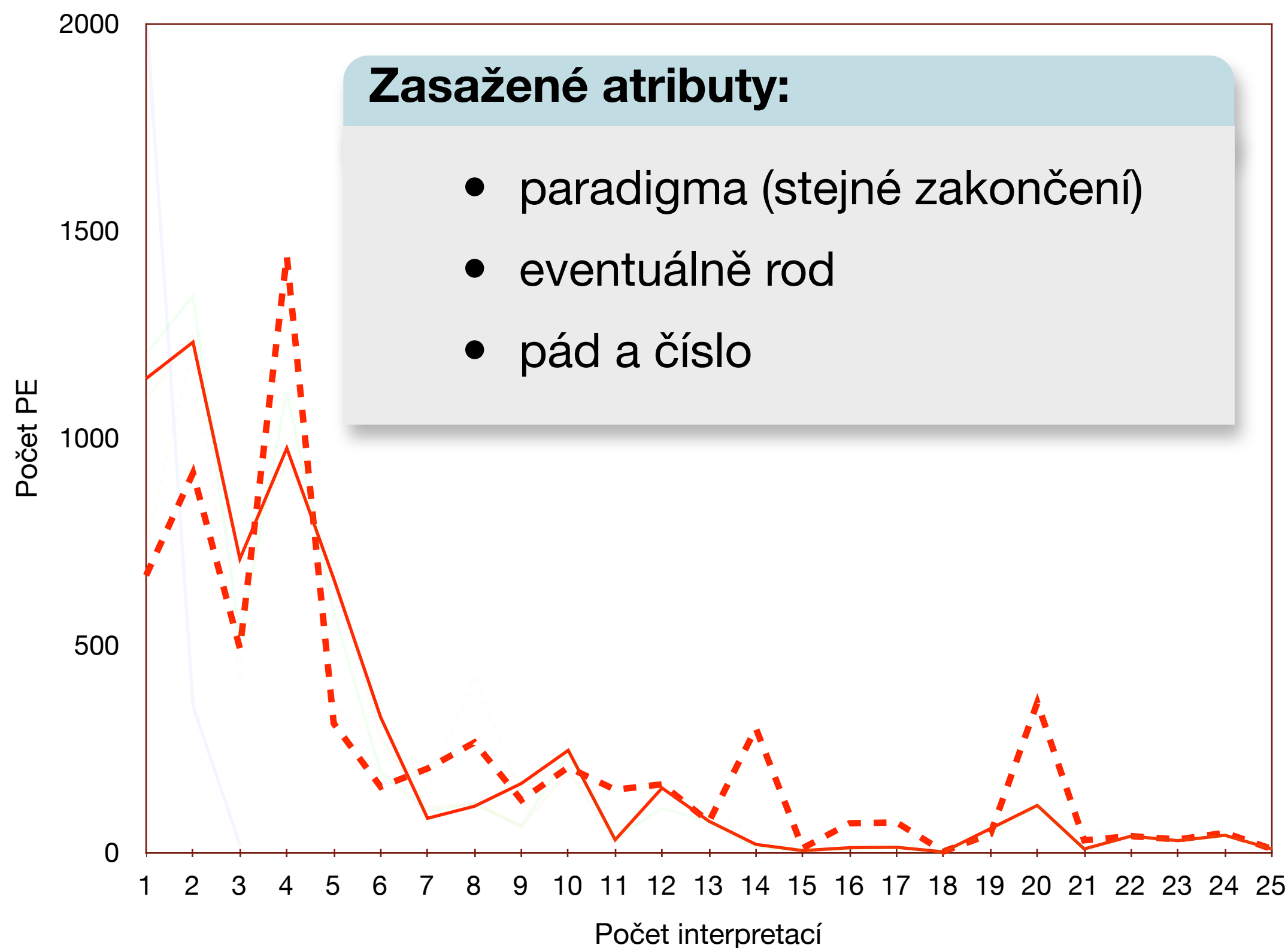


## HOREXT

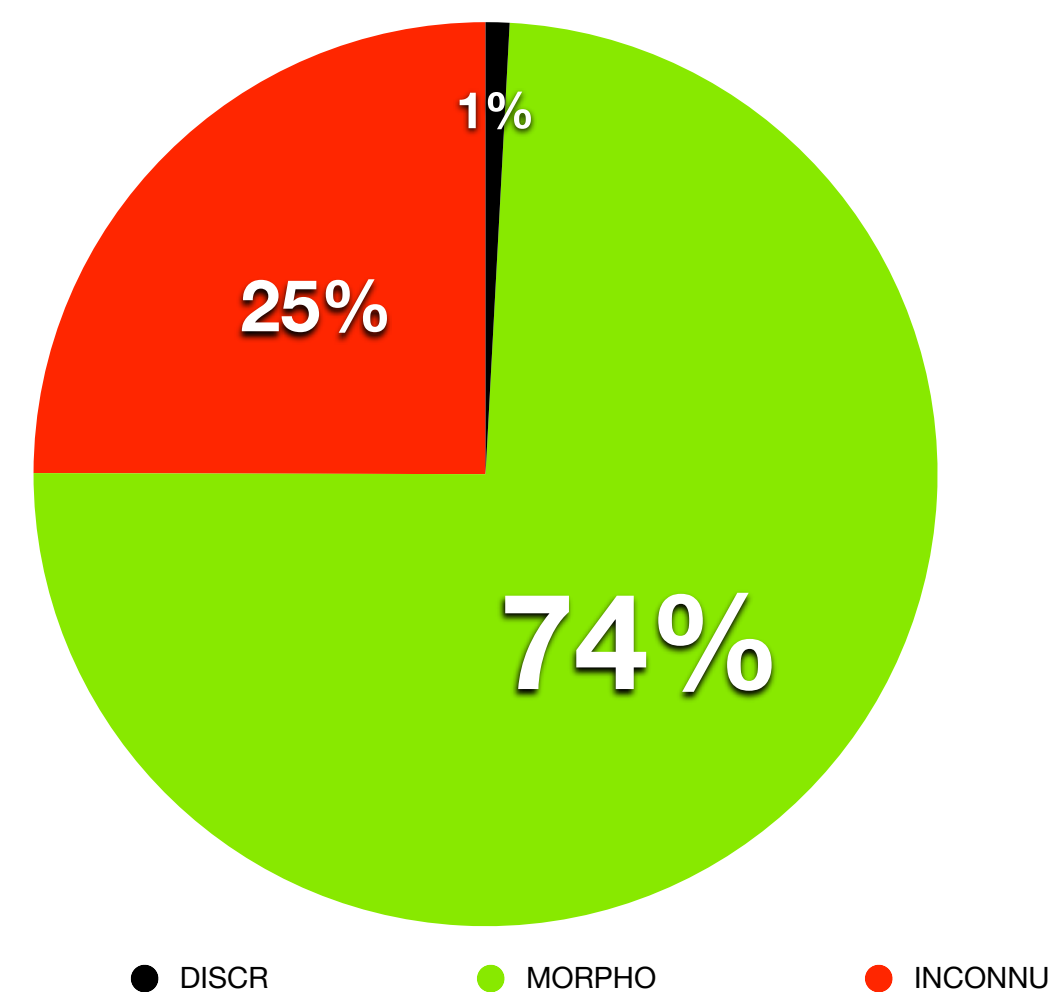


# Horizontální interpretace externí (bez omezení)

