Počítačová lexikografie Makrostruktura

Adam Rambousek

Lexikografické podklady

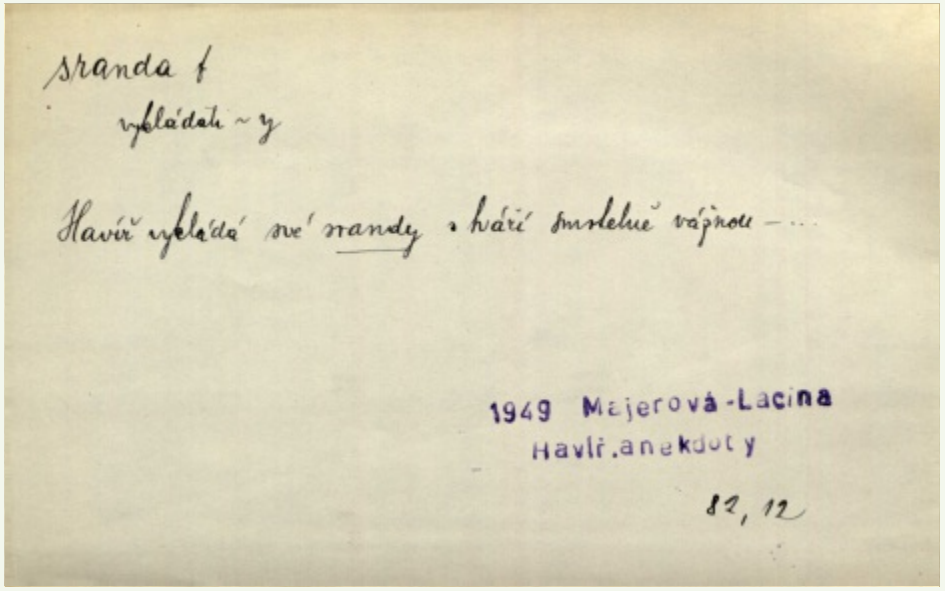
- důkazy o použití jazyka
  - intuice
  - excerpta, výpisky
  - korpusy
- intuice (armchair linguistics)

Intuice

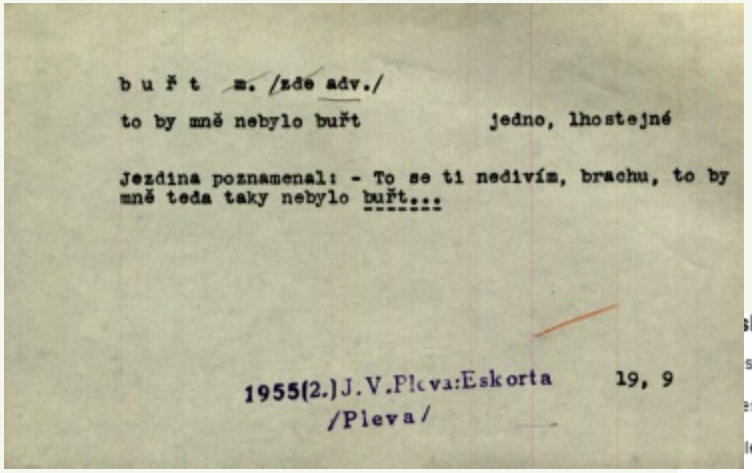
- In the absence of objective evidence, introspection was appealed to instead. But studies in corpus linguistics have shown that introspection is a very flawed technique. We human beings are wired to register the unusual in our minds, generally in a way that is available to conscious recall. But we fail 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



Výpisky



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/

sla

Bleskový filtr

slo	myš	Slovní druh	podst ž
eslí	myš (polohovací to zařízením)		
lek	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

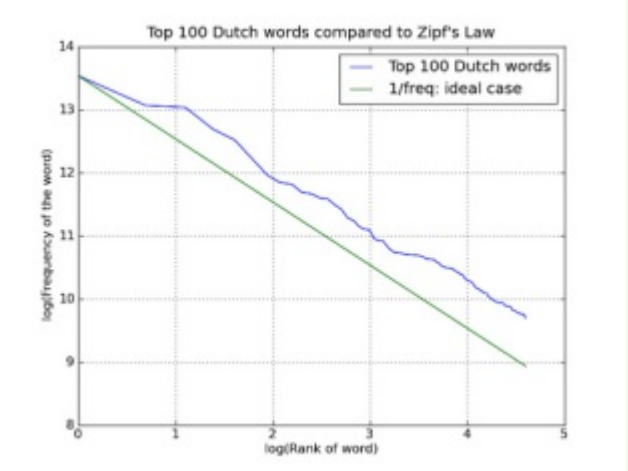
- výhody
  - posuny významu
  - terminologie
  - šíření lexikografie
- nevýhody
  - pracné, časově náročné
  - subjektivní (časté výjimky)

