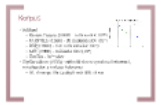
Počítačová lexikografie
Makrostruktura
Adam Rambousek

Podklady

Makrostruktura

Počítačová lexikografie
Makrostruktura

Adam Rambousek

Podklady

Lexikografické podk

Lexikografické podklady

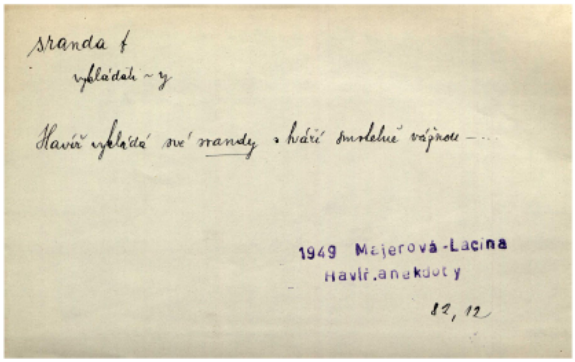
- důkazy o použití jazyka
  - intuice
  - excerpta, výpisky
  - korpusy
- intuice (armchair linguistics)

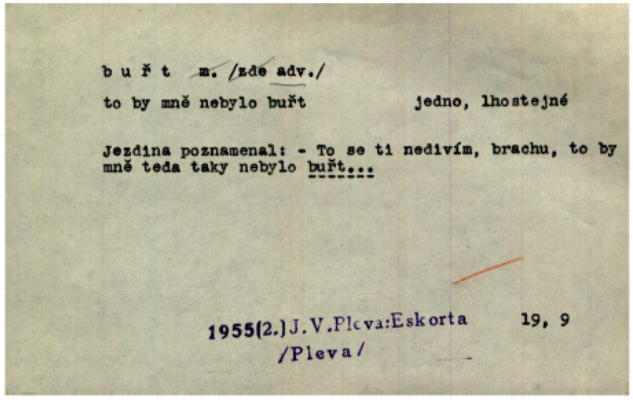
Intuice

- *In the absence of objective evidence, introspection was appealed to instead. But studies in corpus linguistics have shown that introspection is a very flawed technique. We human beings are wired to register the unusual in our minds, generally in a way that is available to conscious recall. But we fail to pay any attention to the commonplace patterns of usage on which we rely so heavily in our everyday communications.* Patrick Hanks (Euralex 2000)
- *Should it ever come about that linguistics can be carried out without the intervention and suffering of a native-speaker analyst, I will probably lose interest in the enterprise.* Charles Fillmore ("Corpus linguistics" or "Computer-aided armchair linguistics")

Výpisky

- *Appeal to the English-speaking and English-reading public*, 1879
- *Návod pro sběratele materiálu k "Slovníku jazyka českého"*, 1911
  - 8 696 850 lístků (1911-1991), neologismy 270 538 záznamů





sranda f

vykládati ~ y

Havíř vykládá své srandy a káří smrtelné vápnou – …

1949 Majerová-Lacina

Havlí.anekdoty

82, 12

b u ř t m. /zde adv./

to by mně nebylo buřt jedno, lhostejné

Jezdina poznamenal: - To se ti nedivím, brachu, to by mně teda taky nebylo buřt...

1955(2.) J.V.Pleva:Eskorta 19, 9

/Pleva/

Detail hesla

Bleskový filtr

Heslo	myš	Slovní druh	podst ž
Podheslí	myš (polohovací to zařízením)		
Titulek	Sbohem Simpsonovi!	Rubrika	Život a vůbec
Autor	Ivan Straka	Postavení autora	redaktor
Mluvčí		Postavení mluvčího	
Kontext	Na redakční poradě bylo odbojným redaktorům nařízeno rozkazem, že musím používat v souvislosti s myší sloveso "poklepat" (kliknout či cvaknout v žádném případě). Byla šance oživit českou terminologii, ať už fonetickým bastardem "kliknout" či ryze českým bufeťáckým "cvaknout". Zvítězilo absurdní "poklepat", ač činnost, kterou s myší (polohovacím to zařízením) provádíme, má k poklepání tak daleko jako Mohamed k hoře.		
Poznámka			
Zdroj	Softwarové noviny	Datum	16.10.2013
Číslo	10	Strana	132	Rok	1995

