PA153 Natural Language Processing

08 - Lexicographic tools and computational lexicography

Karel Pala, Adam Rambousek

Centrum ZPJ, FI MU, Brno

16. listopadu 2015

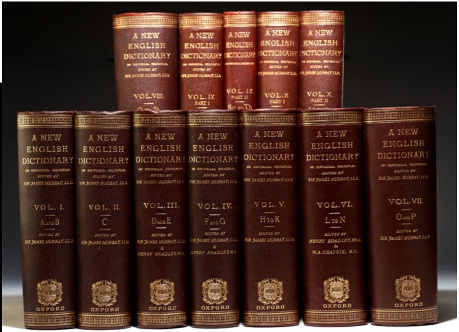
Lexicography

- PLIN035 Computational Lexicography
- subfield of **lexicology**
- 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.)
- practical lexicography
- theoretical lexicography – analysis and description of the lexicon, theory of dictionary components, user groups, evaluation
- *Slovník národního jazyka náleží mezi první potřebnosti vzdělaného člověka.*



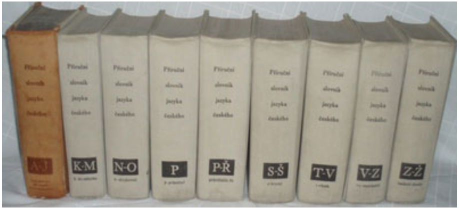
History

- Ebla (Syria) clay tablets, cca 2500-2250 BC
  - Sumerian – Ebla language
- *The Oxford English Dictionary (A New English Dictionary)*
  - 1857, Philological Society, R. C. Trench, criticizing dictionary
  - 1879, James A. H. Murray appointed chief editor
  - 1882-1928, published in 12 volumes, 15 487 pages, 240 000 entries



History

- *Kancelář Slovníku jazyka českého*, 1911
  - volunteers gathering supporting materials
  - excerpts from novels, poems, technical books, journals
  - *Příruční slovník jazyka českého*, 1935-1957
  - 10 824 pages, 250 000 entries
  - quotes by "unwanted authors" censored (Karel Čapek = Lid.nov.)



Future?

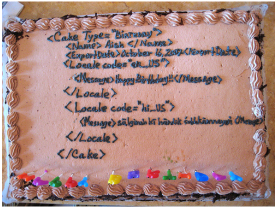
- *Akademický slovník současné češtiny*
  - 2005–2010, lexical database (Praled)
  - 2012–2016, applied research
  - planned 120-150 thousands
  - finished A (2700) to be published in December, B,C in 2017
  - mainly electronic (web, mobile)

Dictionaries and computers

- 1960s – computers are used, lexicographers writing on paper, operators typing into database, Brown Corpus
- 1978, *Longman Dictionary of Contemporary English*
  - 1st with limited definition dictionary, checked automatically
  - special coding for NLP research
- 1980, *COBUILD*, University of Birmingham + Collins
  - contemporary corpus (Bank of English)
  - 1987, *Collins COBUILD English Language Dictionary*
  - 1st dictionary based on corpus data
  - new definition style – full sentence
  - *If a person, animal, or other living thing* **is killed**, *something or someone causes them to die.*
- 1990s – development of specialised dictionary writing systems
- 1987, Text Encoding Initiative

XML

- PB138 Modern Markup Languages
- eXtensible Markup Language – markup (meta)language
- rules for properly formatted document – easy machine processing and information exchange
- actual markup specified by the user (standards, custom)
- elements `<tag>content</tag>`
- without content `<tag></tag>` may be shortened to `<tag/>`
- attributes `<tag attribute="value"/>`

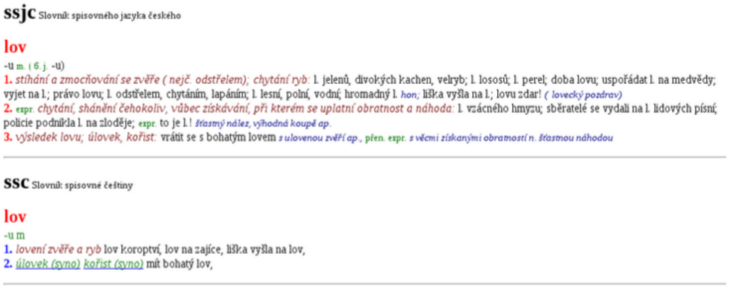


Structure and content description

- **DTD** (Document Type Definition)
  - list of elements and attributes, and their relations
  - no content checking
  - <!ELEMENT meaning (definition, usage+)>
  - <!ATTLIST meaning number CDATA #REQUIRED>
- **XML Schema** (XSD, XML Schema Definition)
  - description of XML document structure and content, schema itself is XML document
  - elements, attributes, structure
  - possibility to define custom content types (e.g. postal address)
  - content checking (e.g. number range, regular expressions, allowed values)

Display

- **XSLT** – eXtensible Stylesheet Language (Transformations)
- converting XML to another format
  - other XML markup, plain text, HTML, LaTeX, PDF
- small templates for parts of XML document, recursive processing of the document
- (functional programming language)


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,

Storing

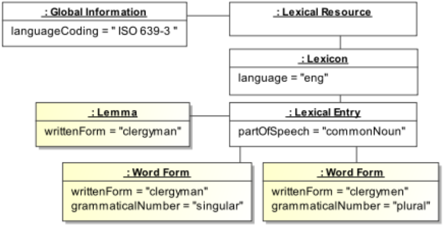
- XML database
- storing XML documents directly
- searching – XPath, XQuery
- e.g. eXist, BaseX, Sedna

