Dictionary Writing Systems and Corpus Query Systems

Session 1, part 1

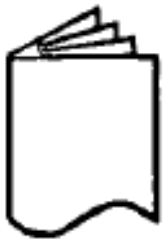
Laurence Urdang

- Dictionaries No. 6 (1984)
- A Lexicographer‘s Adventures in Computing
  - „I clearly recall my first brush with computers: in 1959, I was working on a new dictionary for Random House, and it occurred to me that a comptuer would be ideal for the sorting and manipulation of the various kinds of data I was responsible for generating.“

1966

- "As far as I know, the book we produced, *The Random House Dictionary of the English Language – The Unabridged Edition*, was the first to use computers extensively in its editorial preparation."
- "With all the data appropriately coded, programs enabled the computer to sort all of the bits and pieces into dictionary order."
- "In short, we created a powerful database, consisting of a dictionary with about 260,000 entries which we could access at will."

Euralex bulletin 1990



COMPULEXIS DICTIONARY SYSTEM

The only data-base system specially designed for lexicographers and dictionary publishers.

- Combines data-capture with editorial tools and typesetting.
- Developed by a team of lexicographers and computer specialists, and individually configured to suit the requirements of each dictionary project.
- Quick to learn and easy to use.
- Already successfully applied by Oxford University Press and other major European dictionary publishers.
- Runs on IBM AT or PS/2 or compatibles.

Specialized dictionary software

With the COMPULEXIS dictionary system ...

- The editor is free to concentrate on the "real" work rather than checking punctuation, typography, alphabetical order, etc.
- Proofs which look like the final product can be produced at any time, without external cost.
- Selective check-lists with cross-references and other problem areas can be printed.
- Structural punctuation, frequently used text elements, and numbers and letters used for subdivision can be generated automatically.
- Consistency is dramatically improved.
- Data can be re-used in new editions and spin-off products.

Oxford University Press

What the users think ...

"... We soon realized what fanatastic help we got from the COMPULEXIS system.
... For the editors the COMPULEXIS system has opened the door to all sorts of possibilities.
... The most conservative department in the house has in less than a year completely changed its working methods.
... The more we have worked with the system, the more our enthusiasm has grown."

(Quotes from an unpublished report by the Dictionary Department of a major European publisher)

Gestorlex 1991

- Nov/Dec 1991 issue of Language Industry Monitor
  - official launching of GestorLEX at the October Buchmesse in Frankfurt
  - developed by Danish software house TEXTware
  - hybrid product, a strong database engine providing the platform for a structured editing environment based on SGML
  - it runs under OS/2

SGML & floppy

- facilities to allow more than one person to work on a dictionary at a time
- can select and export a selection of entries on to a floppy disk for further editing on another machine
- selection is “locked” out in the main database although it can be viewed
- GESTORLEX costs DKR 15,000 (c.$2000)

Oxford-DZS 1994

- Oxford-Hachette French Dictionary
  - Oxford University Press
  - 1994, First Edition

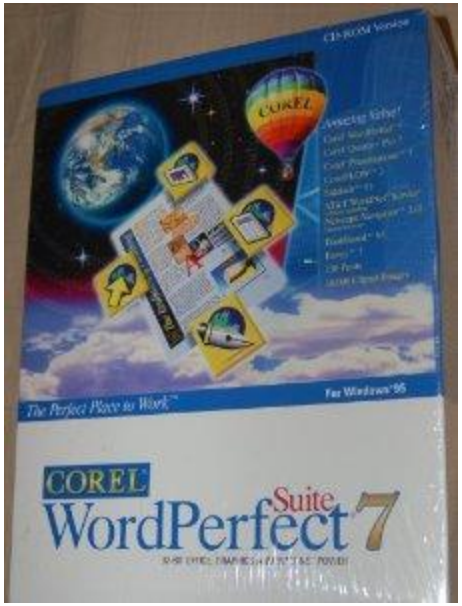
```
<se>
    <hw>abacus</hw>
    <pr><ph>"&b@k@s</ph></pr>
    <hg><ps>n</ps> (<gr>pl</gr>
    <f>-cuses</f>)</hg>
    <s2 num=1>
        <ann><la>Math</la></ann>(<ic>counting
    frame</ic>)
    </s2>;
    <s2 num=2><la>Archit</la> </s2>.
</se>
```