Tweet 0

Error

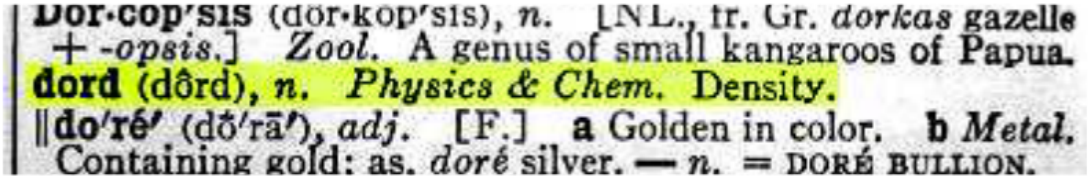
Exportovat

Výpisky

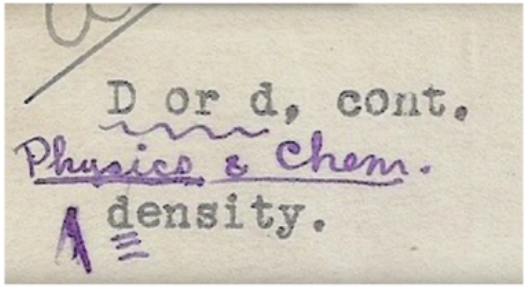
- výhody
  - posuny významu
  - terminologie
  - šíření lexikografie
- nevýhody
  - pracné, časově náročné
  - subjektivní (časté výjimky)

Ghost word

- dord
  - *Webster's New International Dictionary 2nd ed.* (1934)



Dor·cop'sis (dôr·kop'sis), *n.* [NL., fr. Gr. *dorkas* gazelle + *-opsis*.] *Zool.* A genus of small kangaroos of Papua.
dord (dôrd), *n.* *Physics & Chem.* Density.
‖**do'ré'** (dŏ'rā'), *adj.* [F.] **a** Golden in color. **b** *Metal.* Containing gold: as, *doré* silver. — *n.* = DORÉ BULLION.



D or d, cont.
Physics & Chem.
density.

