



Podklady



Makrostruktura





Počítačová lexikografie

Makrostruktura

Adam Rambousek

Podklady



Lexikografické podk

Lexikografické podklady

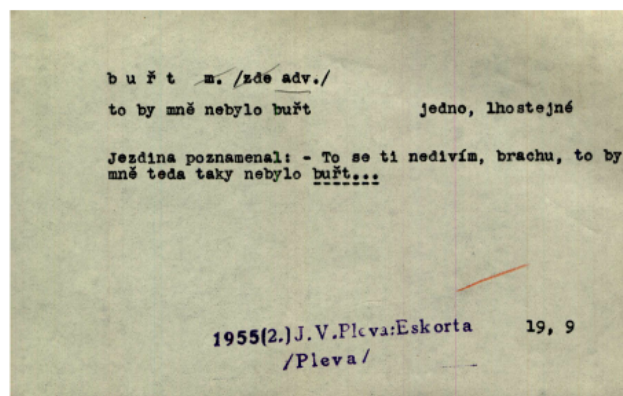
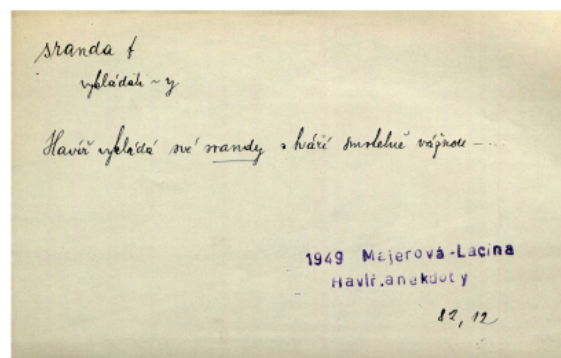
- důkazy o použití jazyka
 - intuice
 - excerpta, výpisky
 - korpusy
- intuice (armchair linguistics)

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



Detail hesla Biskovský 19h

Heslo	mysí	Slovní druh	podst. ž
Podhesí	mysí (pohovácí to zařízení)		
Titulek	Sběhem Simpsonost	Rubrika	Život a vůbec
Autor	Jan Straka	Postavení autora	redaktor
Mluvě		Postavení mluvčího	
Kontext	Na redakční perodě bylo obojím recenzentům nalžene rozkazem, že musíh použít v souvislosti s myší slovesa "poklepat" (nikoliv, či ovakov v žádném případě). Byla šance ožít českou terminologi, at už tovariským asistorem "Mikoum" či ryze českým buřtařským "ovaknout" buřtáři absureri "poklepat", až činnost, kterou s myší (pohovácím to zařízení) provádíme, má k poklepnání tak daleko jako Mohamed k hoře.		
Poznámka			
Zdroj	Setevosová noviny	Datum	16.10.2013
Číslo	10	Strana	132
		Rok	1995

[Tweet](#) Error Exportovat

sranda f

vykládati ~ y

Havíř vykládá své srandy s hárí smrtelně vážnou —...

1949 Majerová-Lacina

Havli. anekdoty

82, 12

b u ř t m. /zde adv./

to by mně nebylo buřt

jedno, lhostejné

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

1955(2.) J. V. Pleva: Eskorta
/Pleva/

19, 9

Bleskový filtr

Detail hesla

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

 Tweet

Error

[Exportovat](#)