Gestorlex / WP7-8 (1995-)

- Problems – data:
  - database not in SGML
- Problems – Gestorlex:
  - expensive
  - lexicographers without OS/2
  - OS/2 demanding on hardware
- Solution – WordPerfect 7 (later 8, 11, X1)
  - worked on Windows (95)
  - affordable (≈60 people working at home – ~~LAN~~)
  - handled SGML format (!) with different layouts

WordPerfect

- general-purpose word processor
  - no management (!) – workflow, distribution etc.
- at the end of the project (custom tools)
  - consistency check (closed sets etc.)
  - proof-reading
  - punctuation
- during project
  - valid SGML (semi-restrictive DTD)
  - macros



Better solutions!

- Publishers home-grown systems: no
- Dedicated DWS candidates:
  - Digital Publishing System (DPS)
    - IDM (Ingénierie Diffusion Multimedia), Paris
  - iLex
    - Erlandsen Media Publishing AB, Malmö
  - TLex Suite (aka TshwaneLex)
    - TshwaneDJe Human Language Technology, Stellenbosch

Summary – DPS

- „perfect system“
- relatively complex to manage
- expensive
- for big projects (many collaborators)
- for big producers (many projects)
- XML less prominent (for lexicographers)
- very powerful management system

Summary – iLex

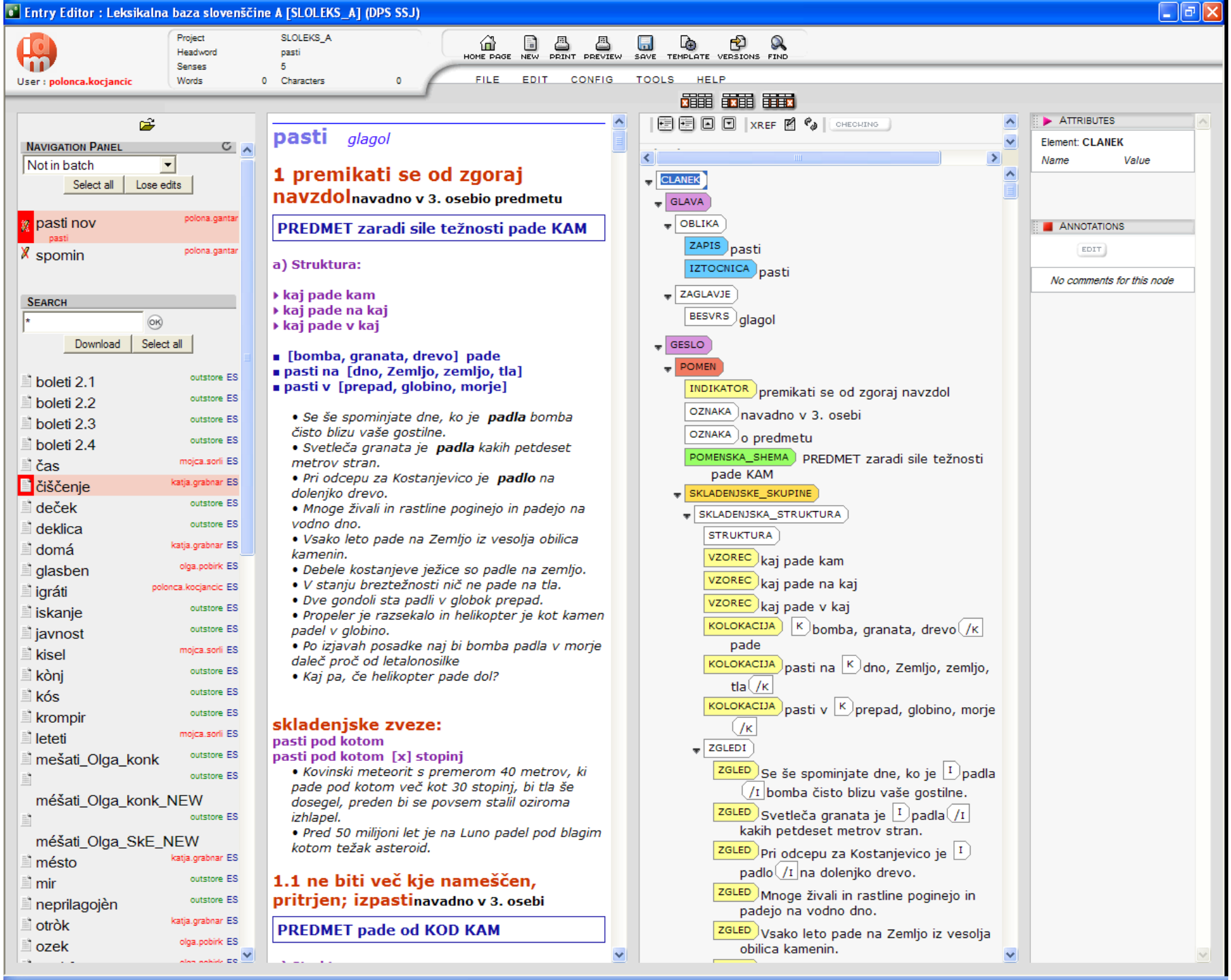
- „handy system“
- advanced computer literate manager
- not cheap but managable (start at ≈400 EUR)
- for collaborative projects
- also for solo projects
- „cafeteria style“
- explicitely XML-based and proud of it

