PA153 Počítačové zpracování přirozeného jazyka

08 - Lexikografické nástroje a počítačová lexikografie

Karel Pala, Adam Rambousek

Centrum ZPJ, FI MU, Brno

11. listopadu 2013

1. Lexikografie
   - Úvod
   - Lexikografie
   - Slovníky a počítače
2. Počítačová lexikografie
   - Reprezentace dat
   - TEI
   - Dictionary Writing Systems
3. Tvorba slovníku
   - Lexikální databáze
   - Slovník

Lexikografie

- PLIN035 Počítačová lexikografie
- podoblast **lexikologie**
- lexicography, **lexikografie**
  - *the activity or occupation of compiling dictionaries* (Oxford d.)
  - *the editing or making of a dictionary* (Merriam-Webster d.)
  - *the job of writing a dictionary* (Macmillan d.)
- praktická lexikografie
- teoretická lexikografie - analýza a popis slovní zásoby, teorie o prvcích slovníku, skupinách uživatelů, hodnocení
- *Slovník národního jazyka náleží mezi první potřebnosti vzdělaného člověka.*

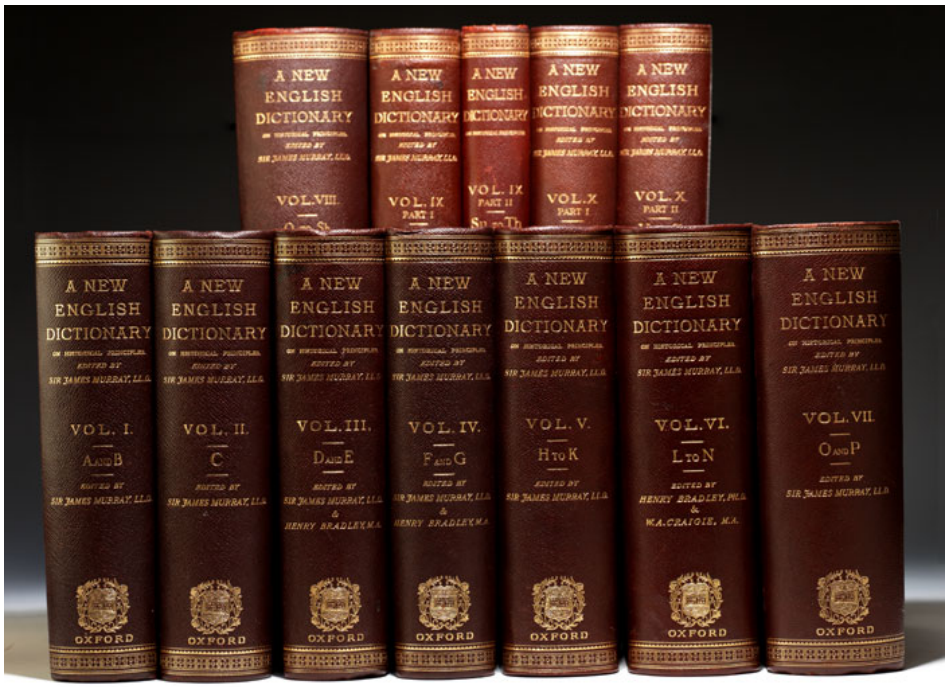


Historie

- hliněné tabulky z Ebla (Sýrie), cca 2500-2250 př.n.l.
  - sumerština - eblaština
- Robert Cawdrey: *A Table Alphabeticall*, 1604
  - první výkladový slovník angličtiny
  - *"hard wordes, borrowed from... for the benefit & helpe of Ladies, Gentlewomen, or any other unskilfull persons"*
- Samuel Johnson: *A Dictionary of the English Language*, 1747-1755
  - moderní slovník, 42 773 hesel
  - *"to preserve the purity and ascertain the meaning our English idiom"*
- Noah Webster: *An American Dictionary of the English Language*, 1828
  - 70 000 hesel, srovnání britské a americké angličtiny
  - odmítal zařazovat do slovníků neslušná slova