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

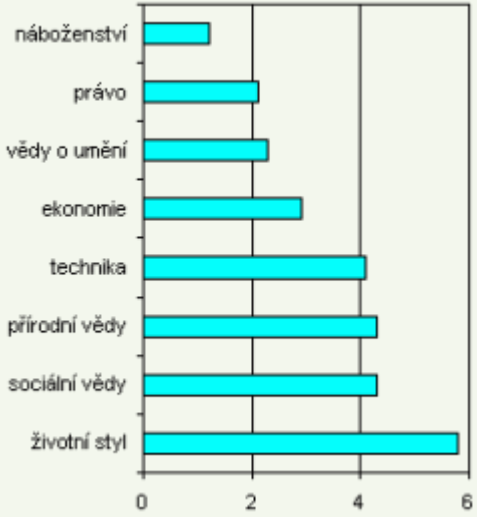
- velikost
  - Brown Corpus (1960) milion slov (10^6)
  - COBUILD (1980) 20 milionů slov (10^7)
  - BNC (1990) 100 milionů slov (10^8)
  - OEC (2000) miliarda slov (10^9)
  - TenTen 10^{10} slov



- Zipfův zákon (1935) několik slov s vysokou frekvencí, mnoho slov s nízkou frekvencí
  - 10. slovo je 10x častější než 100. slovo