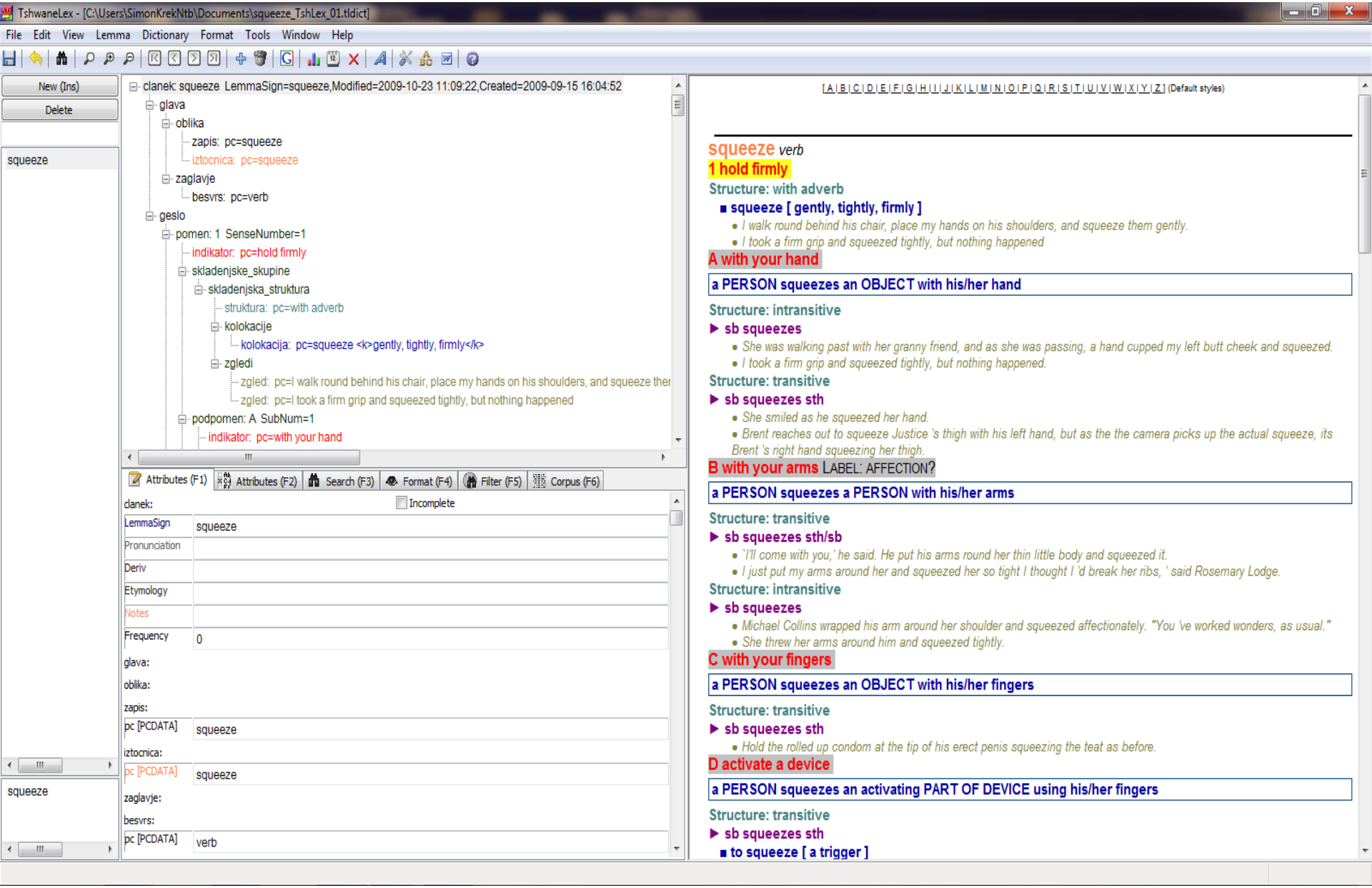
Summary – TLex

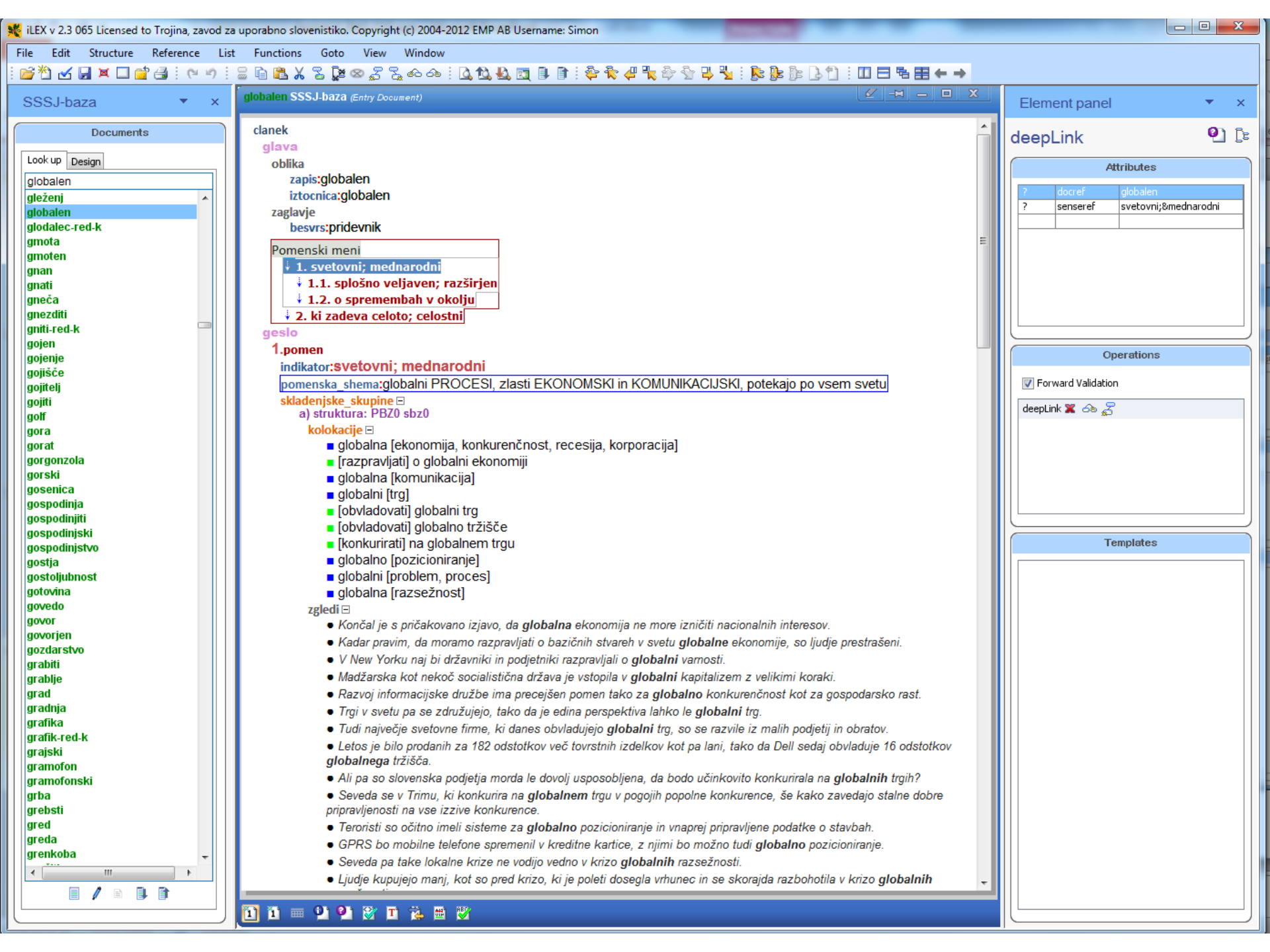
- „one-man band system“
- designed for computer literate lexicographers
- affordable (start at 150 EUR)
- mainly for solo projects
- less for collaboration (?)
- „suite style", "all-in one"
  - dictionary + corpora + terminology etc.
- database with XML output

