

# PA153 Počítačové zpracování přirozeného jazyka

## 04 – Sémantika I (reprezentace lexikálního významu)

Karel Pala, Zuzana Nevěřilová

Centrum ZPJ, FI MU, Brno

1. listopadu 2020

- 1 Lexikální význam
- 2 Slovníkové heslo
- 3 Nalezení významu v kontextu
- 4 Popis lexikálních významů pro ZPJ
  - Sémantické primitivy
  - Sémantické třídy
  - Teorie prototypů
- 5 Shrnutí

# Lexikální význam

lexikální význam (*lexical meaning*): izolovaný význam slova [Oxford Dictionaries, 2013]

- bez ohledu na význam **věty**, ve které se slovo nachází
- bez ohledu na **gramatické kategorie**

jiné významy: gramatický význam, význam slov a význam vět

- *kuře* – *kuřata*
- *frekvence* – *kmitočet*
- Pan profesor *běží* na tramvaj. Gepard *běží* za kořistí.

## └ Lexikální význam

## └ Lexikální význam

lexikální význam (lexical meaning): izolovaný význam slova [Oxford Dictionaries, 2013]

- bez ohledu na význam věty, ve které se slovo nachází
  - bez ohledu na [gramatické kategorie](#)
- jiné významy: gramatický význam, význam slov a význam vět
- kuře – kuřata
  - frekvence – kmitočet
  - Pan profesor běží na tramvaj. Gepard běží za kořistí.

slova kuře a kuřata mají tentýž lexikální význam, ale rozdílný gramatický (singulár, plurál)

frekvence a kmitočet jsou různá slova, která mají tentýž lexikální (i gramatický a dokonce i další) význam

běžet má stejný význam, přestože si představíme celkem jinou činnost (styl, rychlost, terén)

# Lexikální forma a lexikální význam

**Lexikální jednotka** (lexical unit, LU) [Ziková, 2003]:

- reprezentována **lexikální formou**
- asociována s určitým **lexikálním významem**
- má určité **gramatické vlastnosti** (např. tranzitivní sloveso)
- může mít určité **pragmatické vlastnosti** (např. *šmirgl* je expresivní označení pro *tramvaj*)

Lexikální vztahy:

- LU se stejným významem, ale jinou formou **synonymie** (např. šalina, tramvaj, šmirgl)
- LU se stejnou formou, ale jiným významem **homonymie** (např. kolej) nebo **polysémie** (např. ucho)

# Kde najít informace o lexikálním významu?

Slovník/lexikon = soubor lexikálních jednotek (LU)

Slovníky:

- výkladové
- překladové
- současného jazyka (synonym, zkratk, rýmů ...)
- terminologické
- historické
- etymologické
- speciální (frekvenční, retrográdní, valenční)
- ...



## PA153 Zpracování přirozeného jazyka

## └ Lexikální význam

## └ Kde najít informace o lexikálním významu?

## Slovníky:

- výkladové
- příkladové
- současného jazyka (synonym, zkratka, rjmů ...)
- terminologické
- historické
- etymologické
- speciální (frekvenční, retrográdní, valenční)
- ...

Na tomto místě zmiňujeme slovníky jen jako prostředky pro popis lexikálního významu. Více o slovnících, jejich druzích a vytváření slovníků vyučuje předmět *IB047 Úvod do korpusové lingvistiky a počítačové lexikografie*.

# Struktura slovníkového hesla

## bez

-u m. (6. j. -u)

1. *šeřík* (bot.): modrý, bílý b.; kytice bezu

2. *vyšoký keř s květenstvím drobných nažloutlých květů, které dozrávají na podzim v drobné černé bobulky* (bezinky); *bez černý* (bot.): trást bez(em); [x] zůstat pod bezem *neprovdát se*; ob. expr. *jdi mi s tím na b. dej pokoj*; bot. *rod Sambucus*: b. černý; b. hroznatý

3. ob. *květ černého bezu*: vařit čaj z bezu; přen. *odvar z bezového květu*: pít teplý b. (Jir.)