Historie

- *The Oxford English Dictionary (A New English Dictionary)*
  - 1857, Philological Society, R. C. Trench, kritika slovníků
  - 1879, James A. H. Murray jmenován hlavním editorem
  - 1882-1928, vychází 12 svazků, 15 487 stran, 240 000 hesel

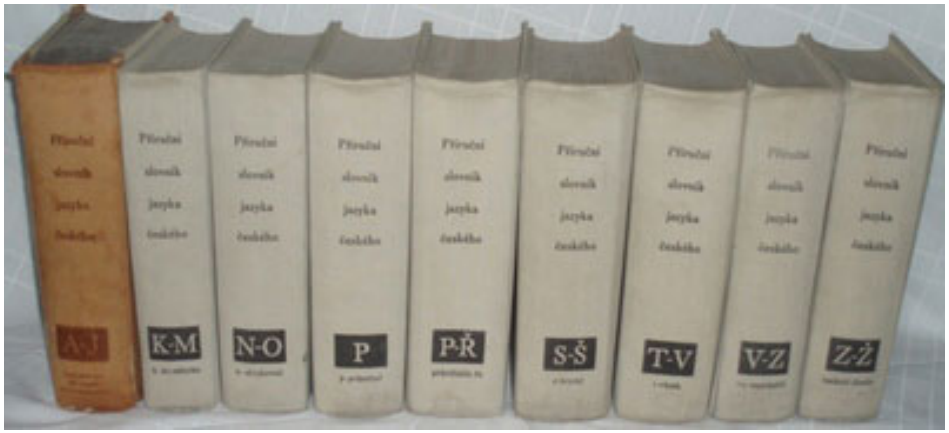


Historie

- Bartoloměj z Chlumce, Klaret, 14. století
  - latinsko-české slovníky, *Vokabulář (gramatický), Bohemář, Glosář*
  - Raro sequens gesta de bestiis cernis honesta.
    Lew leo *wlkque* lupusque le[e]na *lwicze*, *nedvied* ursus
    Ursaque *nedviedicze*, lupa *wlczicze*, dic ovis *owcze*,
    Koza capra, vulpes *lyska*que canicula tysta.
- Daniel Adam z Veleslavína, 16. století
  - *Nomenclator quadrilinguis* + *Silva quadrilinguis*,
    čeština-latina-řečtina-němčina, 958+300 stran, řazeno česky
- Jan Amos Komenský, 17. století
  - *Thesaurus linguae Bohemicae* - latinsko-český, česko-latinský,
    synchronní, diachronní, lexikální, gramatické informace, frazeologie
  - 20 let příprav... požár Lešna

Historie

- Josef Jungmann, *Slovník česko-německý*
  - 1815-1833, vydáno 1835-1839
  - 5 svazků, 4694 stran
  - popisný výkladový slovník
- *Kancelář Slovníku jazyka českého*, 1911
  - sběr slovníkového materiálu, dobrovolníci
  - výpisky z prózy, básní, odborné literatury, publicistických článků
  - *Příruční slovník jazyka českého*, 1935-1957
  - 10 824 stran, 250 000
  - hesel cenzura "nežádoucích spisovatelů"



Slovníky a počítače

- 60. léta - používají se počítače, lexikografové píší na papír, specialisté přepisují do databáze, Brown Corpus
- 1978, *Longman Dictionary of Contemporary English*
  - první s omezeným slovníkem definicí, kontrolováno strojově
  - kódování pro NLP výzkum
- 1980, *COBUILD*, University of Birmingham + Collins
  - korpus současných textů (Bank of English)
  - 1987, *Collins COBUILD English Language Dictionary*
  - první slovník založený na korpusových datech
  - nový styl definice - celé věty
  - *If a person, animal, or other living thing* **is killed**, *something or someone causes them to die.*
- 90. léta - vývoj specializovaných systémů pro tvorbu slovníků
- 1987, Text Encoding Initiative