Other

- Andrea Abel in the coming: *Electronic Lexicography*. Oxford University Press
  - off-the-shelf
    - ABBYY Lingvo Content
  - home-grown
    - DicSy, Norstedts
    - Wissensnetz deutsche Sprache, Duden
    - EELex, Institute for Estonian Language
    - ANW, Institute for Dutch Lexicology
    - Onoma, Woordeboek van die Afrikaanse Taal
    - Dictionary Editor and Browser (DEB), Masaryk University
    - Termania, Amebis (Slovenia)


Entry Editor : Leksikalna baza slovenščine A [SLOLEKS_A] (DPS SSJ)
User : polonca.kocjancic
Project SLOLEKS_A
Headword pasti
Senses 5
Words 0 Characters 0
HOME PAGE NEW PRINT PREVIEW SAVE TEMPLATE VERSIONS FIND
FILE EDIT CONFIG TOOLS HELP
NAVIGATION PANEL
Not in batch
Select all Lose edits
pasti nov polona.gantar
pasti
spomin polona.gantar
SEARCH
Download Select all
boleti 2.1 outstore ES
boleti 2.2 outstore ES
boleti 2.3 outstore ES
boleti 2.4 outstore ES
čas mojca.sorli ES
čiščenje katja.grabnar ES
deček outstore ES
deklica outstore ES
domá katja.grabnar ES
glasben olga.pobirk ES
igráti polonca.kocjancic ES
iskanje outstore ES
javnost outstore ES
kisel mojca.sorli ES
kònj outstore ES
kós outstore ES
krompir outstore ES
leteti mojca.sorli ES
mešati_Olga_konk outstore ES
méšati_Olga_konk_NEW outstore ES
méšati_Olga_SkE_NEW outstore ES
mésto katja.grabnar ES
mir outstore ES
neprilagojèn outstore ES
otròk katja.grabnar ES
ozek olga.pobirk ES
pasti glagol
1 premikati se od zgoraj navzdol navadno v 3. osebio predmetu
PREDMET zaradi sile težnosti pade KAM
a) Struktura:
kaj pade kam
kaj pade na kaj
kaj pade v kaj
[bomba, granata, drevo] pade
pasti na [dno, Zemljo, zemljo, tla]
pasti v [prepad, globino, morje]
• Se še spominjate dne, ko je padla bomba čisto blizu vaše gostilne.
• Svetleča granata je padla kakih petdeset metrov stran.
• Pri odcepu za Kostanjevico je padlo na dolenjko drevo.
• Mnoge živali in rastline poginejo in padejo na vodno dno.
• Vsako leto pade na Zemljo iz vesolja obilica kamenin.
• Debele kostanjeve ježice so padle na zemljo.
• V stanju breztežnosti nič ne pade na tla.
• Dve gondoli sta padli v globok prepad.
• Propeler je razsekalo in helikopter je kot kamen padel v globino.
• Po izjavah posadke naj bi bomba padla v morje daleč proč od letalonosilke
• Kaj pa, če helikopter pade dol?
skladenjske zveze:
pasti pod kotom
pasti pod kotom [x] stopinj
• Kovinski meteorit s premerom 40 metrov, ki pade pod kotom več kot 30 stopinj, bi tla še dosegel, preden bi se povsem stalil oziroma izhlapel.
• Pred 50 milijoni let je na Luno padel pod blagim kotom težak asteroid.
1.1 ne biti več kje nameščen, pritrjen; izpasti navadno v 3. osebi
PREDMET pade od KOD KAM
XREF CHECKING
CLANEK
GLAVA
OBLIKA
ZAPIS pasti
IZTOCNICA pasti
ZAGLAVJE
BESVRS glagol
GESLO
POMEN
INDIKATOR premikati se od zgoraj navzdol
OZNAKA navadno v 3. osebi
OZNAKA o predmetu
POMENSKA_SHEMA PREDMET zaradi sile težnosti pade KAM
SKLADENJSKE_SKUPINE
SKLADENJSKA_STRUKTURA
STRUKTURA
VZOREC kaj pade kam
VZOREC kaj pade na kaj
VZOREC kaj pade v kaj
KOLOKACIJA K bomba, granata, drevo /K pade
KOLOKACIJA pasti na K dno, Zemljo, zemljo, tla /K
KOLOKACIJA pasti v K prepad, globino, morje /K
ZGLEDI
ZGLED Se še spominjate dne, ko je I padla /I bomba čisto blizu vaše gostilne.
ZGLED Svetleča granata je I padla /I kakih petdeset metrov stran.
ZGLED Pri odcepu za Kostanjevico je I padlo /I na dolenjko drevo.
ZGLED Mnoge živali in rastline poginejo in padejo na vodno dno.
ZGLED Vsako leto pade na Zemljo iz vesolja obilica kamenin.
ATTRIBUTES
Element: CLANEK
Name Value
ANNOTATIONS
EDIT
No comments for this node