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



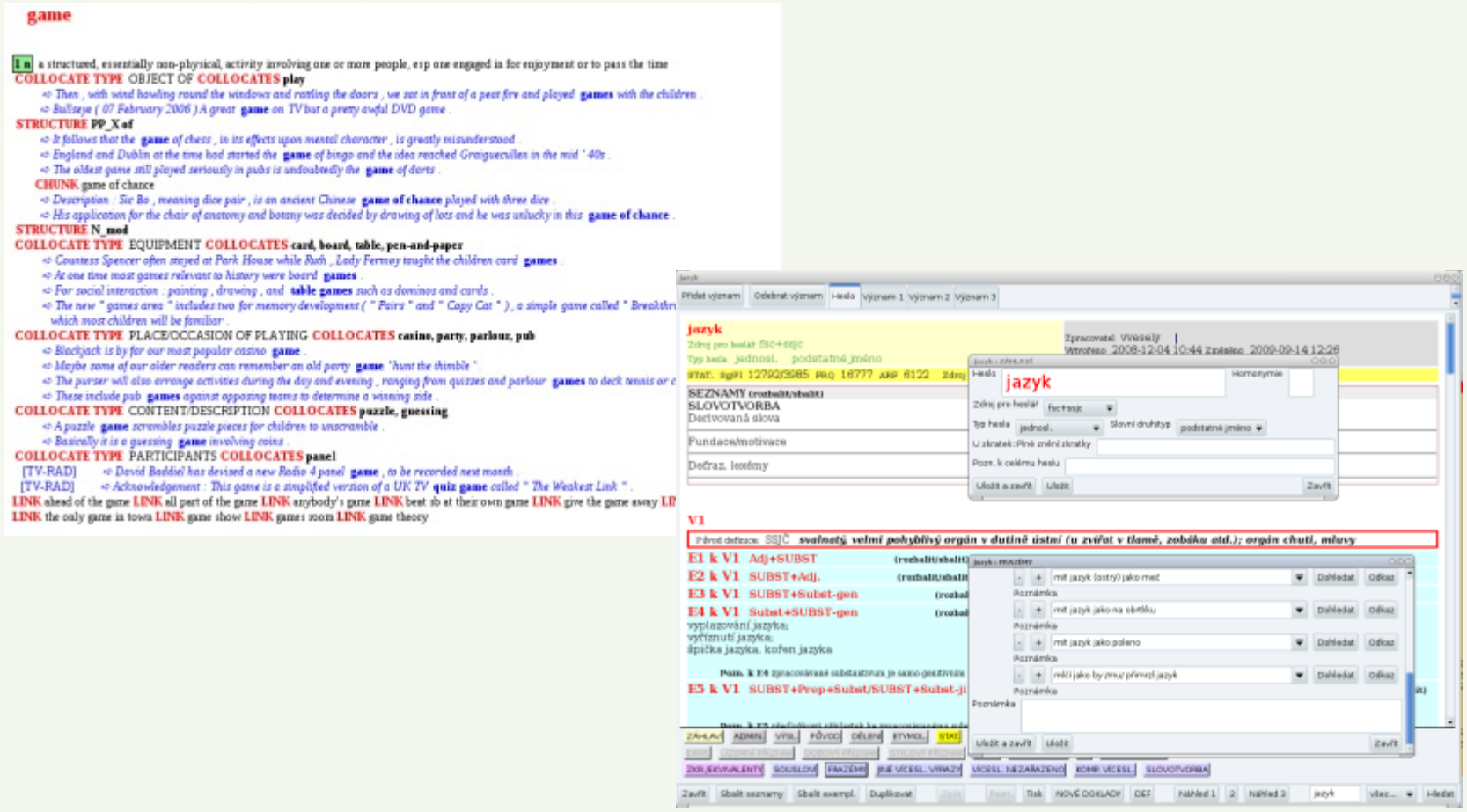
Korpus

- 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)

Lexikální databáze



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

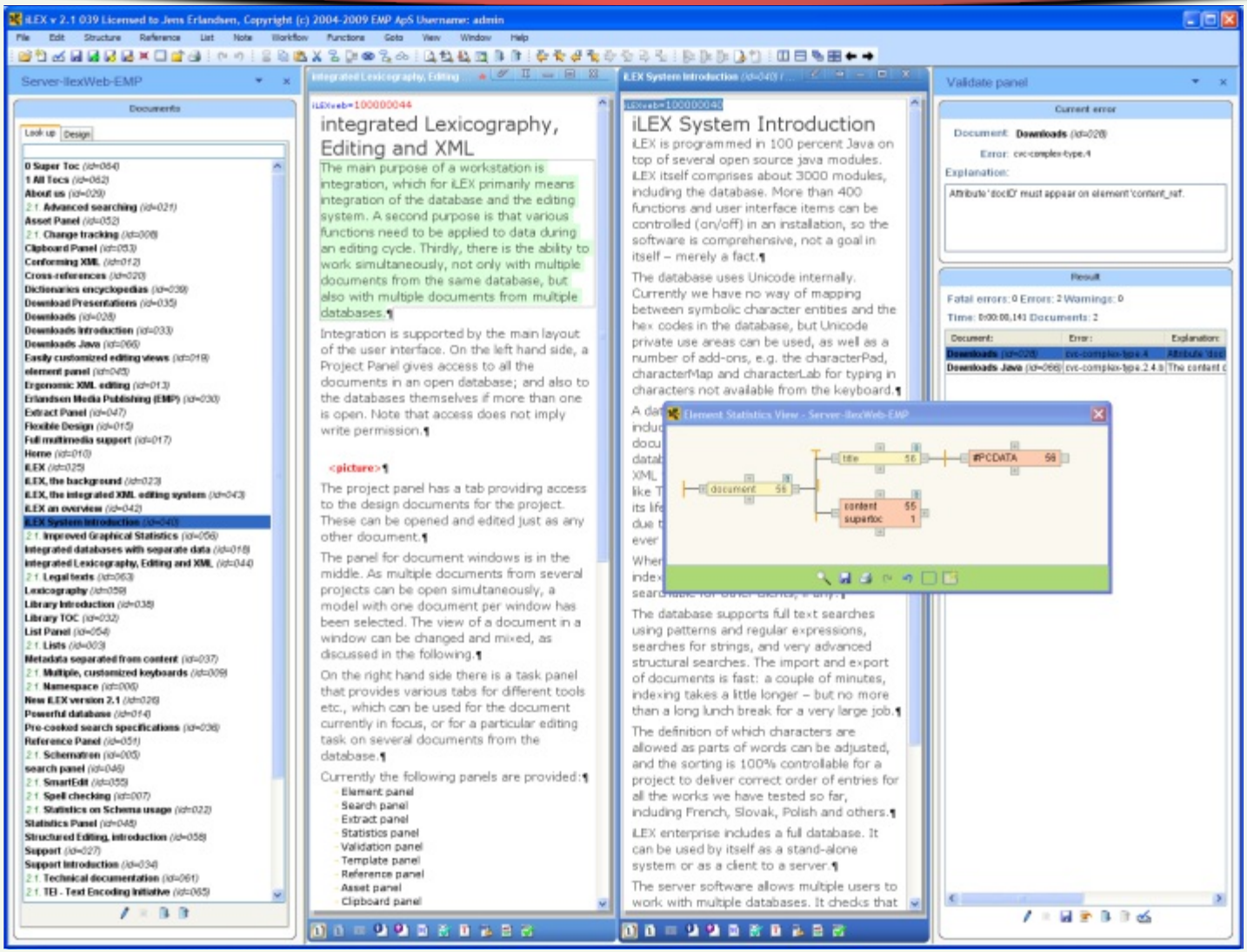
- Achilles
- SSJČ: jm. řeckého reka v Homérově Iliadě: Achillova pata, přen. zranitelné místo; každý člověk má svou Achillovu patu; med. Achillova šlacha upínající se na kost patní;
- SSČ: Achillova pata, zranitelné místo; Achillova šlacha, šlacha lýtkového svalu upínající se na patní kost