Horizontální interpretace externí

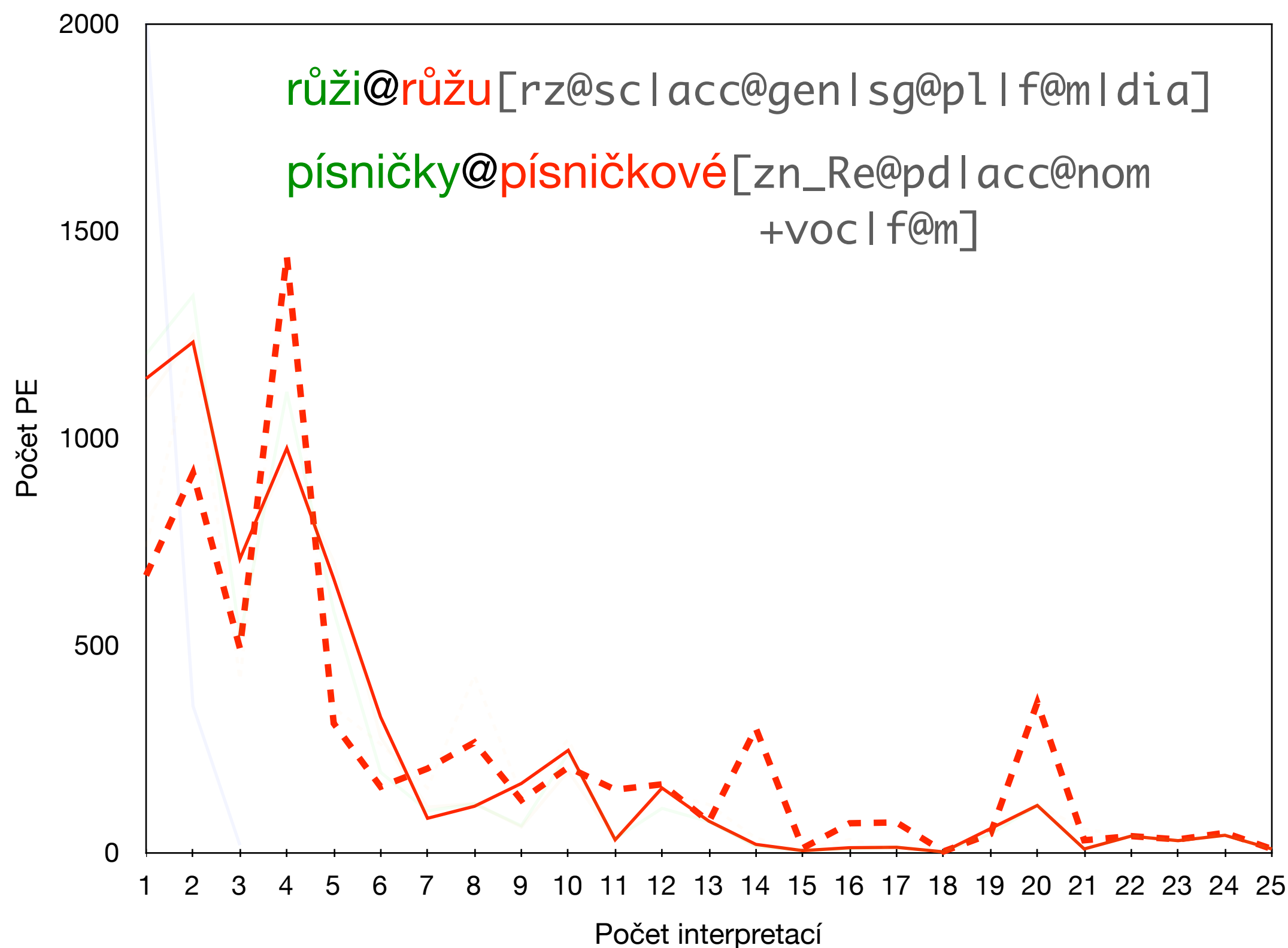


**HOREXT\_norestr**

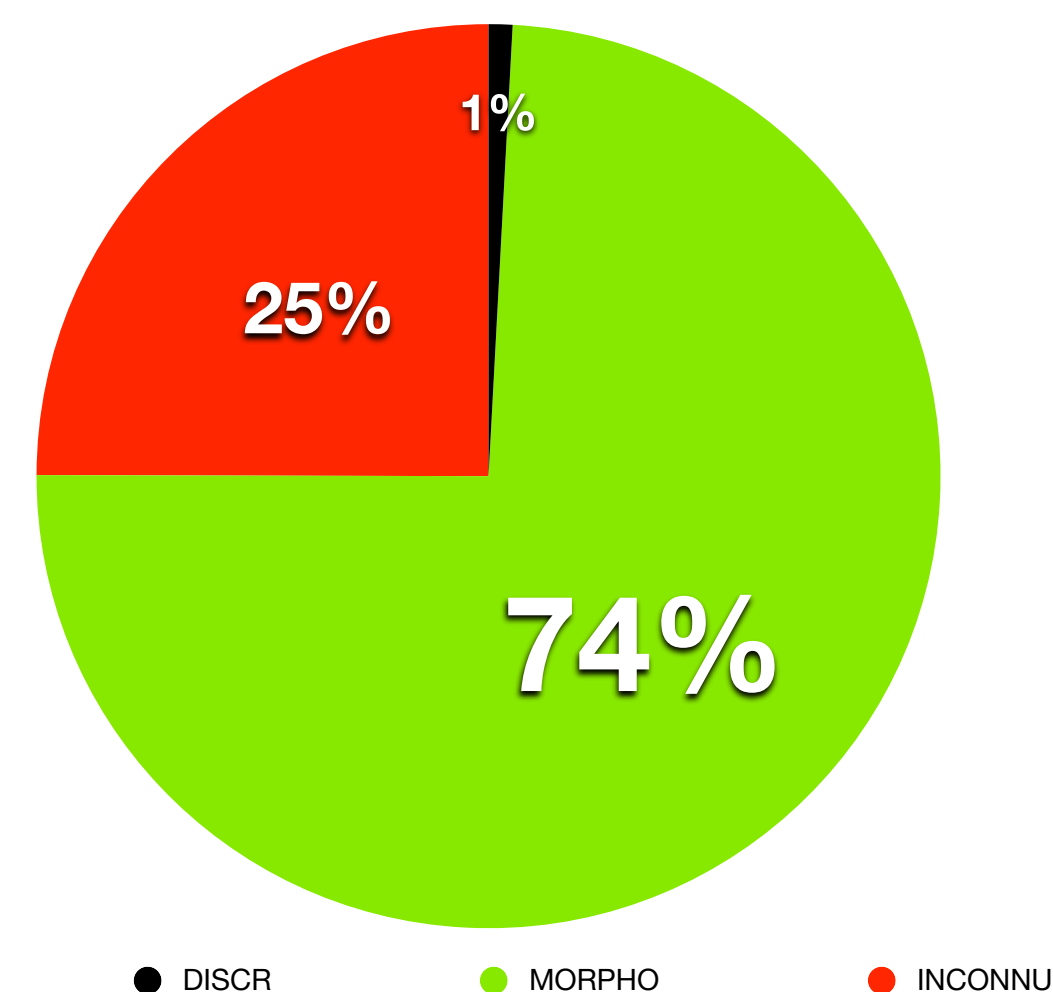