TshwaneLex - [C:\Users\SimonKrekNtb\Documents\squeeze_TshLex_01.tldict]
File Edit View Lemma Dictionary Format Tools Window Help
New (Ins)
Delete
squeeze
squeeze
clanek: squeeze LemmaSign=squeeze,Modified=2009-10-23 11:09:22,Created=2009-09-15 16:04:52
glava
oblika
zapis: pc=squeeze
iztocnica: pc=squeeze
zaglavje
besvrs: pc=verb
geslo
pomen: 1 SenseNumber=1
indikator: pc=hold firmly
skladenjske_skupine
skladenjska_struktura
struktura: pc=with adverb
kolokacije
kolokacija: pc=squeeze <k>gently, tightly, firmly</k>
zgledi
zgled: pc=I walk round behind his chair, place my hands on his shoulders, and squeeze ther
zgled: pc=I took a firm grip and squeezed tightly, but nothing happened
podpomen: A SubNum=1
indikator: pc=with your hand
Attributes (F1)
Attributes (F2)
Search (F3)
Format (F4)
Filter (F5)
Corpus (F6)
clanek:
Incomplete
LemmaSign squeeze
Pronunciation
Deriv
Etymology
Notes
Frequency 0
glava:
oblika:
zapis:
pc [PCDATA] squeeze
iztocnica:
pc [PCDATA] squeeze
zaglavje:
besvrs:
pc [PCDATA] verb
[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] (Default styles)
squeeze verb
1 hold firmly
Structure: with adverb
■ squeeze [gently, tightly, firmly]
• I walk round behind his chair, place my hands on his shoulders, and squeeze them gently.
• I took a firm grip and squeezed tightly, but nothing happened
A with your hand
a PERSON squeezes an OBJECT with his/her hand
Structure: intransitive
▶ sb squeezes
• She was walking past with her granny friend, and as she was passing, a hand cupped my left butt cheek and squeezed.
• I took a firm grip and squeezed tightly, but nothing happened.
Structure: transitive
▶ sb squeezes sth
• She smiled as he squeezed her hand.
• Brent reaches out to squeeze Justice 's thigh with his left hand, but as the the camera picks up the actual squeeze, its Brent 's right hand squeezing her thigh.
B with your arms LABEL: AFFECTION?
a PERSON squeezes a PERSON with his/her arms
Structure: transitive
▶ sb squeezes sth/sb
• `I'll come with you,' he said. He put his arms round her thin little body and squeezed it.
• I just put my arms around her and squeezed her so tight I thought I 'd break her ribs, ' said Rosemary Lodge.
Structure: intransitive
▶ sb squeezes
• Michael Collins wrapped his arm around her shoulder and squeezed affectionately. "You 've worked wonders, as usual."
• She threw her arms around him and squeezed tightly.
C with your fingers
a PERSON squeezes an OBJECT with his/her fingers
Structure: transitive
▶ sb squeezes sth
• Hold the rolled up condom at the tip of his erect penis squeezing the teat as before.
D activate a device
a PERSON squeezes an activating PART OF DEVICE using his/her fingers
Structure: transitive
▶ sb squeezes sth
■ to squeeze [a trigger]