XML

- PB138 Moderní značkovací jazyky
- eXtensible Markup Language - značkovací (meta)jazyk
- pravidla, jak má vypadat správně vytvořený dokument - snadné strojové zpracování a výměna informací
- konkrétní názvy značek určuje uživatel (standardy, vlastní)
- elementy `<značka>obsah</značka>`
- bez obsahu lze `<značka></značka>` zkrátit na `<značka/>`
- atributy `<značka atribut="hodnota"/>`

XML

- správné zanoření značek
  - správně: <a>text
  - špatně: <a>text
- speciální znaky (např. <, >,&) se přepisují na entity (např. <)

Popis struktury a kontrola obsahu

- **DTD** (Document Type Definition)
  - seznam elementů a atributů a vztahy mezi nimi
  - nekontroluje obsah
  - `<!ELEMENT vyznam (definice, priklad+)>`
  - `<!ATTLIST vyznam cislo CDATA #REQUIRED>`
- **XML Schema** (XSD, XML Schema Definition)
  - popis obsahu a struktury XML dokumentu, schéma samotné je XML dokument
  - elementy, atributy, struktura
  - možnost určit vlastní typy obsahu (např. opakující se adresa)
  - kontrola obsahu (např. číselný rozsah, regulární výrazy, povolené hodnoty)

Zobrazení

- **XSLT** – eXtensible Stylesheet Language (Transformations)
- převod XML na jiné formáty
  - jiné XML značkování, text, HTML, LaTeX, PDF
- šablony pro části XML dokumentu, postupné procházení dokumentu
- funkcionální programovací jazyk

ssjc Slovník spisovného jazyka českého

lov

-u m. (6. j. -u)
1. *stíhání a zmocňování se zvěře (nejč. odstřelem); chytání ryb:* l. jelenů, divokých kachen, velryb; l. lososů; l. perel; doba lovu; uspořádat l. na medvědy; vyjet na l.; právo lovu; l. odstřelem, chytáním, lapáním; l. lesní, polní, vodní; hromadný l. *hon;* liška vyšla na l.; lovu zdar! *(lovecký pozdrav)*
2. expr. *chytání, shánění čehokoliv, vůbec získávání, při kterém se uplatní obratnost a náhoda:* l. vzácného hmyzu; sběratelé se vydali na l. lidových písní; policie podnikla l. na zloděje; expr. to je l.! *šťastný nález, výhodná koupě ap.*
3. *výsledek lovu; úlovek, kořist:* vrátit se s bohatým lovem *s ulovenou zvěří ap.*, přen. expr. *s věcmi získanými obratností n. šťastnou náhodou*

SSC Slovník spisovné češtiny

lov

-u m
1. *lovení zvěře a ryb* lov koroptví, lov na zajíce, liška vyšla na lov,
2. *úlovek (syno) kořist (syno)* mít bohatý lov,