TEI

- *Text Encoding Initiative*, `http://www.tei-c.org/`
- *TEI Guidelines* (current version 5, published 2007)
- XML format for semantic description of text documents
- wide range of markup tags
- *TEI Lite* – smaller version, *"90 % needs of 90 % of users "*
- novels, poems, theatre plays, technical reference, dictionaries, corpora, alignment, text revisions, musical notation...
- tools – XSL transformations to LaTeX, docx, EPUB, HTML

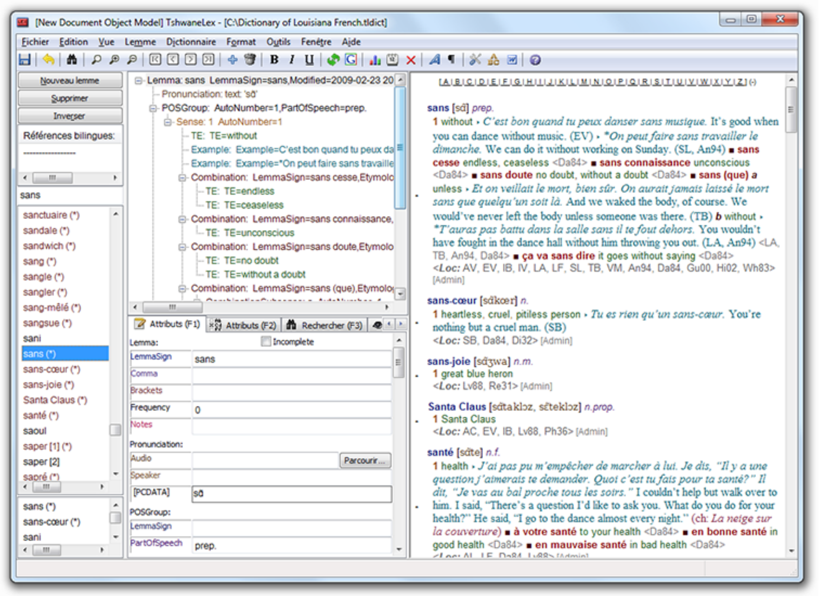
LMF

- *Lexical Markup Framework*, `http://www.lexicalmarkupframework.org/`
- ISO-24613:2008
- common model for lexical resources
- emphasis on machine processing and extensibility
- UML diagram for the lexicon
- core with basic information + extensions for various areas (morphology, syntax, semantics...)



Dictionary Writing Systems

- software application for dictionary creation (usually full process)
- connected to other resources (corpora, analyzers...)
- often custom developed
- commercial (IDM DPS, iLex, TLex, ABBYY Lingvo Content)
- *DEB (Dictionary Editor and Browser)*
  - platform to build dictionary applications
  - client-server, core libraries, specialized modules
  - DEBDict, DEBVisDic, Internetová jazyková příručka, DEBWrite
  - `http://deb.fi.muni.cz`

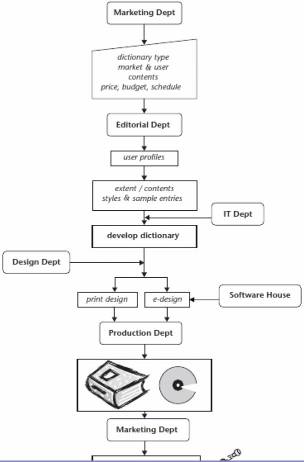

[New Document Object Model] TshwaneLex - [C:\Dictionary of Louisiana French.tldict]
Fichier Edition Vue Lemme Dictionnaire Format Outils Fenêtre Aide
Nouveau lemme
Supprimer
Inverser
Références bilingues:
sans
sanctuaire (*)
sandale (*)
sandwich (*)
sang (*)
sangle (*)
sangler (*)
sang-mêlé (*)
sangsue (*)
sani
sans (*)
sans-cœur (*)
sans-joie (*)
Santa Claus (*)
santé (*)
saoul
saper [1] (*)
saper [2]
sapré (*)
sans (*)
sans-cœur (*)
sani
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.
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 [sɑ̃taklɔz, sɛ̃teklɔz] 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>

Lexical database

- detailed structured database of language
  - (recently) usage examples from corpus
  - grammar
  - valences, patterns
  - language style, usage, region...
  - word relations
- foundation for dictionaries and research
- *PraLeD* (Pražská Lexikální Databáze)
- *DANTE* (Database of ANalysed Texts of English)

Dictionary creation

- dictionary writing is expensive, laborious and time-consuming, competition
- B. T. Sue Atkins, Michael Rundell: *The Oxford Guide to Practical Lexicography*



Dictionary content

- **macrostructure** – entry list (+preface, appendices...)
- heslo^1 = lemma, entry term, heslové slovo, headword
  - noun singular, verb infinitive
  - word parts, collocations
- heslo^2 = heslová stať, entry
- **microstructure** – structure of one entry in the dictionary
  - checked by editing software
  - easier orientation for the reader

Electronic dictionaries

- more information (CD, DVD, web)
  - presentation space
- multimedia, searching, navigation, updates
- longer descriptions, links to further resources
- display information based on user profile
- connection with corpora – ordnet.dk, DWDS.de...
- combining resources, downloading data – Wordnik.com
- user-created content (90-9-1) – Wiktionary, slovnik.zcu.cz...
- Macmillan – switch to digital only
- OED3 – 2000 to 2037, periodical updates
- shift from products to services