# Horizontální interpretace externí (bez omezení)

Horizontální interpretace externí

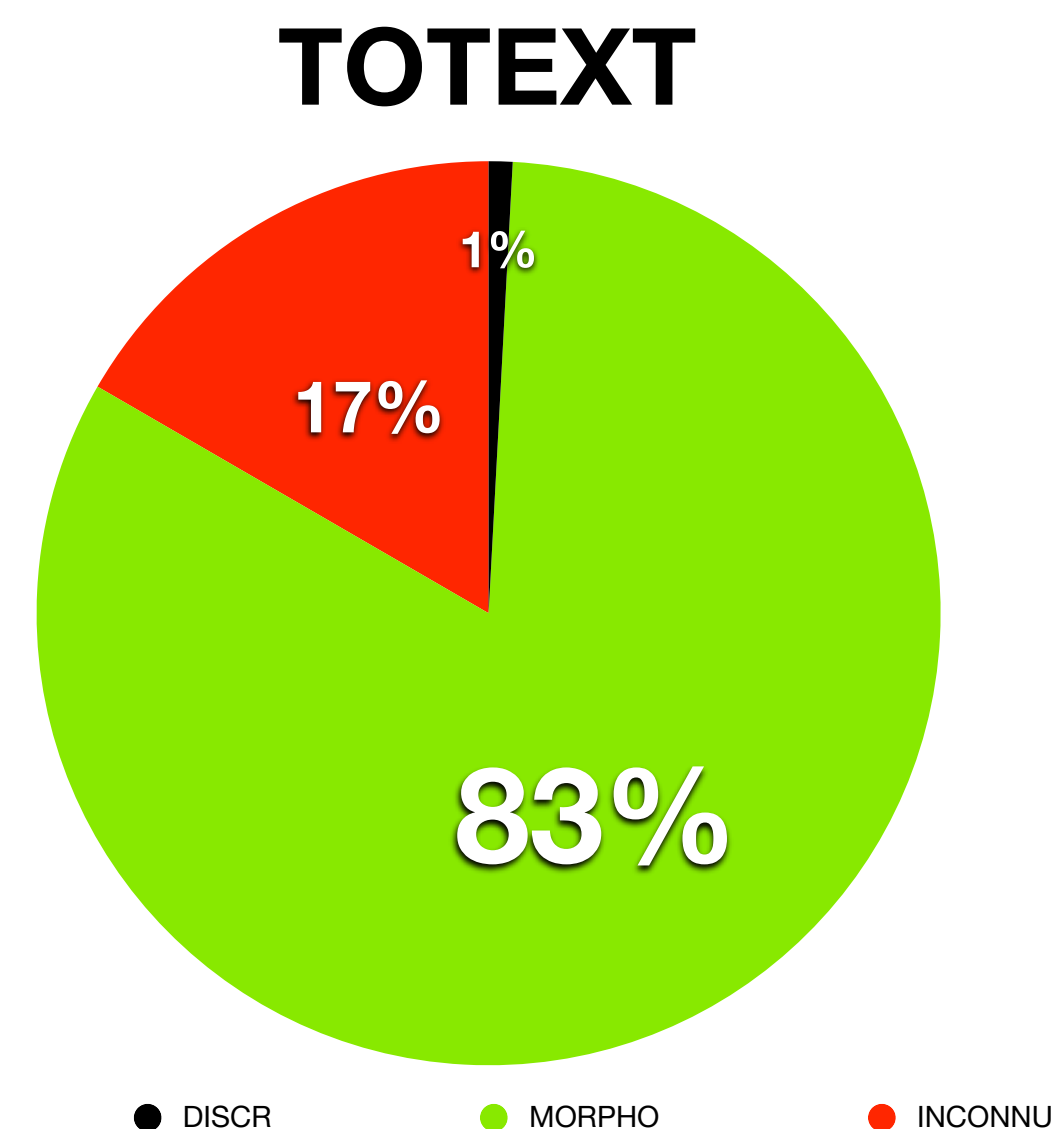
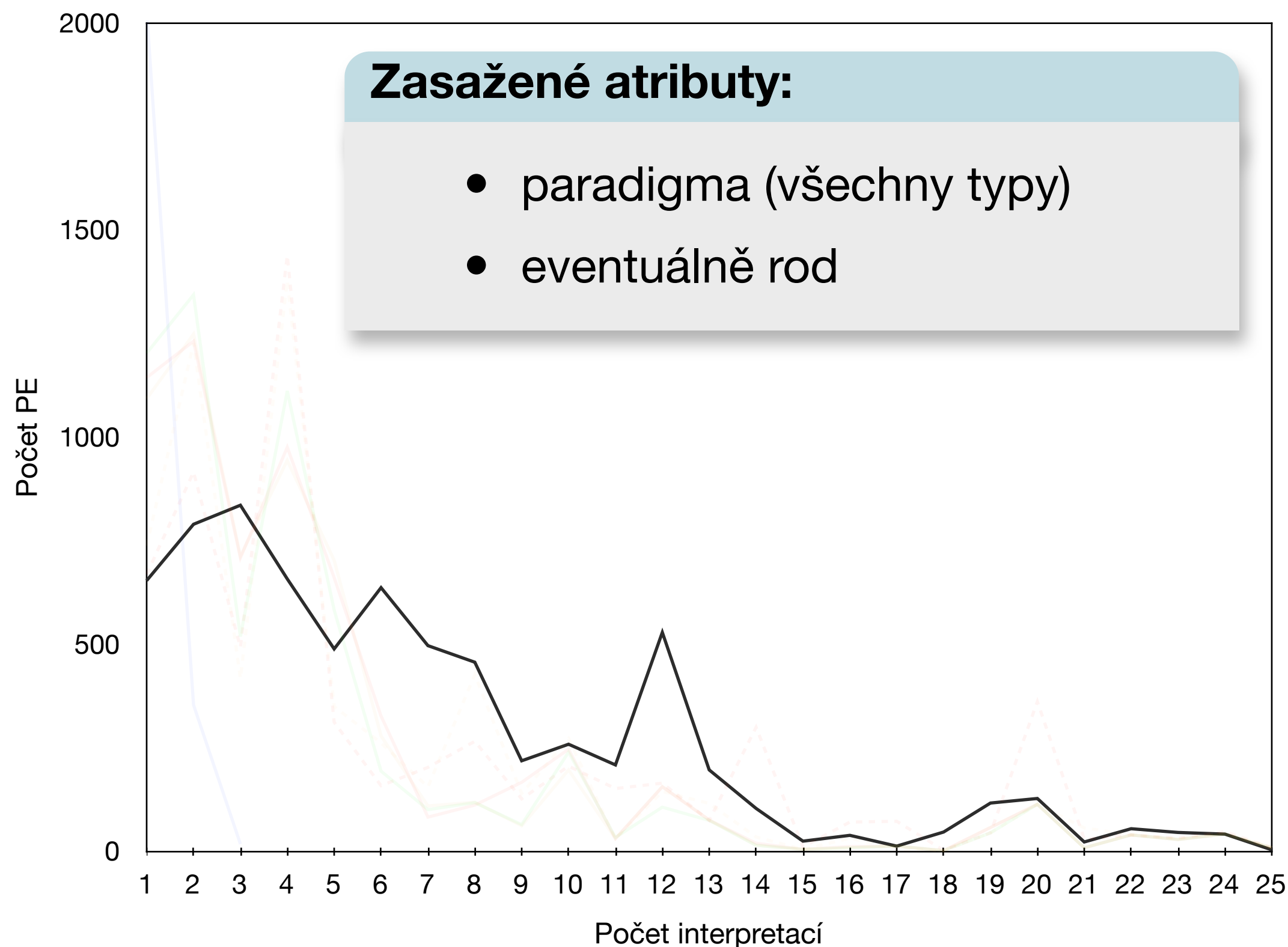