- lexikální forma
- gramatické vlastnosti
- pragmatické vlastnosti
- definice
- kolokace
- příklady užití
- odvozené lexikální formy



# Slovníkové definice a hyperonymie

## bez

-u m. (6. j. -u)

1. *šeřík* (bot.): modrý, bílý b.; kytice bezu
2. *vysoký keř s květenstvím drobných nažloutlých květů, které dozrávají na podzim v drobné černé bobulky (bezinky); bez černý* (bot.): trást bez(em); [x] zůstat pod bezem *neprovdat se*; ob. expr. jdi mi s tím na b. *dej pokoj*; bot. rod *Sambucus*: b. černý; b. hroznatý
3. ob. *květ černého bezu*: vařit čaj z bezu; přen. *odvar z bezového květu*: pít teplý b. (Jir.)

Definice pomocí **synonym**:

bez = šeřík

Definice klasická:

bez = vysoký keř s květenstvím drobných nažloutlých květů... [Havránek et al., 1960]

- genus proximum (nejbližší rod)
- differentia specifica (druhové rozdíly)

hyperonymie

## kulhati

ned. (1. j. -ám, rozk. -ej, přech. přít. -aje)

1. *chodit tak, že váha těla se nepřenáší stejnoměrně na obě nohy, levou nohu*

troponymie



# Nalezení významu v kontextu

někdy (ve skutečnosti velmi často) jen se znalostí lexikálního významu nevystačíme

⇒ je třeba znát kontext

lexikální desambiguace (Word Sense Disambiguation)

funkce:  $(w, c) \rightarrow s$

- $w \in \mathcal{W}$  – množina slov
- $c \in \mathcal{C}$  – množina kontextů
- $s \in \mathcal{S}$  – množina významů

## Naivní Leskův algoritmus: kočka (SSJČ) [Lesk, 1986]

- 1 malá kočkovitá šelma, chovaná v domácnostech, na venkově zvl. pro hubení myší; kočka domácí (zool.); šedivá, černá, tříbarevná k.; hladká srst kočky; k. mňouká, přede; k. číhá na myš; k. chytá ptáky; angorská k.; být falešný, úlisný jako k.; přen. expr. je to k. falešník; to děvče je k. lichotné, úlisné; [x] jsou na sebe jako pes a k. nenávidí se. . .
- 2 malá n. středně velká šelma s hustým kožichem; zool. rod Felis: k. plavá; k. divoká; k. domácí
- 3 samice kočkovité šelmy vůbec; rysí k.; lví k.; expr. každá kočkovitá šelma vůbec (tygr, levhart aj.)
- 4 ob. kožišina na límci, kolem krku n. ramen
- 5 kocovina (Haš.)
- 6 věc připomínající někt. vlastnost u kočky: bot. velký trs ostřic vystupující z rašeliniště (na blatech); tech. pojízdný vozík jeřábu se zdvihacím ústrojím
- 7 druh důtek; devítiočasá k.

## Naivní Leskův algoritmus: vstup

Aminokyselina DL-methionin okyseluje moč, čímž chrání močové ústrojí psů i *koček* (důležitá vlastnost zvláště u kastrováných jedinců).

{aminokyselina, DL-methionin, okyselovat, moč, čímž, chránit, močový, ústrojí, pes, i, důležitý, vlastnost, zvláště, u, kastrováný, jedinec}

{aminokyselina, což, DL-methionin, důležitý, chránit, i, jedinec, kastrováný, moč, močový, okyselovat, pes, u, ústrojí, vlastnost, zvláště}

## Leskův algoritmus: naivní

{aminokyselina, což, DL-methionin, důležitý, chránit, i, jedinec, kastrovaný, moč, močový, okyselovat, **pes**, **u**, **ústrojí**, **vlastnost**, **zvláště**}

**1:** {a, angorský, být, černý, číhat, děvče, domácí, domácnost, expresivně, falešník, falešný, hladký, hubení, chovaný, chytat, jako, kočkovitý, lichotný, malý, mňoukat, myš, na, nenávidět, **pes**, pro, přeneseně, příst, pták, se, srst, šedivý, šelma, to, tříbarevný, úlisný, v, venkov, zoologicky, **zvláště**}

**2:** {divoký, domácí, Felis, hustý, kožich, malý, nebo, plavý, rod, s, středně, šelma, velký, zoologicky}

⋮

**6:** {bláto, botanicky, jeřáb, na, některý, ostřice, pojízdný, připomínající, rašeliniště, s, technicky, trs, **u**, **ústrojí**, věc, velký, vozík, **vlastnost**, vystupující, z, zdvihací}