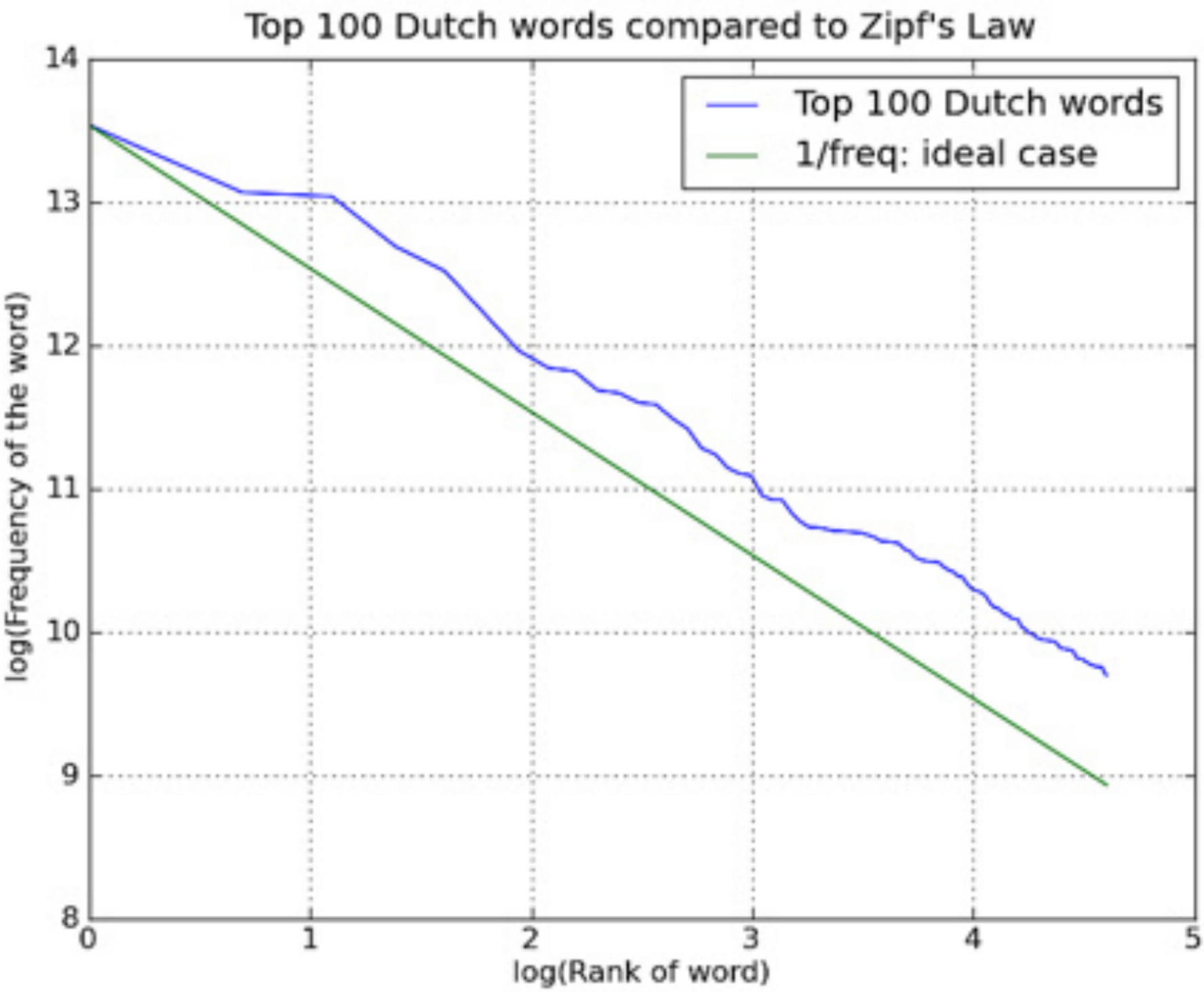
  - 1939 objeveno, 1940 smazáno v knize

Korpus

- IB047 Úvod do korpusové lingvistiky
- *a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language as a source of data for linguistic research*
- dokonalý korpus neexistuje
  - korpus je jen vzorek jazyka
  - obsahuje i nespisovný jazyk
  - čas a náklady na výrobu

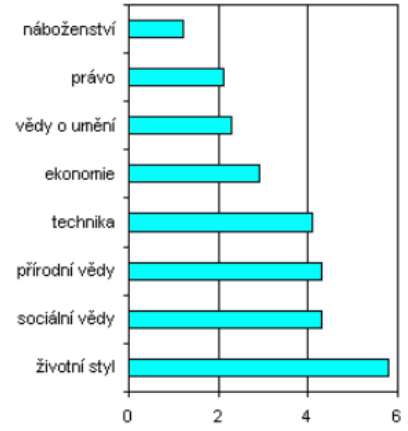
Korpus

- velikost
  - *Brown Corpus* (1960) - milion slov (10^6)
  - *COBUILD* (1980) - 20 milionů slov (10^7)
  - *BNC* (1990) - 100 milionů slov (10^8)
  - *OEC* (2000) - miliarda slov (10^9)
  - *TenTen* - 10^{10} slov
- Zipfův zákon (1935) - několik slov s vysokou frekvencí, mnoho slov s nízkou frekvencí
  - 10. slovo je 10x častější než 100. slovo


Top 100 Dutch words compared to Zipf's Law
Top 100 Dutch words
1/freq: ideal case
log(Frequency of the word)
log(Rank of word)
14
13
12
11
10
9
8
0
1
2
3
4
5