- všechna slova v definici musejí být v hesláři

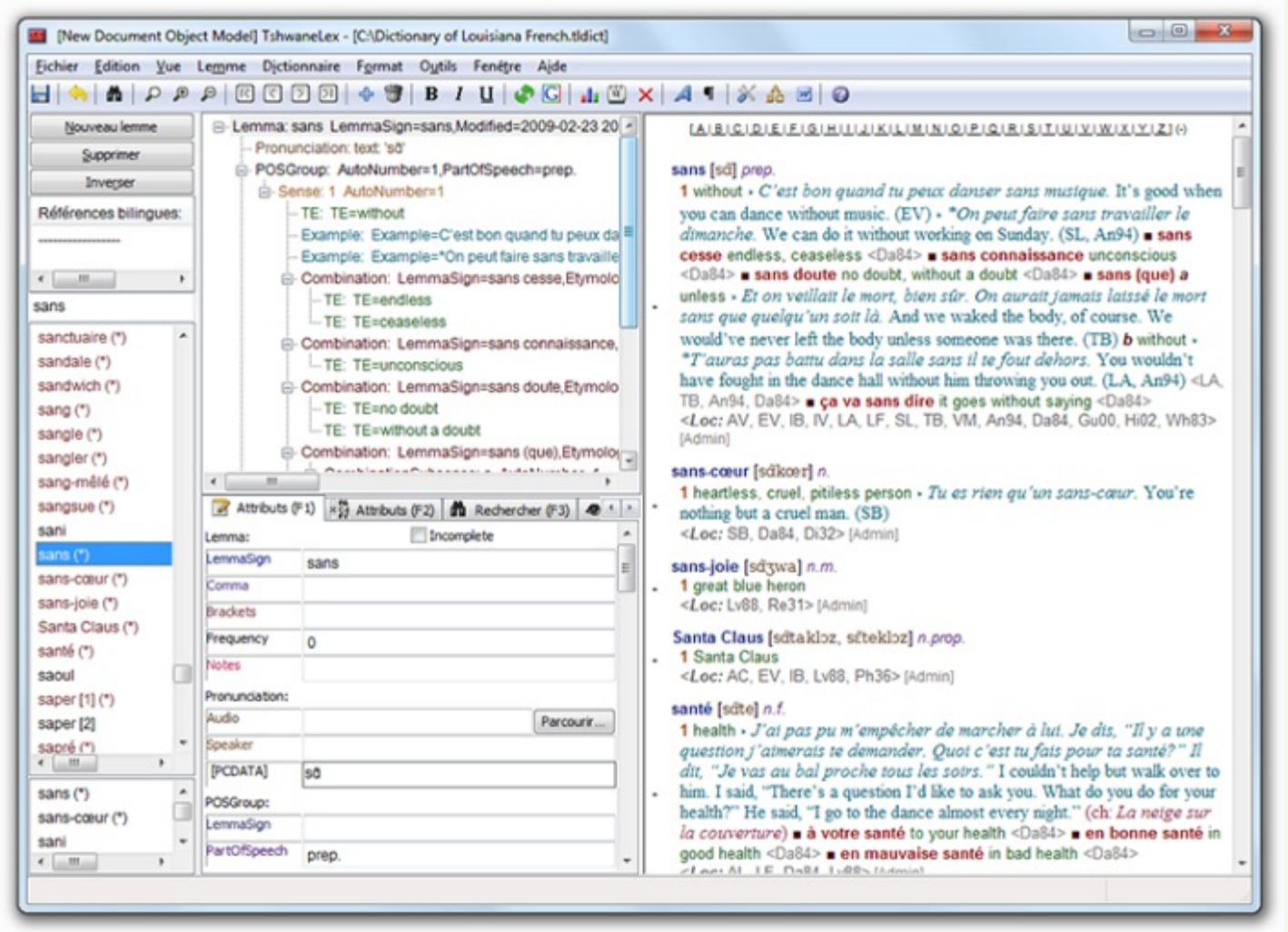
Dictionary Writing Systems

- 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)
- nekomerční (Glossword, Matapuna)
- DEB (Dictionary Editor and Browser)