**7:** {devítiocasá, druh, důtky}

## Leskův algoritmus: naivní

{aminokyselina, což, DL-methionin, důležitý, chránit, i, jedinec, kastrovaný, moč, močový, okyselovat, pes, u, ústrojí, vlastnost, zvláště}

$$D_1 = \{\text{pes, zvláště}\}$$

$$D_2 = \{\}$$

$$D_3 = \{\}$$

$$D_4 = \{\}$$

$$D_5 = \{\}$$

$$D_6 = \{\text{u, ústrojí, vlastnost}\}$$

$$D_7 = \{\}$$

věc připomínající někt. vlastnost u kočky: bot. velký trs ostřic vystupující z rašeliniště (na blatech); tech. pojízdný vozík jeřábu se zdvihacím ústrojím

## PA153 Zpracování přirozeného jazyka

## └─ Nalezení významu v kontextu

## └─ Leskův algoritmus: naivní

## Leskův algoritmus: naivní

{aminokyselina, což, DL-methionin, důležitý, chránit, i, jedínce, kastrovany, moč, močový, okyselovat, pes, u, ústrojí, vlastnost, zvířítě}

$D_1 = \{\text{pes, zvířítě}\}$

$D_2 = \{\}$

$D_3 = \{\}$

$D_4 = \{\}$

$D_5 = \{\}$

$D_6 = \{\text{u, ústrojí, vlastnost}\}$

$D_7 = \{\}$

více připomínající někt. vlastnost u kočky: bot, velký trs ostříc vystupující z rástliniček (na blátech); tech. pojízdný vozík jízdu se zdvihacím ústrojím

Naivní L. algoritmus určil, že význam slova kočka v uvedené větě je 6. Je to spíš náhoda podpořená tím, že u významů 1 a 6 v SSJč také nejvíc textu.

Vylepšené verze L. algoritmu některá slova nepočítají, přidávají slovům váhy (např. pomocí TF-IDF), zohledňují vzdálenost od desambiguovaného slova



## Naivní Leskův algoritmus: kočka (SSJČ) [Lesk, 1986]

- 1 malá kočkovitá šelma, chovaná v domácnostech, na venkově zvl. pro hubení myší; kočka domácí (zool.); šedivá, černá, tříbarevná k.; hladká srst kočky; k. mňouká, přede; k. číhá na myš; k. chytá ptáky; angorská k.; být falešný, úlisný jako k.; přen. expr. je to k. falešník; to děvče je k. lichotné, úlisné; [x] jsou na sebe jako pes a k. nenávidí se. . .
- 2 malá n. středně velká šelma s hustým kožichem; zool. rod Felis: k. plavá; k. divoká; k. domácí
- 3 samice kočkovité šelmy vůbec; rysí k.; lví k.; expr. každá kočkovitá šelma vůbec (tygr, levhart aj.)
- 4 ob. kožišina na límci, kolem krku n. ramen
- 5 kocovina (Haš.)
- 6 věc připomínající někt. vlastnost u kočky: bot. velký trs ostřic vystupující z rašeliniště (na blatech); tech. pojízdný vozík jeřábu se zdvihacím ústrojím
- 7 druh důtek; devítiočasá k.

# Slabiny WSD

$(w, c) \rightarrow s$

- $w \in \mathcal{W}$  – množina slov
- $c \in \mathcal{C}$  – množina kontextů
- $s \in \mathcal{S}$  – množina významů

Všechny algoritmy WSD závisejí na inventáři a popisu významů.

Kolik významů má slovo *kočka*?

- SSJČ: 7
- SSČ: 2
- PSJČ: 10
- Slovník českých synonym: 4
- Český WordNet: 3

$(w, c) \rightarrow s$ 

- $w \in W$  – množina slov
- $c \in C$  – množina kontextů
- $s \in S$  – množina významů

Všechny algoritmy WSD závisí na inventáři a popisu významů.

Kolik významů má slovo ločka?