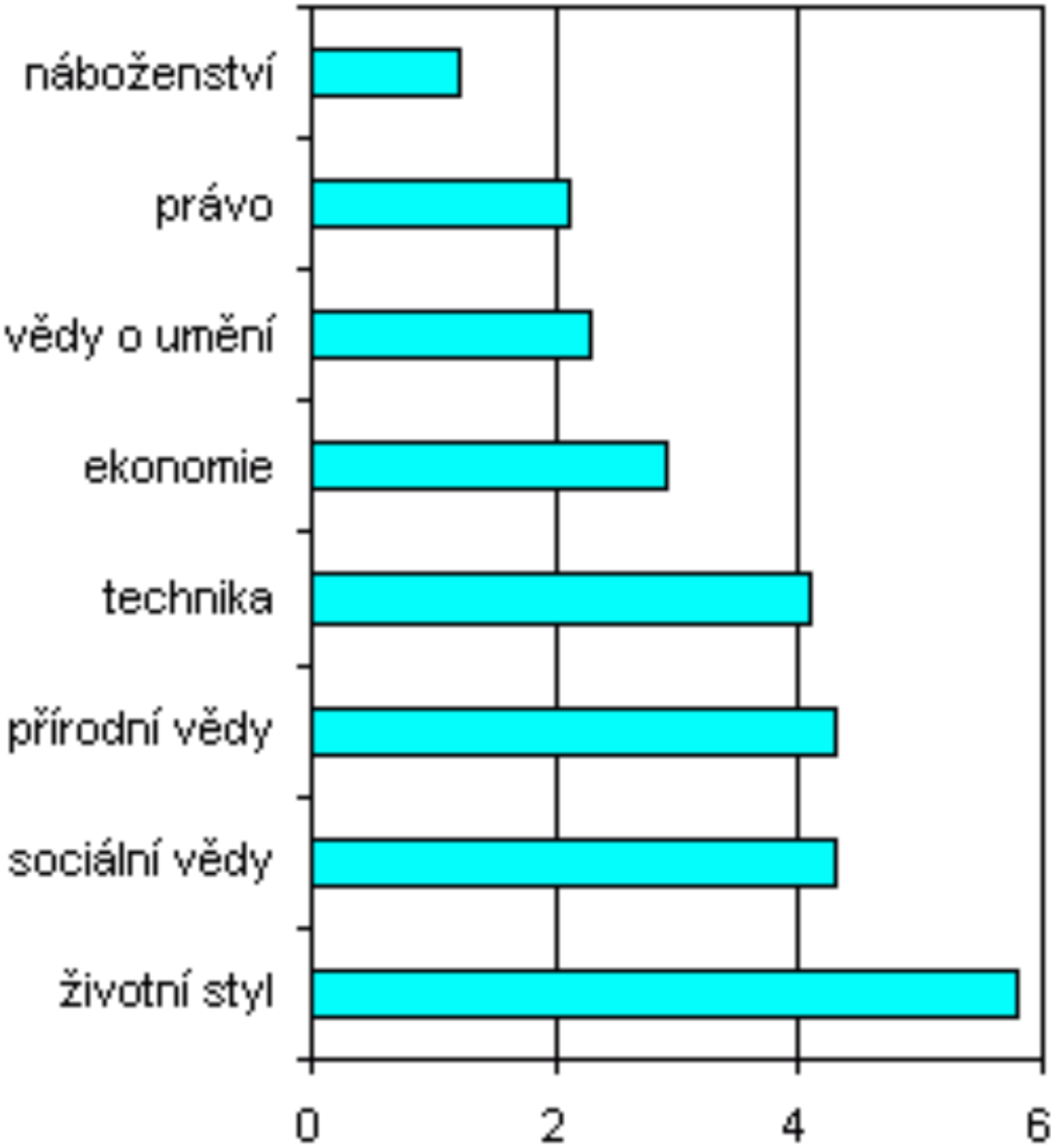
Korpus

- vyvážený
  - *Linguistic Data Consortium* - anglický korpus z článků Associated Press a New York Times
  - *BNC* - Journal of Gastroenterology, mucosa x unfortunate
- co zahrnout a v jakém poměru?
- BNC
  - 90% written, 10% spoken; 75% informative, 25% imaginative
- SYN2000 (100 milionů slov)
  - 60% publicistika, 25% odborná, 15% beletrie
- SYN2005 (100 milionů slov)
  - 40% beletrie, 27% odborná, 33% publicistika