Ukládání

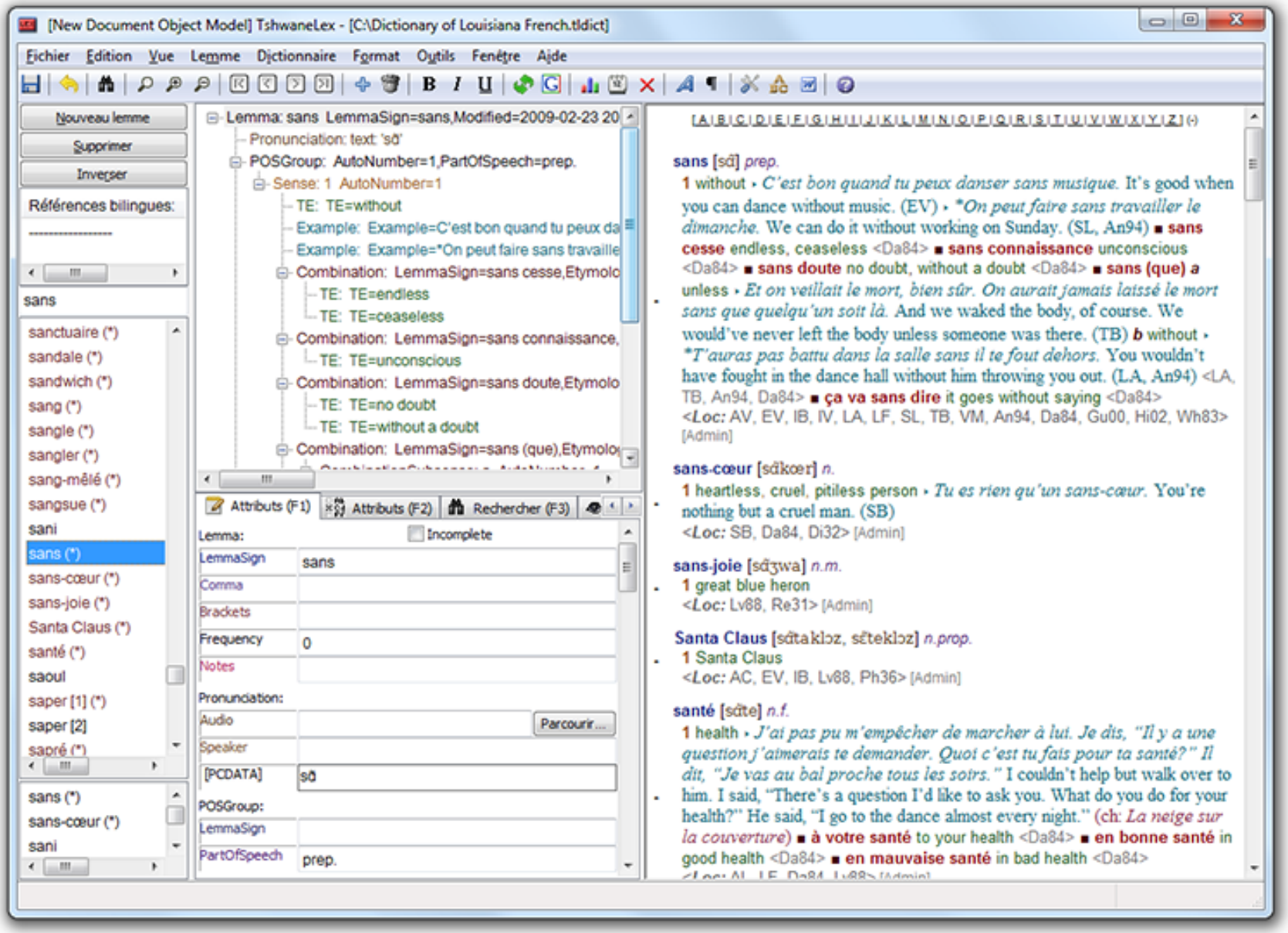
- XML databáze
- ukládají se přímo XML dokumenty
- vyhledávání - XPath, XQuery
- např. eXist, BaseX, Sedna

TEI

- *Text Encoding Initiative*, `http://www.tei-c.org/`
- *TEI Guidelines* (aktuálně verze 5 z roku 2007)
- XML formát pro sémantický popis textových dokumentů
- velký rozsah značek
- *TEI Lite* – osekaná verze, *"90 % potřeb 90 % uživatelů"*
- romány, poezie, divadelní hry, dokumentace, slovníky, korpusy, grafy, rukopisy, zarovnání, odkazy, změny textu, notové zápisy...
- nástroje - sada XSLT pro převod na LaTeX, docx, EPUB, HTML

Dictionary Writing Systems

- aplikace pro tvorbu slovníků (obvykle celý proces tvorby)
- často vlastní
- komerční
  - *IDM DPS* - klient-server (Windows)
  - *iLex* - jádro a dokupované moduly, samostatně nebo klient-server, mobily (Windows, Linux, Mac)
  - *TLex* - online, offline (Windows, Mac)
- *DEB (Dictionary Editor and Browser)*
  - platforma pro slovníkové aplikace
  - klient-server, základní knihovny, speciální moduly
  - DEBDict, DEBVisDic, Internetová jazyková příručka
  - `http://deb.fi.muni.cz`