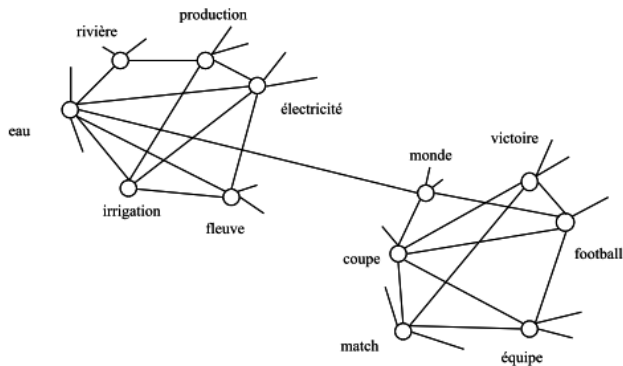
- SSJČ: 7
- SSČ: 2
- PSJČ: 10
- Slovník českých synonym: 4
- Český WordNet: 3

Leskův a. je jednoduchý i ve svých pokročilejších verzích, zajímavý algoritmus nabídl [Yarowsky, 1995]. Jde o alg. strojového učení, kdy se v prvním průchodu určí kolokace, které naprosto jistě souvisejí s konkrétním významem slova. V dalších průchodech se vypočítávají další slova, která signalizují konkrétní význam slova.

# WSD nebo WSD

Algoritmy, které nepočítají s pevným inventářem významů, jen s kontextem:

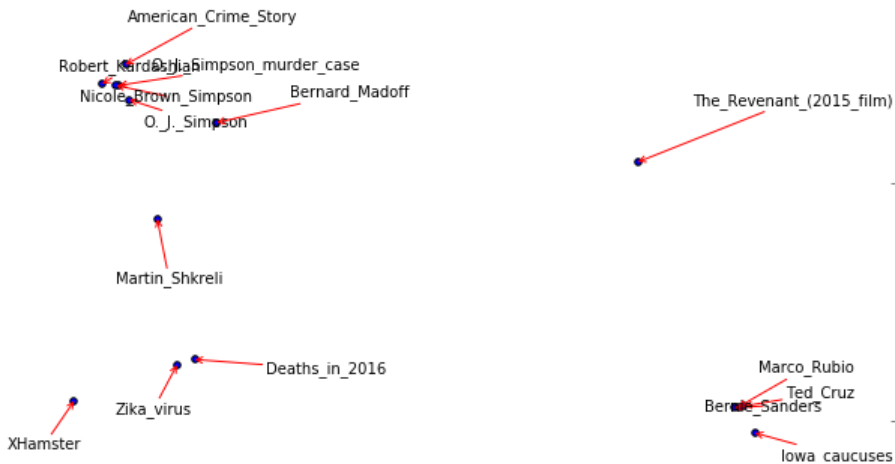
## Word Sense Discrimination



[Véronis, 2004]

# Word Sense Discrimination

Algoritmy, které nepočítají s pevným inventářem významů, jen s kontextem: **word embeddings**



# Komponentová analýza (Componential analysis)

= popis významů slov pomocí množiny sémantických rysů (primitiv), které jsou buď přítomny, nebo nepřítomny, nebo irelevantní pro daný význam:

- muž = +ADULT +MALE
- chlapec = -ADULT +MALE
- batole = -ADULT ±MALE

# Komponentová analýza (Componential analysis) I

| označení | popis           | příklad                        |
|----------|-----------------|--------------------------------|
| T        | tempus, čas     | den, rok, leden, soumrak       |
| L        | locus, místo    | dům, chrám, světadíl, břeh     |
| BYT      | bytost          | víla                           |
| HUM      | člověk          | střežda, rada, bača            |
| ANIM     | zvíře           | pes, slon, velbloud            |
| PLANT    | rostlina        | strom, kosatec                 |
| QUA      | vlastnost       | nespokojenec, povýšenec + HUM  |
| FEN      | fenomén         | úkaz, zázrak                   |
| ENT      | entita          | protiklad, argument            |
| OBJ      | objekt, předmět | stůl, krb, ale i dům (OBJ + L) |

# Komponentová analýza (Componential analysis) II

| označení | popis                   | příklad                            |
|----------|-------------------------|------------------------------------|
| INF      | informace               | telefonát, článek, vzkaz           |
| EMO      | emoce                   | cit, radost, strach, neklid, úsměv |
| INS      | instrument, nástroj     | nůž, šíp hřeben                    |
| MACH     | stroj, aparát, zařízení | počítač                            |
| PROC     | proces                  | zážeh, postup, pokrok              |
| MOT      | pohyb                   | běh, let, pád                      |
| AKT      | aktivita, činnost       | boj, odboj, příchod                |
| MAT      | materiál                | hlína, dřevo                       |
| BP       | část těla (body part)   | prst, krk                          |
| ORG      | organizace, instituce   | vláda                              |