e

a

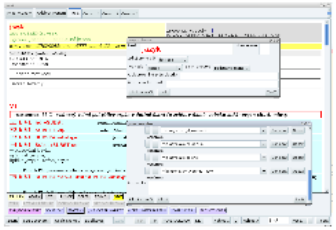

náboženství
právo
vědy o umění
ekonomie
technika
přírodní vědy
sociální vědy
životní styl
0
2
4
6

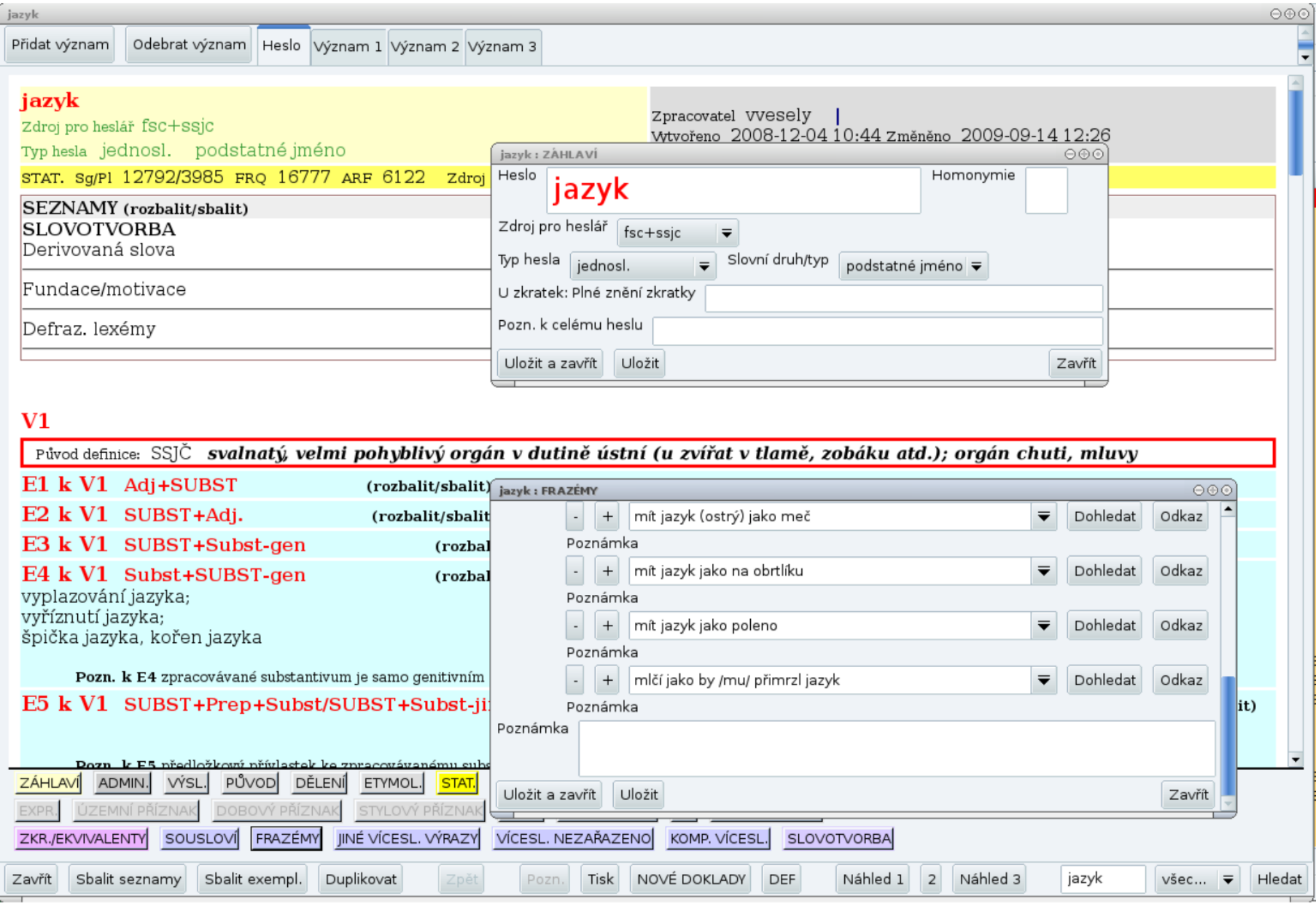
Korpus

- získání
- převod
- značkování (formální, lingvistické)
- z webu: Sketch Engine, WebBootCaT