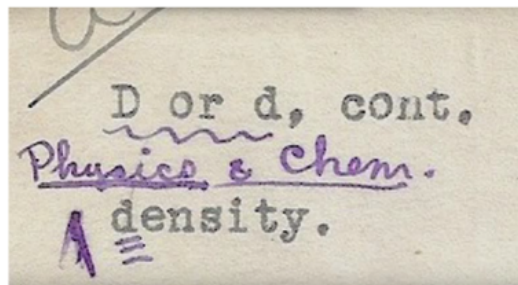
Výpisky

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



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

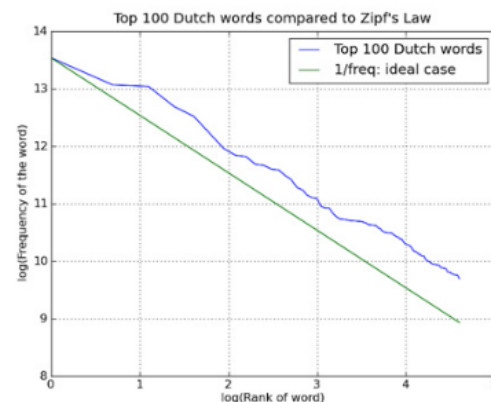
- 1939 objeveno, 1940 smazáno v knize

Korpus

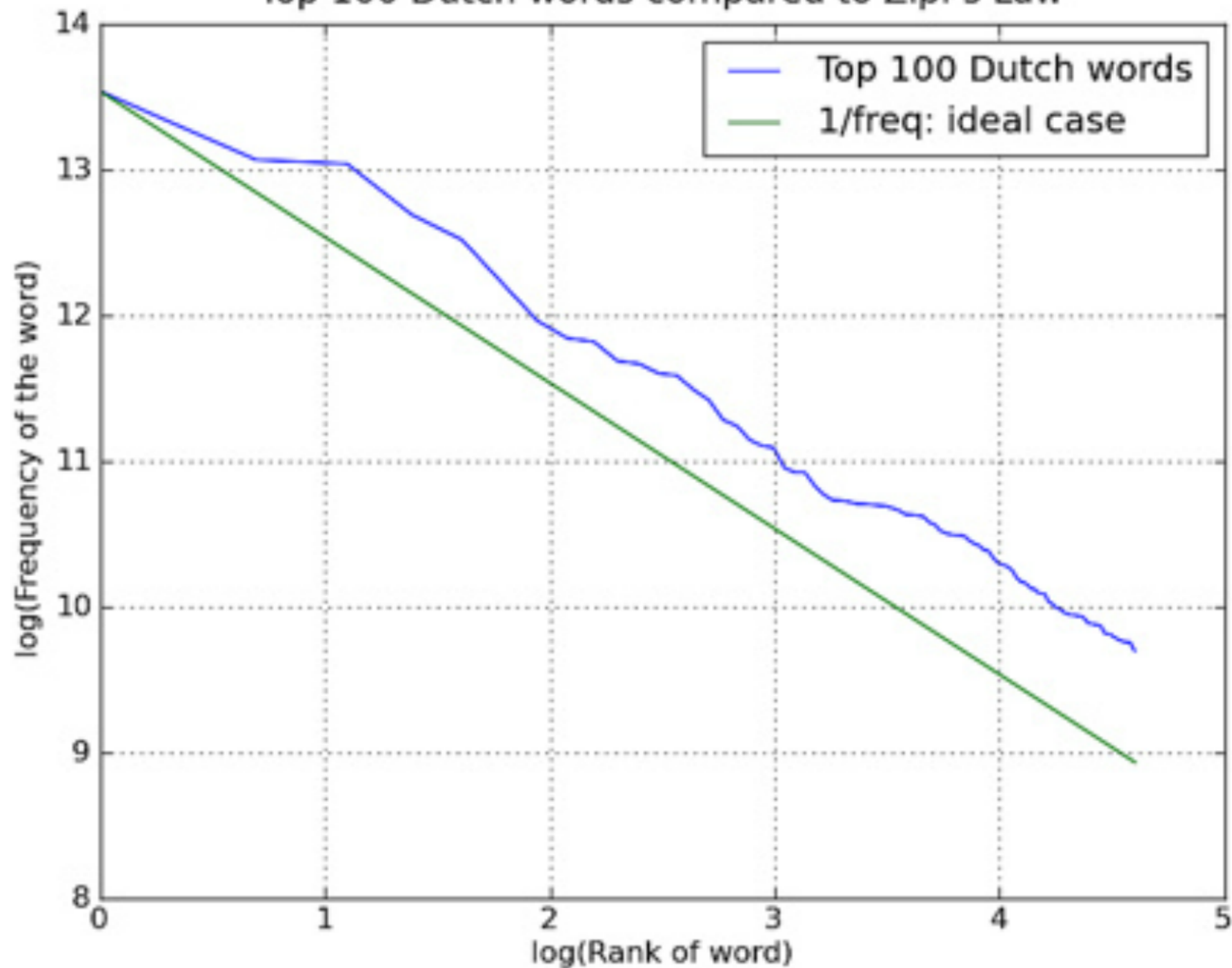
- IB047 Úvod do korpusové lingvistiky
- *a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language as a source of data for linguistic research*
- dokonalý korpus neexistuje
 - korpus je jen vzorek jazyka
 - obsahuje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