[New Document Object Model] TshwaneLex - [C:\Dictionary of Louisiana French.tldict]
Fichier Edition Vue Lemme Dictionnaire Format Outils Fenêtre Aide
Nouveau lemme
Supprimer
Inverser
Références bilingues:
sans
sanctuaire (*)
sandale (*)
sandwich (*)
sang (*)
sangle (*)
sangler (*)
sang-mêlé (*)
sangsue (*)
sani
sans (*)
sans-cœur (*)
sans-joie (*)
Santa Claus (*)
santé (*)
saoul
saper [1] (*)
saper [2]
Lemma: sans LemmaSign=sans,Modified=2009-02-23 20
Pronunciation: text: 'sɑ̃'
POSGroup: AutoNumber=1,PartOfSpeech=prep.
Sense: 1 AutoNumber=1
TE: TE=without
Example: Example=C'est bon quand tu peux da
Example: Example="On peut faire sans travaille
Combination: LemmaSign=sans cesse,Etymolo
TE: TE=endless
TE: TE=ceaseless
Combination: LemmaSign=sans connaissance,
TE: TE=unconscious
Combination: LemmaSign=sans doute,Etymolo
TE: TE=no doubt
TE: TE=without a doubt
Combination: LemmaSign=sans (que),Etymolo
Attributs (F1)
Attributs (F2)
Rechercher (F3)
Lemma:
Incomplete
LemmaSign
sans
Comma
Brackets
Frequency
0
Notes
Pronunciation:
Audio
Parcourir...
Speaker
[PCDATA]
sɑ̃
POSGroup:
LemmaSign
PartOfSpeech
prep.
[A|B|C|D|E|F|G|H|I|J|K|L|M|N|O|P|Q|R|S|T|U|V|W|X|Y|Z]
sans [sɑ̃] prep.
1 without ▸ C'est bon quand tu peux danser sans musique. It's good when you can dance without music. (EV) ▸ *On peut faire sans travailler le dimanche. We can do it without working on Sunday. (SL, An94) ■ sans cesse endless, ceaseless <Da84> ■ sans connaissance unconscious <Da84> ■ sans doute no doubt, without a doubt <Da84> ■ sans (que) a unless ▸ Et on veillait le mort, bien sûr. On aurait jamais laissé le mort sans que quelqu'un soit là. And we waked the body, of course. We would've never left the body unless someone was there. (TB) b without ▸ *T'auras pas battu dans la salle sans il te fout dehors. You wouldn't have fought in the dance hall without him throwing you out. (LA, An94) <LA, TB, An94, Da84> ■ ça va sans dire it goes without saying <Da84>
<Loc: AV, EV, IB, IV, LA, LF, SL, TB, VM, An94, Da84, Gu00, Hi02, Wh83> [Admin]
sans-cœur [sɑ̃kœr] n.
1 heartless, cruel, pitiless person ▸ Tu es rien qu'un sans-cœur. You're nothing but a cruel man. (SB)
<Loc: SB, Da84, Di32> [Admin]
sans-joie [sɑ̃ʒwa] n.m.
1 great blue heron
<Loc: Lv88, Re31> [Admin]
Santa Claus [sɑ̃takloz, sɑ̃teklɔz] n.prop.
1 Santa Claus
<Loc: AC, EV, IB, Lv88, Ph36> [Admin]
santé [sɑ̃te] n.f.
1 health ▸ J'ai pas pu m'empêcher de marcher à lui. Je dis, "Il y a une question j'aimerais te demander. Quoi c'est tu fais pour ta santé?" Il dit, "Je vas au bal proche tous les soirs." I couldn't help but walk over to him. I said, "There's a question I'd like to ask you. What do you do for your health?" He said, "I go to the dance almost every night." (ch: La neige sur la couverture) ■ à votre santé to your health <Da84> ■ en bonne santé in good health <Da84> ■ en mauvaise santé in bad health <Da84>