Lexikální databáze

- podrobná strukturovaná jazyková databáze
  - (nyní obvykle) doklady z korpusu
  - gramatické údaje
  - valence, vzory
  - styl, užití, oblast...
  - vztahy mezi slovy
- podklad pro slovníky a výzkum
- *PraLeD* (Pražská Lexikální Databáze)
- *DANTE* (Database of ANalysed Texts of English)




jazyk
Přidat význam
Odebrat význam
Heslo
Význam 1
Význam 2
Význam 3
jazyk
Zdroj pro heslář fsc+ssjc
Typ hesla jednosl. podstatné jméno
Zpracovatel vvesely
Vytvořeno 2008-12-04 10:44 Změněno 2009-09-14 12:26
STAT. Sg/Pl 12792/3985 FRQ 16777 ARF 6122 Zdroj
SEZNAMY (rozbalit/sbalit)
SLOVOTVORBA
Derivovaná slova
Fundace/motivace
Defraz. lexémy
jazyk : ZÁHLAVÍ
Heslo
jazyk
Homonymie
Zdroj pro heslář
fsc+ssjc
Typ hesla
jednosl.
Slovní druh/typ
podstatné jméno
U zkratek: Plné znění zkratky
Pozn. k celému heslu
Uložit a zavřít
Uložit
Zavřít
V1
Původ definice: SSJČ svalnatý, velmi pohyblivý orgán v dutině ústní (u zvířat v tlamě, zobáku atd.); orgán chuti, mluvy
E1 k V1 Adj+SUBST (rozbalit/sbalit)
E2 k V1 SUBST+Adj. (rozbalit/sbalit
E3 k V1 SUBST+Subst-gen (rozbal
E4 k V1 Subst+SUBST-gen (rozbal
vyplazování jazyka;
vyřínutí jazyka;
špička jazyka, kořen jazyka
Pozn. k E4 zpracovávané substantivum je samo genitivním
E5 k V1 SUBST+Prep+Subst/SUBST+Subst-ji
jazyk : FRÁZÉMY
mít jazyk (ostrý) jako meč
Dohledat
Odkaz
Poznámka
mít jazyk jako na obrtlíku
mít jazyk jako poleno
mlčí jako by /mu/ přimrzl jazyk
Poznámka
Uložit a zavřít
Uložit
Zavřít
ZÁHLAVÍ
ADMIN.
VÝSL.
PŮVOD
DĚLENÍ
ETYMOL.
STAT.
EXPR.
ÚZEMNÍ PŘÍZNAK
DOBOVÝ PŘÍZNAK
STYLOVÝ PŘÍZNAK
ZKR./EKVIVALENTY
SOUSLOVÍ
FRAZÉMY
JINÉ VÍCESL. VÝRAZY
VÍCESL. NEZAŘAZENO
KOMP. VÍCESL.
SLOVOTVORBA
Zavřít
Sbalit seznamy
Sbalit exempl.
Duplikovat
Zpět
Pozn.
Tisk
NOVÉ DOKLADY
DEF
Náhled 1
2
Náhled 3
jazyk
všec...
Hledat

game

1 n a structured, essentially non-physical, activity involving one or more people, esp one engaged in for enjoyment or to pass the time

COLLOCATE TYPE OBJECT OF **COLLOCATES** **play**

- ➪ *Then , with wind howling round the windows and rattling the doors , we sat in front of a peat fire and played* **games** *with the children .*
- ➪ *Bullseye (07 February 2006) A great* **game** *on TV but a pretty awful DVD game .*