**HOREXT\_norestr**

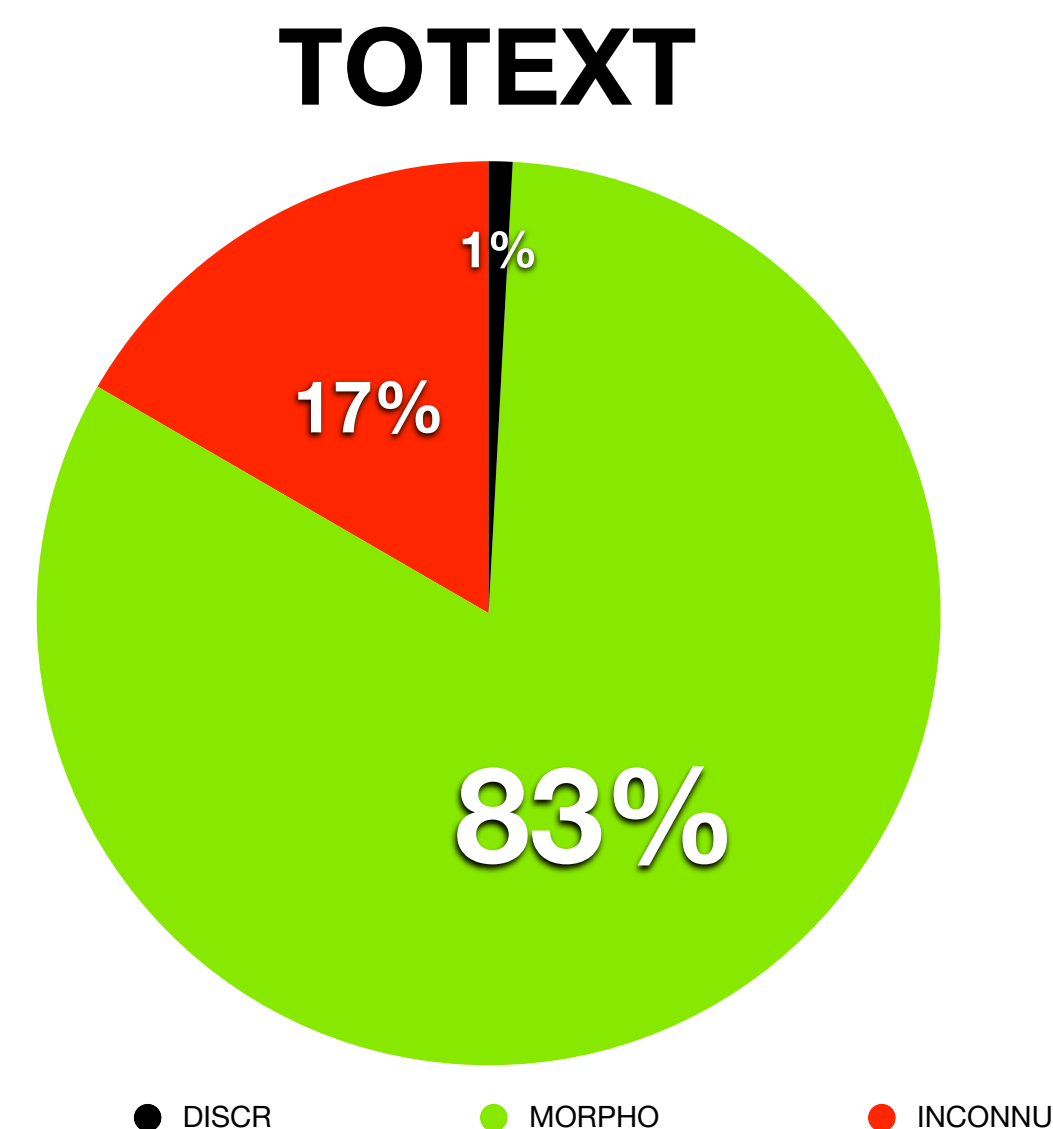
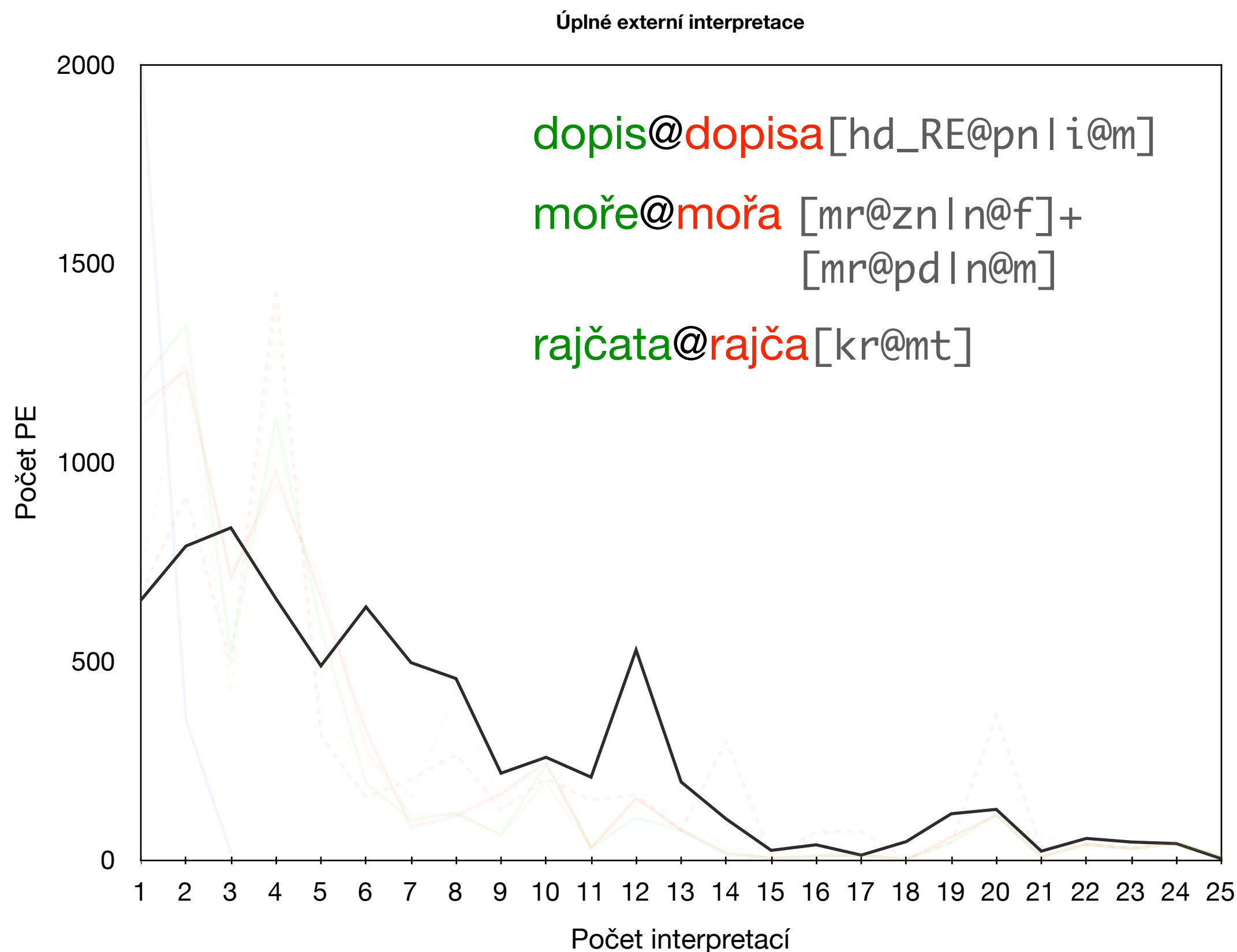