Lexikální databáze

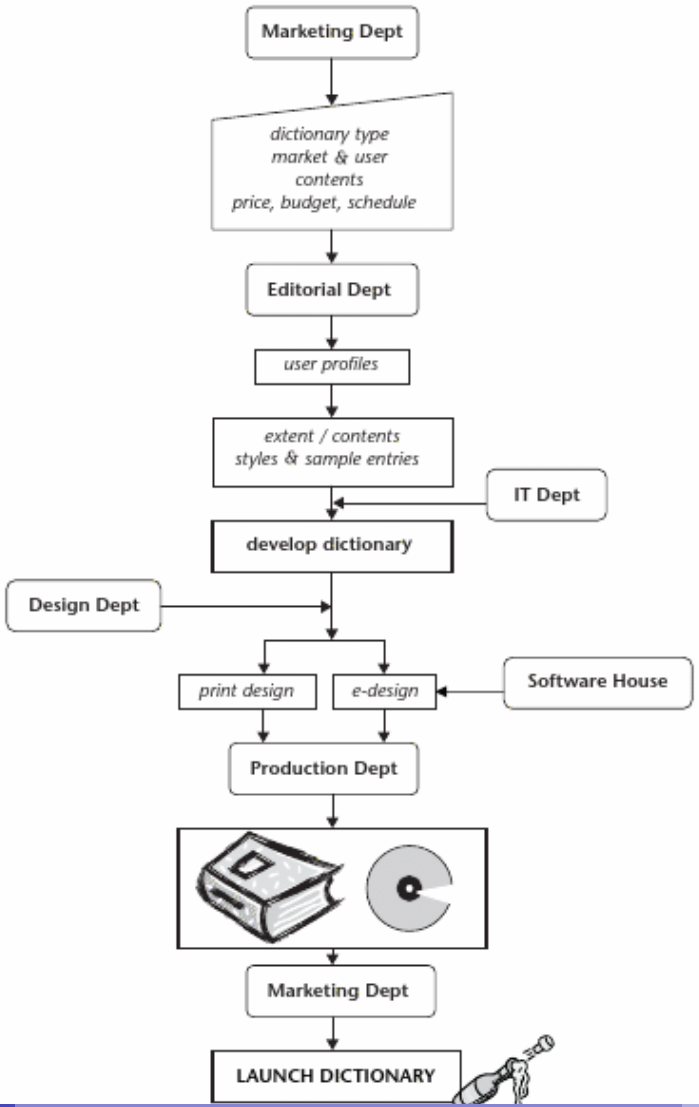
- podrobná strukturovaná jazyková databáze
  - (nyní obvykle) doklady z korpusu
  - gramatické údaje
  - valence, vzory
  - styl, užití, oblast...
  - vztahy mezi slovy
- podklad pro slovníky a výzkum
- *PraLeD* (Pražská Lexikální Databáze)
- *DANTE* (Database of ANalysed Texts of English)

Tvorba slovníku

- tvorba slovníků je drahá, náročná a trvá dlouho, konkurence
- grant nebo se musí vyplatit

Tvorba slovníku

- B. T. Sue Atkins, Michael Rundell: *The Oxford Guide to Practical Lexicography*



Tvorba slovníku

- co chybí? → druh slovníku a jeho uživatelé
- rozpočet a časový plán
- uživatelské profily, Style guide
- editační software (výroba nebo nastavení)
- korpus (vývoj, prohledávání)
- procesy
- *píšeme slovník*
- vzhled a sazba (tisk, digitální)
- výroba
- propagace
- prodej
- profit

Obsah slovníku

- **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
- **mikrostruktura** – struktura jednoho záznamu ve slovníku
  - kontrola pomocí softwaru
  - usnadnění orientace pro čtenáře

Elektronické slovníky

- více informací (CD, DVD, web)
- multimédia
- delší vysvětlující články, odkazy na další zdroje
  - materiály pro učitele, pro studenty
  - přibalený korpus
- vyhledávání
- navigace
- zobrazování údajů podle profilu uživatele (časté operace)