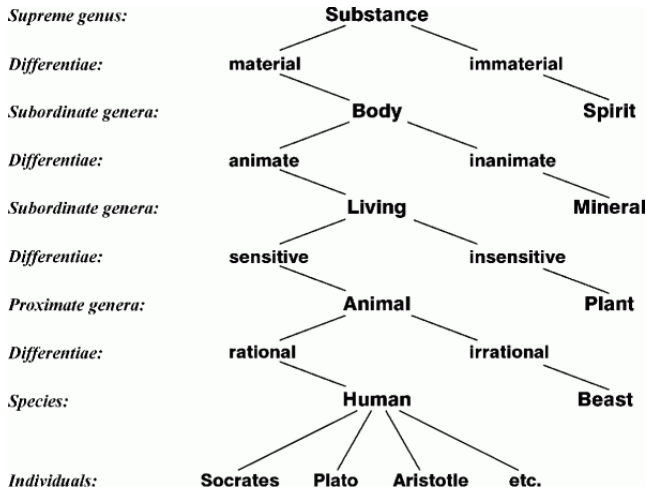
# Sémantické třídy

= skupiny slov, která sdílejí nějakou sémantickou vlastnost

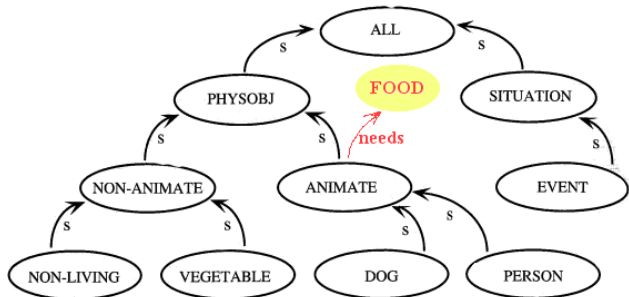
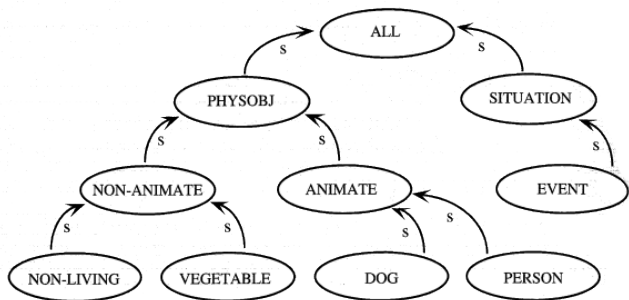
obratlovec – savec – šelma – psovitá šelma – pes – pudl – trpasličí pudl

taxonomie, hierarchie tříd

# Sémantické třídy, Porfyriův strom



# Sémantické třídy, sémantické sítě, odvozování



WordNet (Princeton WordNet, PWN) – lexikální síť

- původně nástroj k ověření teorie o uspořádání lidské paměti (G. A. Miller, od r. 1985)
- počítačově dobře zpracovatelný zdroj informací o významech slov a vztazích mezi významy [Fellbaum, 1998]
- jednotkou je synonymická řada (synonymical set, synset)
- synsety jsou spojeny relacemi:
  - ▶ hyperonymie/hyponymie: vůz, automobil – dodávka
  - ▶ holonymie/meronymie (part of, member of): vůz, automobil – tlumič; orchestr – houslista
  - ▶ troponymie: šeptat – mluvit
  - ▶ near-antonym: den – noc
  - ▶ odvození: velikost – velký
- slovní druhy: substantiva, adjektiva, verba, adverbia

# WordNet

angličtina: PWN (117 tis. synsetů)

projekty EuroWordNet (holandština, italština, španělština, němčina, francouzština, čeština, estonština)

- ILI - InterLingual Index
- Top Ontology (63 kategorií)
- Base Concepts

projekty (BalkaNet), při kterých vznikají wordnety pro další jazyky, koordinátorem databází je Global WordNet Association (GWA)  
současný český W.: 28 tis. synsetů

# WordNet není jediný

Ontologie = explicitní specifikace sdílené konceptualizace

- firemní o.
- všeobecné o. SUMO/MILO (Suggested Upper Merged Ontology, Mid-Level Ontology)
- common sense o. ConceptNet