# Úplné externí interpretace

Úplné externí interpretace

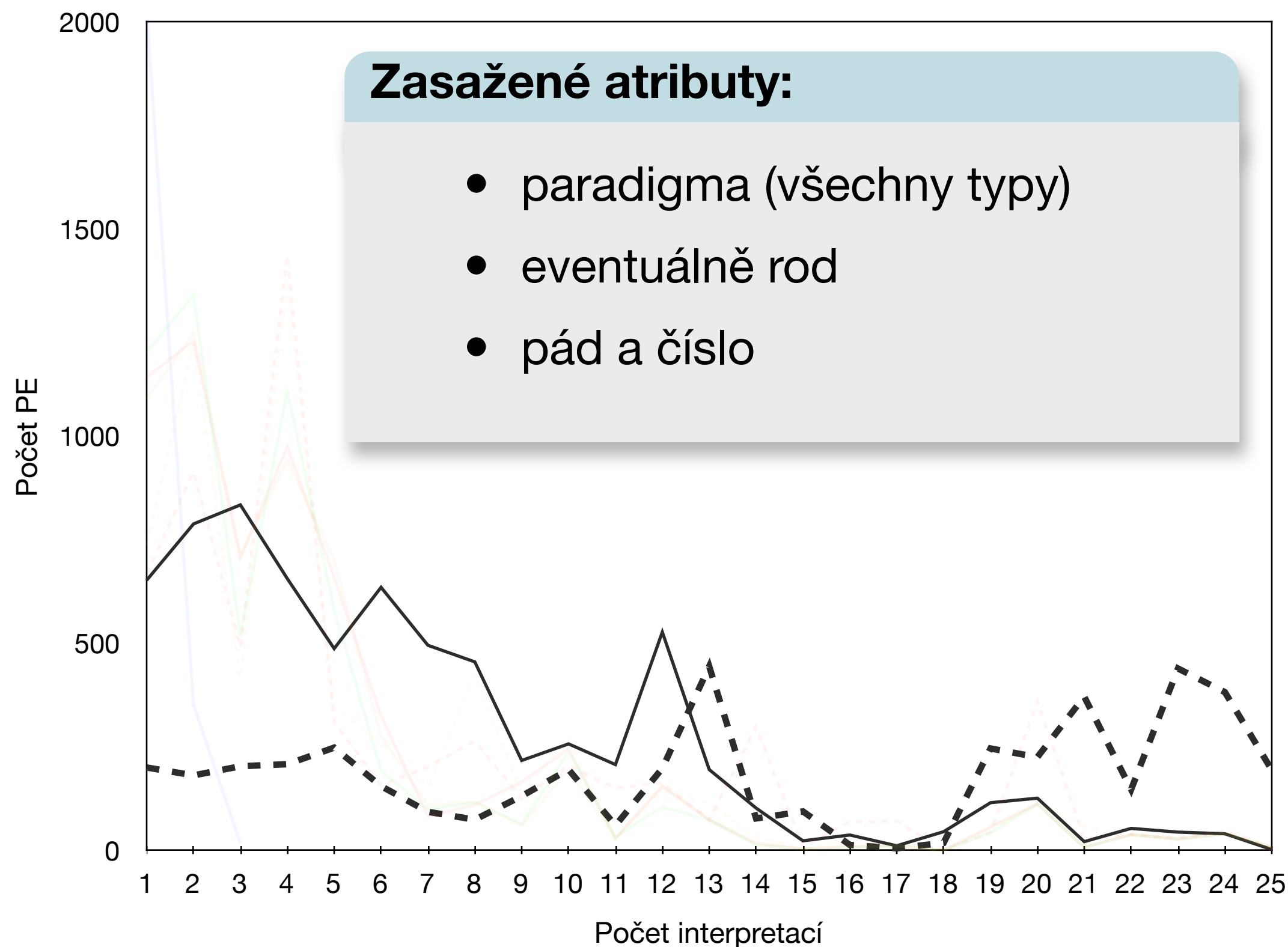


# Úplné externí interpretace

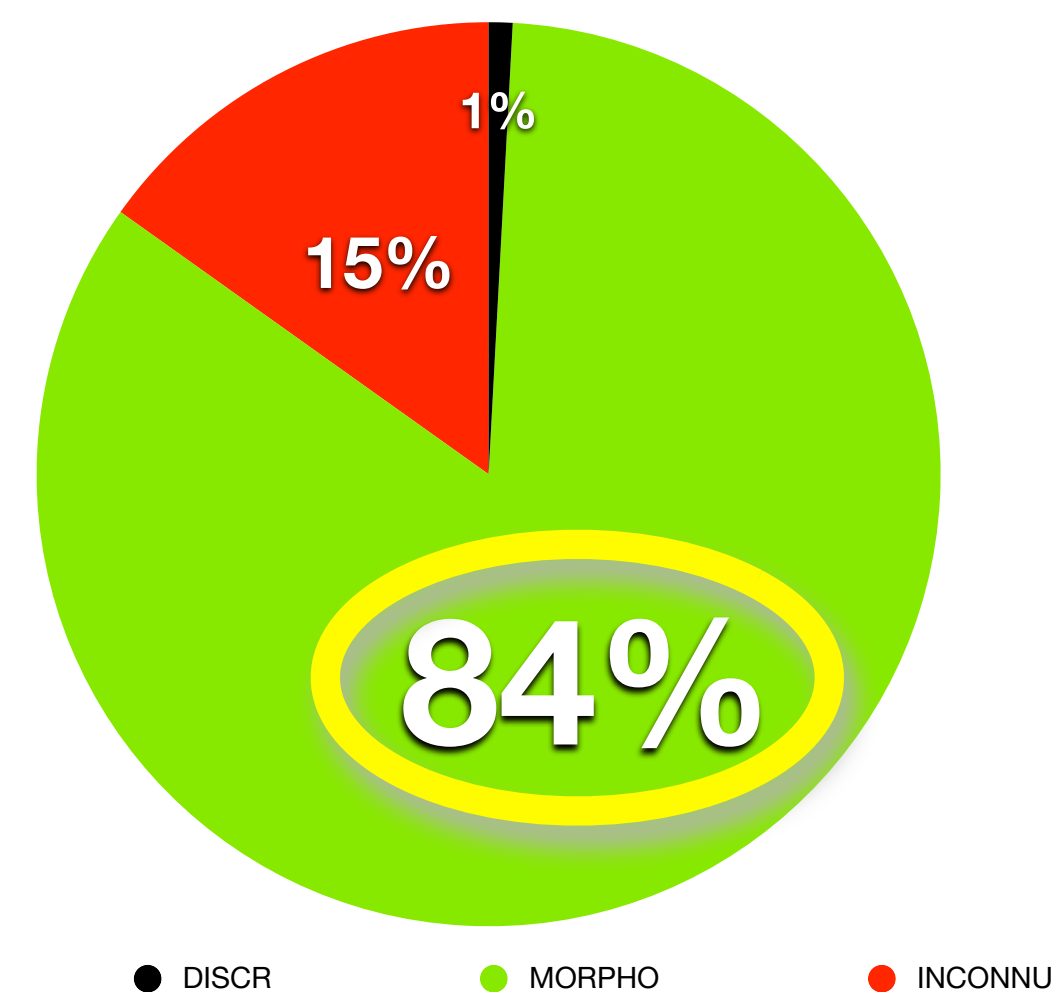


# Úplné externí interpretace (neomezené)

Úplné externí interpretace

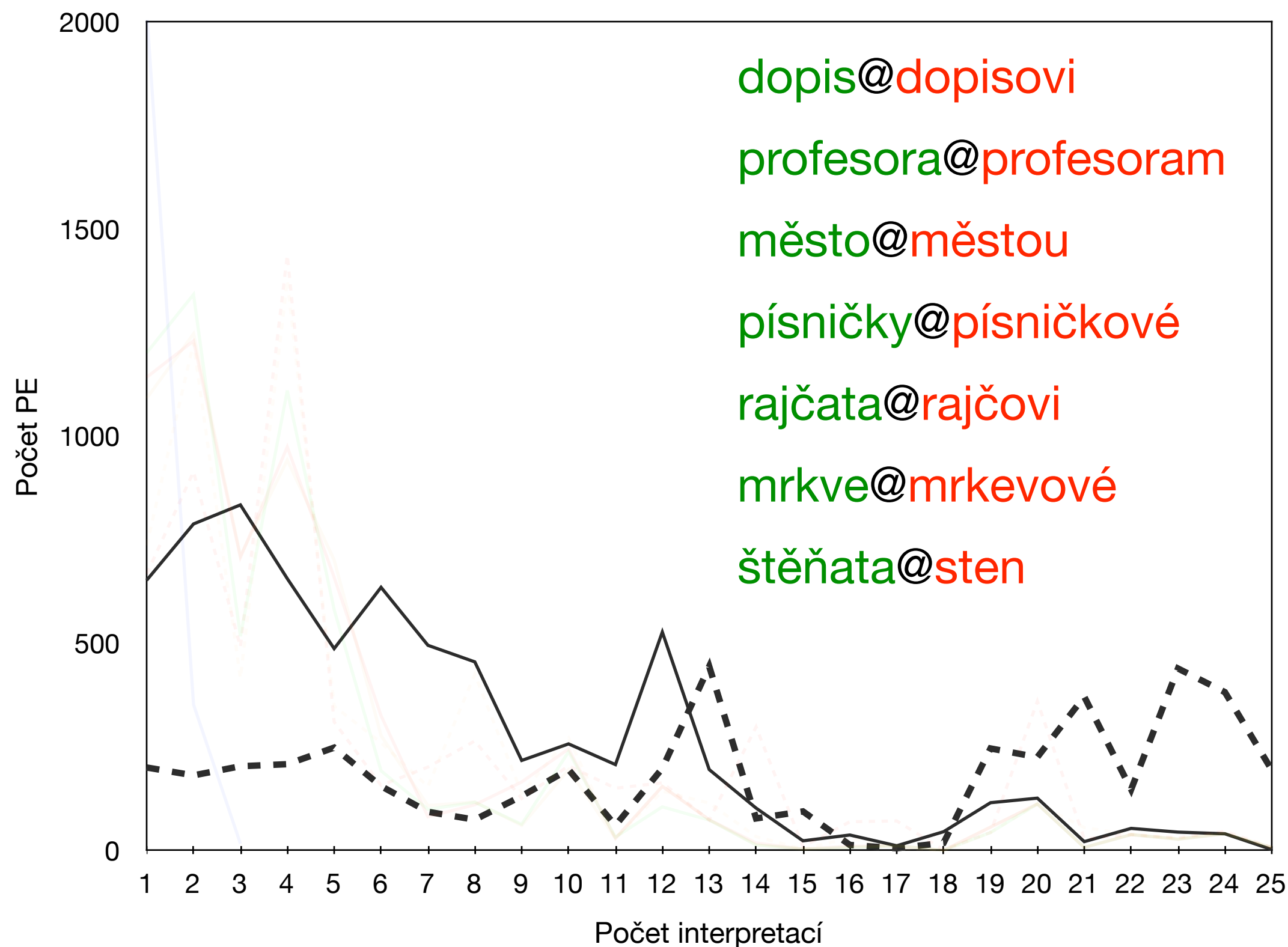


TOTEXT\_norestr

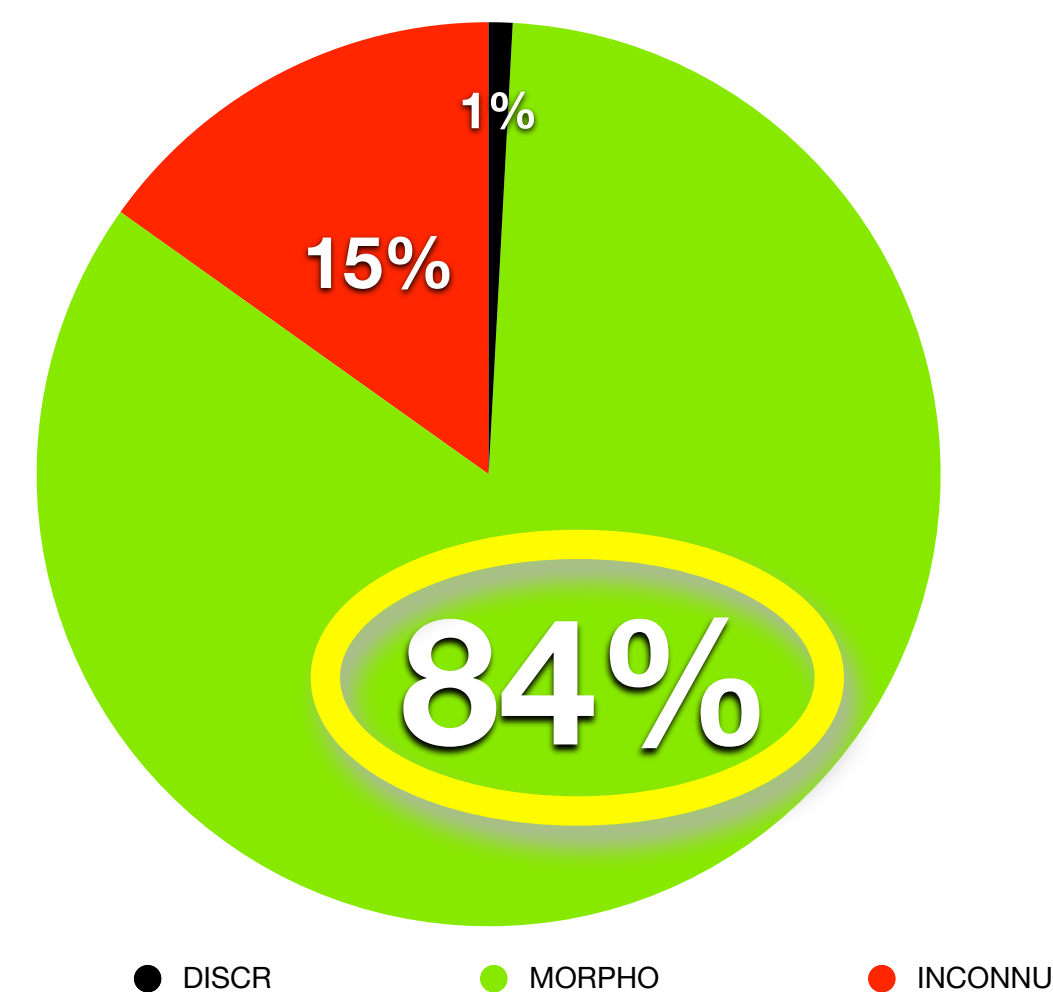


# Úplné externí interpretace (neomezené)

Úplné externí interpretace



TOTEXT\_norestr



# Nerozpoznané PE (15 %)

## (1) překlepy v kmenu i v koncovce (potenciální morfologická interpretace)

- ▶ přidání, odebrání, nahrazení znaku (Levenshteinova vzdálenost 1-2)

manželku@malzelku manželku@manželhu dopis@dopisp turistu@turistok

### Možné řešení

- ▶ regularizovat PE s překlepem v kmenu a porovnat s nejbližší hypotetickou formou

## (2) lexikální chyby

- ▶ pozorovatelná lexikální motivace chyby
- ▶ záměna flexe a derivace (*deminutiva, přechýlené formy, adjektiva odvozená od substantiv, atd.*)

koláč@koláček Francouz@francouzský profesor@profesorka kuře@kuřecí kostel@kostelků

# Nerozpoznané PE (15 %)

## (3) chyby způsobené vlivem uživatelského rozhraní

- ▶ zadání produkce od nesprávné FR

kavárny@Prahya   vína@vínasvač   strom@stromylék

- ▶ parazitní řetězce

vína@vinosg

## (4) jiné, zatím neprozkoumané ...

manželkamě   přitalé  
přteléá

# Filtrace interpretací

**(1) Komprese hodnot** pro identický atribut zasažený v několika interpretacích

| <i>stavením</i>   | $\neq$ | <i>stavení</i>                       |
|-------------------|--------|--------------------------------------|
| <i>cas : inst</i> |        | <i>cas : nom gen dat acc voc loc</i> |
| <i>num : sg</i>   |        | <i>num : sg</i>                      |
| <i>gen : n</i>    |        | <i>gen : n</i>                       |
| <i>pdgm : st</i>  |        | <i>pdgm : st</i>                     |
| <i>alt : sans</i> |        | <i>alt : sans</i>                    |
| <i>dia : 0</i>    |        | <i>dia : 0</i>                       |