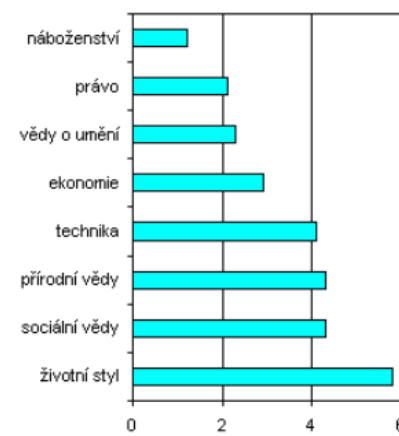


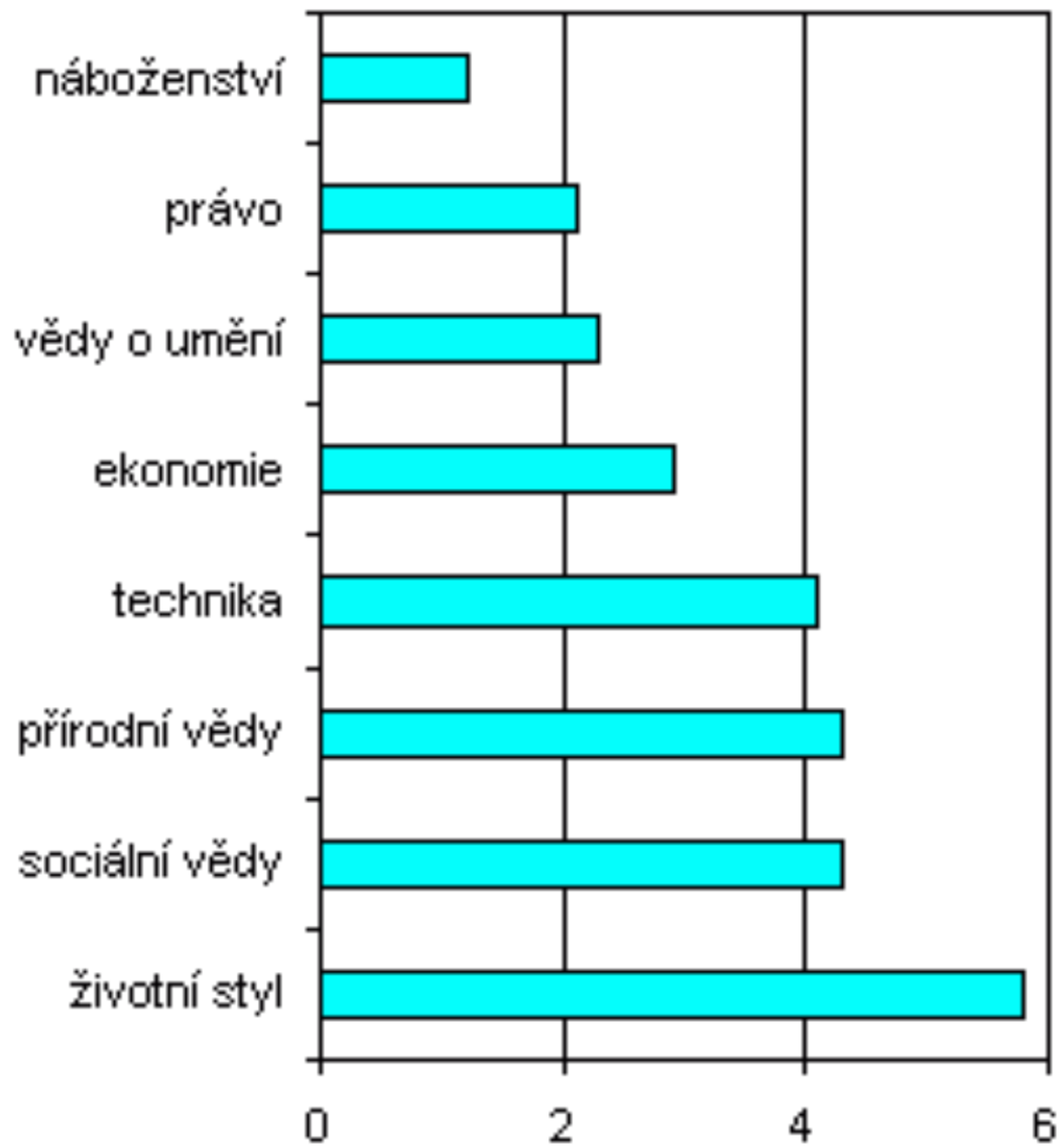
Top 100 Dutch words compared to Zipf's Law