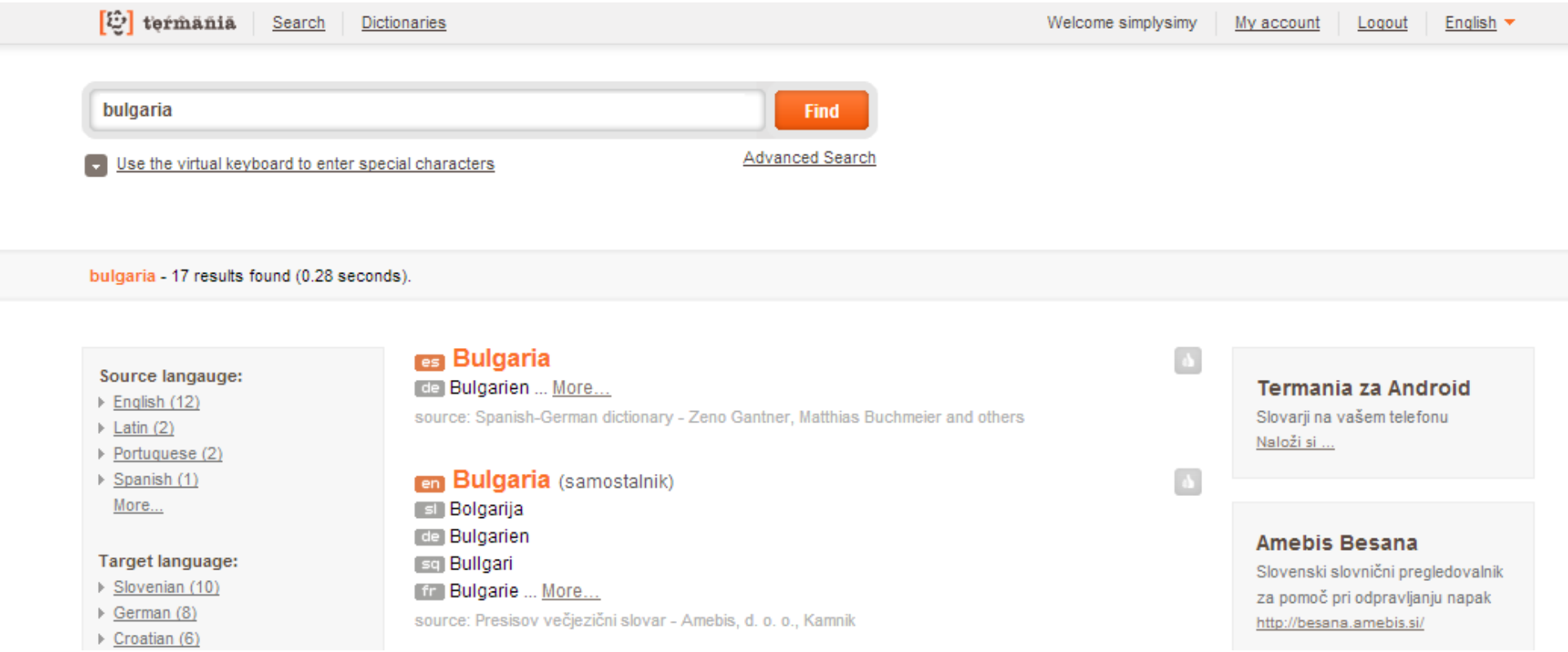

iLEX v 2.3 065 Licensed to Trojina, zavod za uporabno slovenistiko. Copyright (c) 2004-2012 EMP AB Username: Simon
File Edit Structure Reference List Functions Goto View Window
SSSJ-baza
Documents
Look up Design
globalen
gleženj
globalen
globalec-red-k
gmota
gmoten
gnan
gnati
gneča
gnezditi
gniti-red-k
gojen
gojenje
gojišče
gojitelj
gojiti
golf
gora
gorat
gorgonzola
gorski
gosenica
gospodinja
gospodinjiti
gospodinjski
gospodinjstvo
gostja
gostoljubnost
gotovina
govedo
govor
govorjen
gozdarstvo
grabiti
grablje
grad
gradnja
grafika
grafik-red-k
grajski
gramofon
gramofonski
grba
grebsti
gred
greda
grenkoba
globalen SSSJ-baza (Entry Document)
clanek
glava
oblika
zapis:globalen
iztocnica:globalen
zaglavje
besvrs:pridevnik
Pomenski meni
1. svetovni; mednarodni
1.1. splošno veljaven; razširjen
1.2. o spremembah v okolju
2. ki zadeva celoto; celostni
geslo
1.pomen
indikator:svetovni; mednarodni
pomenska_shema:globalni PROCESI, zlasti EKONOMSKI in KOMUNIKACIJSKI, potekajo po vsem svetu
skladenjske_skupine
a) struktura: PBZ0 sbz0
kolokacije
globalna [ekonomija, konkurenčnost, recesija, korporacija]
[razpravljati] o globalni ekonomiji
globalna [komunikacija]
globalni [trg]
[obvladovati] globalni trg
[obvladovati] globalno tržišče
[konkurirati] na globalnem trgu
globalno [pozicioniranje]
globalni [problem, proces]
globalna [razsežnost]
zgledi
Končal je s pričakovano izjavo, da globalna ekonomija ne more izničiti nacionalnih interesov.
Kadar pravim, da moramo razpravljati o bazičnih stvareh v svetu globalne ekonomije, so ljudje prestrašeni.
V New Yorku naj bi državniki in podjetniki razpravljali o globalni varnosti.
Madžarska kot nekoč socialistična država je vstopila v globalni kapitalizem z velikimi koraki.
Razvoj informacijske družbe ima precejšen pomen tako za globalno konkurenčnost kot za gospodarsko rast.