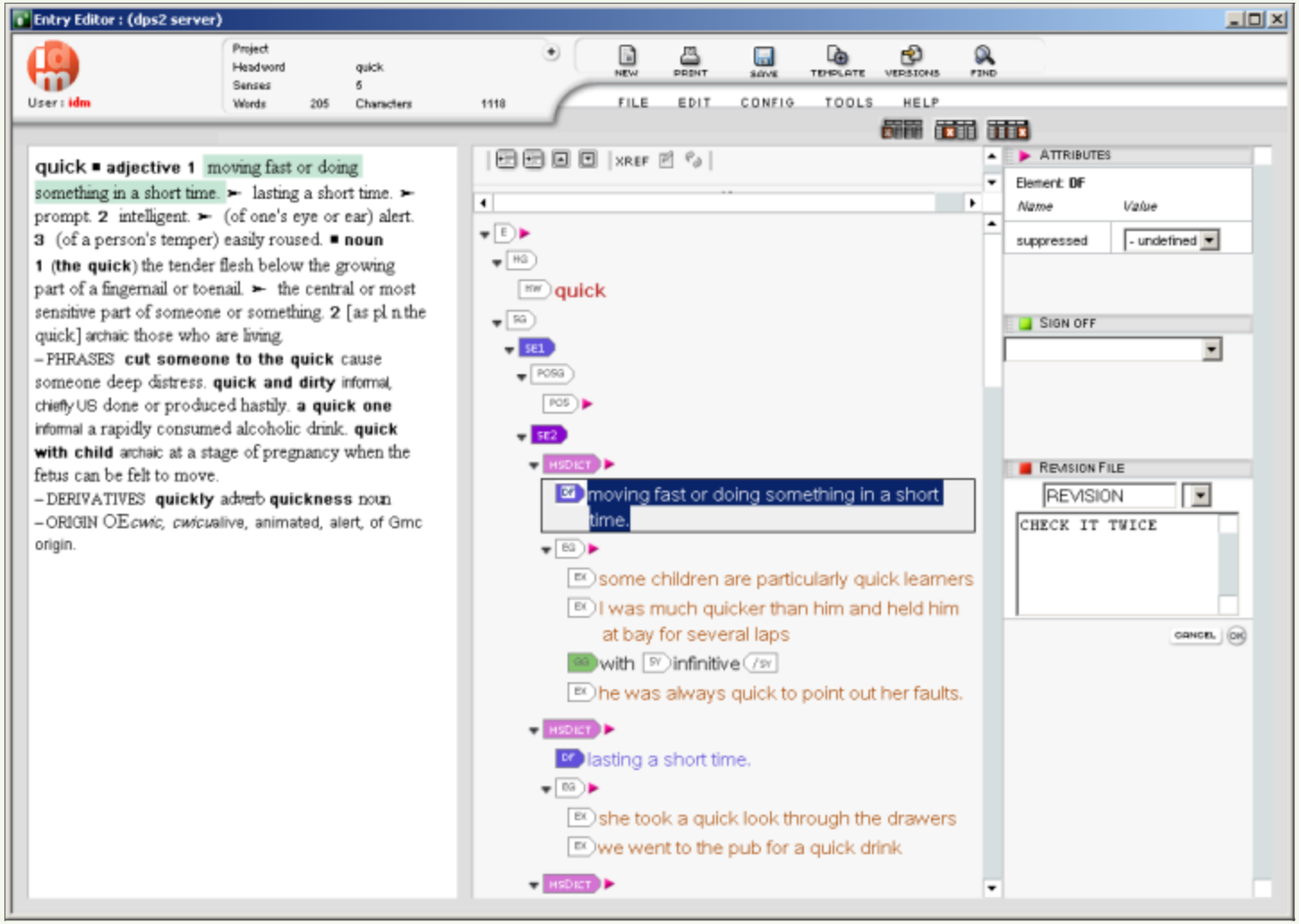
iLex


iLEX v 2.1 039 Licensed to Jens Erlandsen, Copyright (c) 2004-2009 EMP ApS Username: admin
File Edit Structure Reference List Note Workflow Functions Goto View Window Help
Server-IlexWeb-EMP
Documents
Look up Design
integrated Lexicography, Editing and XML
The main purpose of a workstation is integration, which for iLEX primarily means integration of the database and the editing system. A second purpose is that various functions need to be applied to data during an editing cycle. Thirdly, there is the ability to work simultaneously, not only with multiple documents from the same database, but also with multiple documents from multiple databases.
Integration is supported by the main layout of the user interface. On the left hand side, a Project Panel gives access to all the documents in an open database; and also to the databases themselves if more than one is open. Note that access does not imply write permission.
<picture>
The project panel has a tab providing access to the design documents for the project. These can be opened and edited just as any other document.
The panel for document windows is in the middle. As multiple documents from several projects can be open simultaneously, a model with one document per window has been selected. The view of a document in a window can be changed and mixed, as discussed in the following.
On the right hand side there is a task panel that provides various tabs for different tools etc., which can be used for the document currently in focus, or for a particular editing task on several documents from the database.
Currently the following panels are provided:
Element panel
Search panel
Extract panel
Statistics panel
Validation panel
Template panel
Reference panel
Asset panel
Clipboard panel
iLEX System Introduction
iLEX is programmed in 100 percent Java on top of several open source java modules. iLEX itself comprises about 3000 modules, including the database. More than 400 functions and user interface items can be controlled (on/off) in an installation, so the software is comprehensive, not a goal in itself – merely a fact.
The database uses Unicode internally. Currently we have no way of mapping between symbolic character entities and the hex codes in the database, but Unicode private use areas can be used, as well as a number of add-ons, e.g. the characterPad, characterMap and characterLab for typing in characters not available from the keyboard.