# Ontologie a datové formáty (ontologické jazyky)

- predikátová logika 1. řádu a rozšíření
- Rodina KIF (Knowledge Interchange Format)
- Rodina RDF (Resource Description Framework), „jazyky sémantického webu“: RDF, RDFS, OWL, DAML

## PA153 Zpracování přirozeného jazyka

└ Popis lexikálních významů pro ZPJ

└ Sémantické třídy

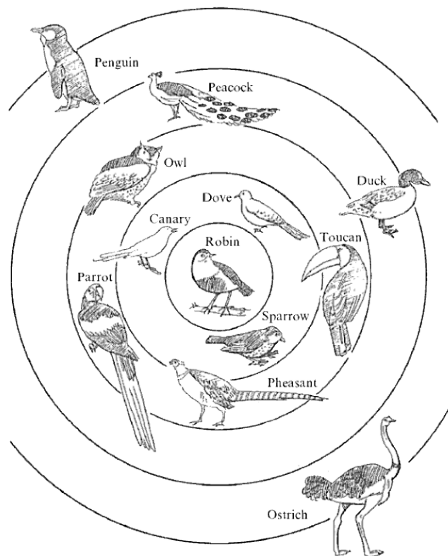
└ Ontologie a datové formáty (ontologické jazyky)

- predikátová logika 1. řádu a rozšíření
- Rodina KIF (Knowledge Interchange Format)
- Rodina RDF (Resource Description Framework), „jazyky sémantického webu“: RDF, RDFS, OWL, DAML

O jazycích SW příště, v souvislosti s popisem významu věty.



# Teorie prototypů: co je to ptáček?



Aitchison, 2003 in [Goddard, 2011]

# Teorie prototypů

E. Rosch dokázala, že lidé uvažují o vlastnostech třídy jako o vlastnostech typického zástupce třídy.

t. prototypů se uplatňuje v popisu typických situací (rámce, skripty)  
vzdálenost mezi koncepty: *židle je víc nábytek než sporák*

# Shrnutí

|                  |                                   |
|------------------|-----------------------------------|
| gramatika        | slovní druh, gramatické kategorie |
| syntax           | větný člen                        |
| <b>sémantika</b> | <b>sémantická třída</b>           |
| pragmatiky       |                                   |

popis lexikálního významu:

- pro uživatele jazyka: slovníky
- pro počítačové programy: specializované zdroje (sém. rysy, ontologie, prototypy)

rozlišení lexikálního významu:

- pro uživatele jazyka: číslo významu
- pro počítačové programy: WSD, vzdálenost mezi koncepty

# Odkazy I



Fellbaum, C. (1998).

*WordNet: An Electronic Lexical Database (Language, Speech, and Communication).*

The MIT Press.

Published: Hardcover.



Goddard, C. (2011).

*Semantic Analysis: A Practical Introduction.*

Oxford Textbooks in Linguistics. Oxford University Press.



Havránek, B. et al. (1960).

*Slovník spisovného jazyka českého (Dictionary of Written Czech, SSJČ).*

Academia, Praha, 1st edition.

electronic version, created in the Institute of Czech Language, Czech Academy of Sciences Prague in cooperation with Faculty of Informatics, Masaryk University Brno.

## Odkazy II



Lesk, M. (1986).

Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone.

In *Proceedings of the 5th annual international conference on Systems documentation*, SIGDOC '86, pages 24–26, New York, NY, USA. ACM.



Oxford Dictionaries (2013).

Lexical meaning. Oxford Dictionaries.  
online.

<http://oxforddictionaries.com/definition/english/lexical-meaning> (accessed October 03, 2013).



Véronis, J. (2004).

Hyperlex: Lexical cartography for information retrieval.

In *Computer Speech and Language: Special Issue on Word Sense Disambiguation*, page 23.

# Odkazy III



Yarowsky, D. (1995).

Unsupervised word sense disambiguation rivaling supervised methods.  
In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, ACL '95, pages 189–196, Stroudsburg, PA, USA. Association for Computational Linguistics.



Ziková, M. (2003).

Současný český jazyk: Tvoření slov.  
online.

[http://www.phil.muni.cz/cest/lide/zikova/CJA009\\_1.rtf](http://www.phil.muni.cz/cest/lide/zikova/CJA009_1.rtf)  
(accessed October 03, 2013).