VERT ; st ; **inst@nom|gen|dat|acc|voc|loc** ; sg ; n ; sans ; ok ; 0 ;

# Filtrace interpretací

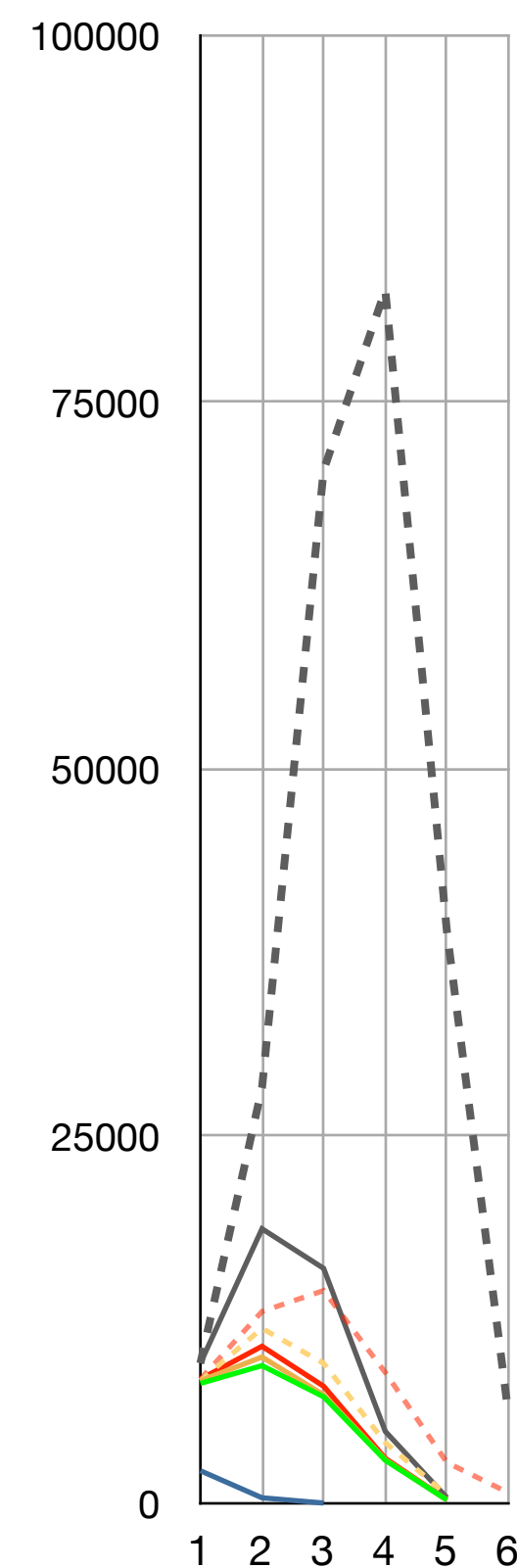
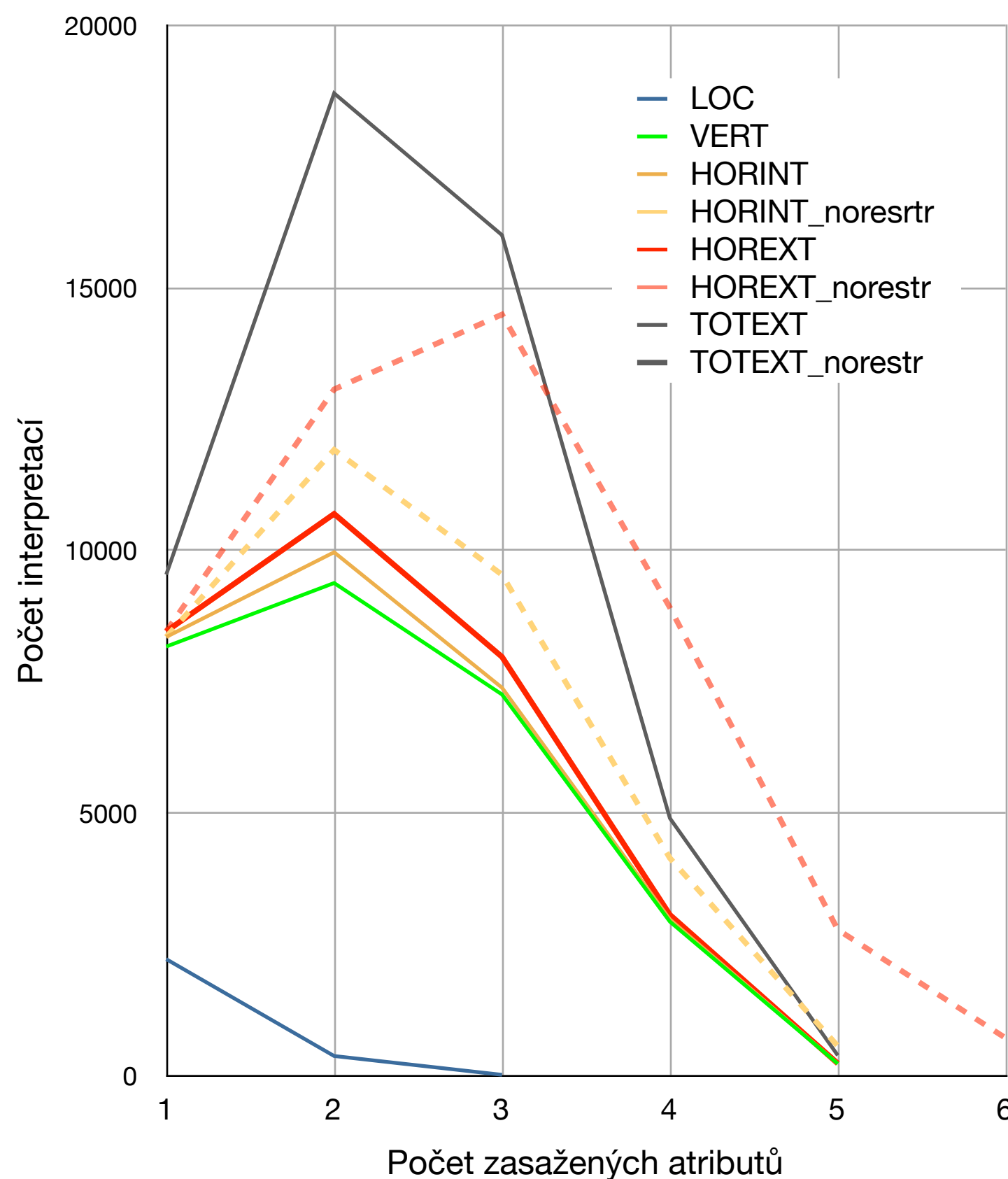
## (2) Počet atributů zasažených chybou