The database supports full text searches using patterns and regular expressions, searches for strings, and very advanced structural searches. The import and export of documents is fast: a couple of minutes, indexing takes a little longer – but no more than a long lunch break for a very large job.
The definition of which characters are allowed as parts of words can be adjusted, and the sorting is 100% controllable for a project to deliver correct order of entries for all the works we have tested so far, including French, Slovak, Polish and others.
iLEX enterprise includes a full database. It can be used by itself as a stand-alone system or as a client to a server.
The server software allows multiple users to work with multiple databases. It checks that
Element Statistics View - Server-IlexWeb-EMP
Validate panel
Current error
Document: Downloads (id=028)
Error: cvc-complex-type.4
Explanation:
Attribute 'docID' must appear on element 'content_ref'.
Result
Fatal errors: 0 Errors: 2 Warnings: 0
Time: 0:00:00.141 Documents: 2

TLex


[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 (*)
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, 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>

IDM DPS


Entry Editor : (dps2 server)
User : idm
Project
Headword quick
Senses 5
Words 205 Characters 1118
NEW PRINT SAVE TEMPLATE VERSIONS FIND
FILE EDIT CONFIG TOOLS HELP
quick ■ adjective 1 moving fast or doing something in a short time. ➢ lasting a short time. ➢ prompt. 2 intelligent. ➢ (of one's eye or ear) alert. 3 (of a person's temper) easily roused. ■ noun 1 (the quick) the tender flesh below the growing part of a fingernail or toenail. ➢ the central or most sensitive part of someone or something. 2 [as pl. n the quick] archaic those who are living.
– PHRASES cut someone to the quick cause someone deep distress. quick and dirty informal, chiefly US done or produced hastily. a quick one informal a rapidly consumed alcoholic drink. quick with child archaic at a stage of pregnancy when the fetus can be felt to move.
– DERIVATIVES quickly adverb quickness noun
– ORIGIN OE cwic, cwicu alive, animated, alert, of Gmc origin.
XREF
quick
moving fast or doing something in a short time.
some children are particularly quick learners
I was much quicker than him and held him at bay for several laps
with infinitive
he was always quick to point out her faults.
lasting a short time.
she took a quick look through the drawers
we went to the pub for a quick drink
ATTRIBUTES
Element: DF
Name Value
suppressed - undefined
SIGN OFF
REVISION FILE
REVISION
CHECK IT TWICE
CANCEL OK