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

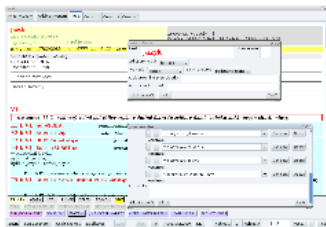


Korpus

- 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

Zdroj pro heslář fsc+ssjc

Typ hesla jednosl. podstatné jméno

STAT. Sg/Pl 12792/3985 FRQ 16777 ARF 6122 Zdroj

SEZNAMY (rozbalit/sbalit)

SLOVOTVORBA

Derivovaná slova

Fundace/motivace

Defraz. lexémy

Zpracovatel veselý

Vytvořeno 2008-12-04 10:44 Změněno 2009-09-14 12:26

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 (rozbalit/sbalit)

E4 k V1 Subst+SUBST-gen (rozbalit/sbalit)

vyplazování jazyka;
vyříznutí jazyka;
špička jazyka, kořen jazyka

Pozn. k E4 zpracovávané substantivum je samo genitivním

E5 k V1 SUBST+Prep+Subst/SUBST+Subst-ji (rozbalit/sbalit)

Pozn. k E5 předložkový přízvuček ke zpracovávanému sub

jazyk : FRAZÉMY

- + mít jazyk (ostrý) jako meč Dohledat Odkaz

Poznámka

- + mít jazyk jako na obrtlíku Dohledat Odkaz

Poznámka

- + mít jazyk jako poleno Dohledat Odkaz

Poznámka

- + mlčí jako by /mu/ přimrzl jazyk Dohledat Odkaz

Poznámka

Uložit a zavřít Uložit Zavřít

ZÁHLAVÍ ADMIN. VÝSL. PŮVOD DĚLENÍ ETYMOLOG. STAT.

EXPR. ÚZEMNÍ PŘÍZNAK DOBOVÝ PŘÍZNAK STYLOVÝ PŘÍZNAK

ZKR./EKVIVALENTY SOUSLOVÍ FRAZÉMY JINÉ VÍCESL. VÝRAZY VÍCESL. NEZAŘAZENO KOMP. VÍCESL. SLOVOTVORBA

game

In 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 with **games** 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** the only game in town **LINK** game show **LINK** games room **LINK** game theory

Makrostruktura

Makrostruktura

- heslář (+předmluva, přílohy...)
- heslo¹ = lemma, entry term, heslové slovo, headword
 - obvykle nominativ sg., slovesa v infinitivu
 - části slov, spojení slov

Makrostruktura

- heslář (+předmluva, přílohy...)
- heslo¹ = lemma, entry term, heslové slovo, headword
 - obvykle nominativ sg., slovesa v infinitivu
 - části slov, spojení slov
- heslo² = heslová stať, entry

Heslář

- rozsah
- výběr podle oboru a typu
- obecný jazyk: frekvence
- hledání v elektronické verzi

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