STRUCTURE **PP_X of**

- ➪ *It follows that the* **game** *of chess , in its effects upon mental character , is greatly misunderstood .*
- ➪ *England and Dublin at the time had started the* **game** *of bingo and the idea reached Graiguecullen in the mid ' 40s .*
- ➪ *The oldest game still played seriously in pubs is undoubtedly the* **game** *of darts .*

CHUNK game of chance

- ➪ *Description : Sic Bo , meaning dice pair , is an ancient Chinese* **game of chance** *played with three dice .*
- ➪ *His application for the chair of anatomy and botany was decided by drawing of lots and he was unlucky in this* **game of chance** *.*

STRUCTURE **N_mod**

COLLOCATE TYPE EQUIPMENT **COLLOCATES** **card, board, table, pen-and-paper**

- ➪ *Countess Spencer often stayed at Park House while Ruth , Lady Fermoy taught the children card* **games** *.*
- ➪ *At one time most games relevant to history were board* **games** *.*
- ➪ *For social interaction : painting , drawing , and* **table games** *such as dominos and cards .*
- ➪ *The new " games area " includes two for memory development (" Pairs " and " Copy Cat ") , a simple game called " Breakthrough " which helps wi which most children will be familiar .*

COLLOCATE TYPE PLACE/OCCASION OF PLAYING **COLLOCATES** **casino, party, parlour, pub**

- ➪ *Blackjack is by far our most popular casino* **game** *.*
- ➪ *Maybe some of our older readers can remember an old party* **game** *' hunt the thimble ' .*
- ➪ *The purser will also arrange activities during the day and evening , ranging from quizzes and parlour* **games** *to deck tennis or cricket .*
- ➪ *These include pub* **games** *against opposing teams to determine a winning side .*

COLLOCATE TYPE CONTENT/DESCRIPTION **COLLOCATES** **puzzle, guessing**

- ➪ *A puzzle* **game** *scrambles puzzle pieces for children to unscramble .*
- ➪ *Basically it is a guessing* **game** *involving coins .*

COLLOCATE TYPE PARTICIPANTS **COLLOCATES** **panel**

[TV-RAD] ➪ *David Baddiel has devised a new Radio 4 panel* **game** *, to be recorded next month .*

[TV-RAD] ➪ *Acknowledgement : This game is a simplified version of a UK TV* **quiz game** *called " The Weakest Link " .*

LINK ahead of the game **LINK** all part of the game **LINK** anybody's game **LINK** beat sb at their own game **LINK** give the game away **LINK** on the game **LINK** **LINK** the only game in town **LINK** game show **LINK** games room **LINK** game theory

Makrostruktura

Makrostruktura

- heslář (+předmluva, přílohy...)
- heslo[1] = lemma, entry term, heslové slovo, headword
  - obvykle nominativ sg., slovesa v infinitivu
  - části slov, spojení slov

Makrostruktura

- heslář (+předmluva, přílohy...)
- heslo[1] = lemma, entry term, heslové slovo, headword
  - obvykle nominativ sg., slovesa v infinitivu
  - části slov, spojení slov
- heslo[2] = heslová stať, entry

Heslář

- rozsah
- výběr podle oboru a typu
- obecný jazyk: frekvence

Heslář

- obecná slova
  - běžná slova (varianty)
  - zkratky
  - části slov
  - víceslovné výrazy
- vlastní jména
  - osoby, místa, metonymie, národnosti/skupiny, organizace, náboženství, předměty
- zkratky vs. plné názvy
- slovní spojení samostatně?

Heslář

- Achilles
- SSJČ: *jm. řeckého reka v Homérově Iliadě*: Achillova pata, *přen. zranitelné místo*; každý člověk má svou Achillovu patu; *med. Achillova šlacha upínající se na kost patní*;
- SSČ: Achillova pata, *zranitelné místo*; Achillova šlacha, *šlacha lýtkového svalu upínající se na patní kost*
- všechna slova v definici musejí být v hesláři