\* interpretace, která obsahuje nejmenší počet atributů zasažených chybou je považována za pravděpodobnější

| <i>Olze</i>                  | $\neq$ | <i>Olge</i>              |   | <i>Olze</i>                  | $\neq$ | <i>Olge</i>              |
|------------------------------|--------|--------------------------|---|------------------------------|--------|--------------------------|
| <i>cas : dat</i>             |        | <i>cas : dat</i>         |   | <b><i>cas : dat</i></b>      |        | <b><i>cas : loc</i></b>  |
| <i>num : sg</i>              |        | <i>num : sg</i>          |   | <i>num : sg</i>              |        | <i>num : sg</i>          |
| <i>gen : f</i>               |        | <i>gen : f</i>           |   | <i>gen : f</i>               |        | <i>gen : f</i>           |
| <b><i>alt : g &gt; z</i></b> |        | <b><i>alt : sans</i></b> | + | <b><i>alt : g &gt; z</i></b> |        | <b><i>alt : sans</i></b> |
| <i>pdgm : zn_Re</i>          |        | <i>pdgm : zn_Re</i>      |   | <i>pdgm : zn_Re</i>          |        | <i>pdgm : zn_Re</i>      |
| <i>dia : 1</i>               |        | <i>dia : 1</i>           |   | <i>dia : 1</i>               |        | <i>dia : 1</i>           |

LOC ; zn\_Re ; dat ; sg ; f ; **k > c@sans** ; ok ; 0 ;

# Počet zasažených atributů v různých interpretacích



# Pertinence rysu pro diagnostiku

## (3) Pertinence rysu pro diagnostiku (pro stejný počet zasažených atributů)

- ▶ číslo > rod a vzor > pád
- ▶ diakritika ? ostatní atributy (např. *Každou neděli dávám ženě růží, jmen@jmén*)

Toho *\*pese* se nemusíš bát.  
pes sg. *psa*

*psa* ≠

|                    |
|--------------------|
| <i>tagLex</i> : N  |
| <i>cas</i> : gen   |
| <i>num</i> : sg    |
| <i>gen</i> : ma    |
| <i>pdgm</i> : pn   |
| <i>alt</i> : e > # |

| ① <i>pese</i>     | ② <i>pese</i>     |
|-------------------|-------------------|
| <i>tagLex</i> : N | <i>tagLex</i> : N |
| <b>cas: voc</b>   | <i>cas</i> : gen  |
| <i>num</i> : sg   | <i>num</i> : sg   |
| <i>gen</i> : ma   | <i>gen</i> : ma   |
| <i>pdgm</i> : pn  | <b>pdgm : mz</b>  |
| <b>alt : sans</b> | <b>alt : sans</b> |

# Pertinence rysu pro diagnostiku

## (3) Pertinence rysu pro diagnostiku (pro stejný počet zasažených atributů)

- ▶ *číslo > rod a vzor > pád*
- ▶ *diakritika ? ostatní atributy* (např. *Každou neděli dávám ženě růží, jmen@jmén*)

Vzít v úvahu **syntaktická kritéria**:

- ▶ dát přednost interpretaci ***chyba v pádě*** před jinou interpretací

Vzít v úvahu **gramatický rod francouzského ekvivalentu FR**:

- ▶ dát přednost interpretaci ***chyba v rodě*** před jinou interpretací
- ▶ časté interference z francouzštiny:  
***sůl*** f. 'le sel' / ***stůl*** ma. 'la table' / ***kuře*** n. 'le poulet' / ***moře*** n. 'la mer'

# Frekvence jednotlivých typů chyb po filtraci

| jediná interpretace            |      |  | dvě interpretace |               |      |
|--------------------------------|------|--|------------------|---------------|------|
| typ chyby                      | %    |  | typ chyby        |               | %    |
| locale                         | 28,9 |  | cas              | genre         | 7,0  |
| cas                            | 26,8 |  | locale           | cas           | 5,4  |
| nombre                         | 9,8  |  | type pdgm        | genre         | 2,7  |
| genre                          | 5,4  |  | nombre           | genre         | 2,1  |
| type pdgm                      | 4,3  |  | nombre           | type pdgm     | 1,2  |
| cas et nombre                  | 3,7  |  | locale           | nombre        | 0,9  |
| Total                          | 78,9 |  | cas              | type pdgm     | 0,5  |
|                                |      |  | cas et nombre    | genre         | 0,4  |
|                                |      |  | cas et nombre    | type pdgm     | 0,3  |
| chyby v diakritice a alternaci |      |  | locale           | cas et nombre | 0,3  |
| typ chyby                      | %    |  | locale           | cas et genre  | 0,3  |
| diacritique                    | 32,6 |  | Total            |               | 21,1 |
| alternance                     | 12,1 |  |                  |               |      |

# Feedback

1 Francouzští **studentové** **studenti** se učí česky a jde jim...

2 Všichni **pilní** žáci jsou ve škole a studují gramatiku.

3 **Studentky** čtou knihu a překládají některé pasáže.

4 Moji **spolužáky** **spolužáci** se chtějí učit česky, rusky,

5 Některí **žáci** studují francouzštinu univerzitní knihovně.

6 Závodu **žáci** se učí česky, rusky, Chuchli v Praze.

7 Elegantní **páni** nosí klobouk.

8 Naši **profesoři** jsou nároční.

9 V jižních Čechách jsou krásná historická **město** **města** – například Třeboň, Český Krumlov nebo Tábor.

10 Petr a Pavel jsou nerozluční **přítelé** **přátelé**.

**erreur** ⚠

Il y a plusieurs interprétations possibles de votre erreur :

1) Le cas et le nombre sont corrects mais peut être que vous vous trompez de genre et que vous déclinez le mot d'après le type *hrad*.

2) Le nombre et le genre sont corrects mais il y a une erreur de cas.

**erreur** ⚠

Il y a plusieurs interprétations possibles de votre erreur :

1) Le cas et le nombre sont corrects mais peut être que vous vous trompez de genre et que vous déclinez le mot d'après le type *hrad*.

2) Le nombre et le genre sont corrects mais il y a une erreur de cas.

- Co je pro studenta pertinentní?
- Stručný nebo obsažný popis?
- Čtou to?

# Shrnutí

- relativně jednoduchá metoda s relativně **uspokojivými výsledky**
- sebraný korpus chyb umožňuje **vyhodnocovat** diagnostiku a dává podněty k jejímu zdokonalování
- **detailní analýza** jednoho úseku české morfologie v produkcích nerodilých mluvčích, která by mohla být užitečná pro zpracování **volných produkcí** (morfologická a syntaktická analýza) a pro výzkum **osvojování češtiny**

# Perspektivy automatické diagnostiky

- **internacionalizace** uživatelského rozhraní platformy CETLEF (anglická verze)
- **zveřejnění korpusu chyb** s chybovou anotací.
- rozšíření diagnostiky na českou **konjugaci** a pravidelnou **derivaci**
- hledání řešení pro vytvoření **jazykově nezávislé metody** (práce na slovenštině)

# Děkuji za pozornost

## Bibliografie

- BEESLEY, Kenneth R. and KARTTUNEN, Lauri (2003). Finite-State Morphology, Stanford, California.
- HEIFT, Trude & SCHULZE, Mathias (2007). Errors and Intelligence in Computer-Assisted Language Learning: Parsers and Pedagogues. UK : D. Reidel Publishing Company.
- MEURERS, Detmar (2012). Natural Language Processing and Language Learning in Chapelle, ed. Encyclopedia of Applied Linguistics, Blackwell
- MEURERS, Detmar (2009). On the Automatic Analysis of Learner Language. Introduction to the Special Issue. CALICO Journal, 26(3):469–473.
- NERBONNE, John (2003). Natural language processing in computer-assisted language learning, in The Oxford Handbook of Computational Linguistics, Mitkov, Ruslan (Ed). Oxford University Press, 670–698.
- ŠMILAUER, Ivan (2011). Diagnostic et désambiguïsation des erreurs de flexion dans un outil ELIAO. In Autour des langues et du langage, Perspectives pluridisciplinaires, Presses Universitaire de Grenoble, 19-26.
- ŠMILAUER, Ivan (2008). Acquisition du tchèque par les francophones : analyse automatique des erreurs de déclinaison. PBML, 90:33-56.