Trgi v svetu pa se združujejo, tako da je edina perspektiva lahko le globalni trg.
Tudi največje svetovne firme, ki danes obvladujejo globalni trg, so se razvile iz malih podjetij in obratov.
Letos je bilo prodanih za 182 odstotkov več tovrstnih izdelkov kot pa lani, tako da Dell sedaj obvladuje 16 odstotkov globalnega tržišča.
Ali pa so slovenska podjetja morda le dovolj usposobljena, da bodo učinkovito konkurirala na globalnih trgih?
Seveda se v Trimu, ki konkurira na globalnem trgu v pogojih popolne konkurence, še kako zavedajo stalne dobre pripravljenosti na vse izzive konkurence.
Teroristi so očitno imeli sisteme za globalno pozicioniranje in vnaprej pripravljene podatke o stavbah.
GPRS bo mobilne telefone spremenil v kreditne kartice, z njimi bo možno tudi globalno pozicioniranje.
Seveda pa take lokalne krize ne vodijo vedno v krizo globalnih razsežnosti.
Ljudje kupujejo manj, kot so pred krizo, ki je poleti dosegla vrhunec in se skorajda razbohotila v krizo globalnih
Element panel
deepLink
Attributes
? docref globalen
? senseref svetovni;&mednarodni
Operations
Forward Validation
deepLink
Templates

Issues

- management tool
  - control workflow, monitor progress, distribute batches, backups etc.
- publishing tool
  - reusability, direct transfer (online, DTP)
- data management tool
  - reversing (bilinguals), data extraction etc.
- future
  - integration with browsers
  - online editing (mobile, cloud computing)

Termania – quick demo

- www.termania.net



Corpus Query Systems

Session 1, part 2

Topics

- We will talk about:
  - CQS & lexicographers
    - personal experience from Sara to the Sketch Engine
  - CQS & language learning/teaching
    - Gigafida interface
- We won't talk about
  - corpus design & annotation
  - corpora & NLP

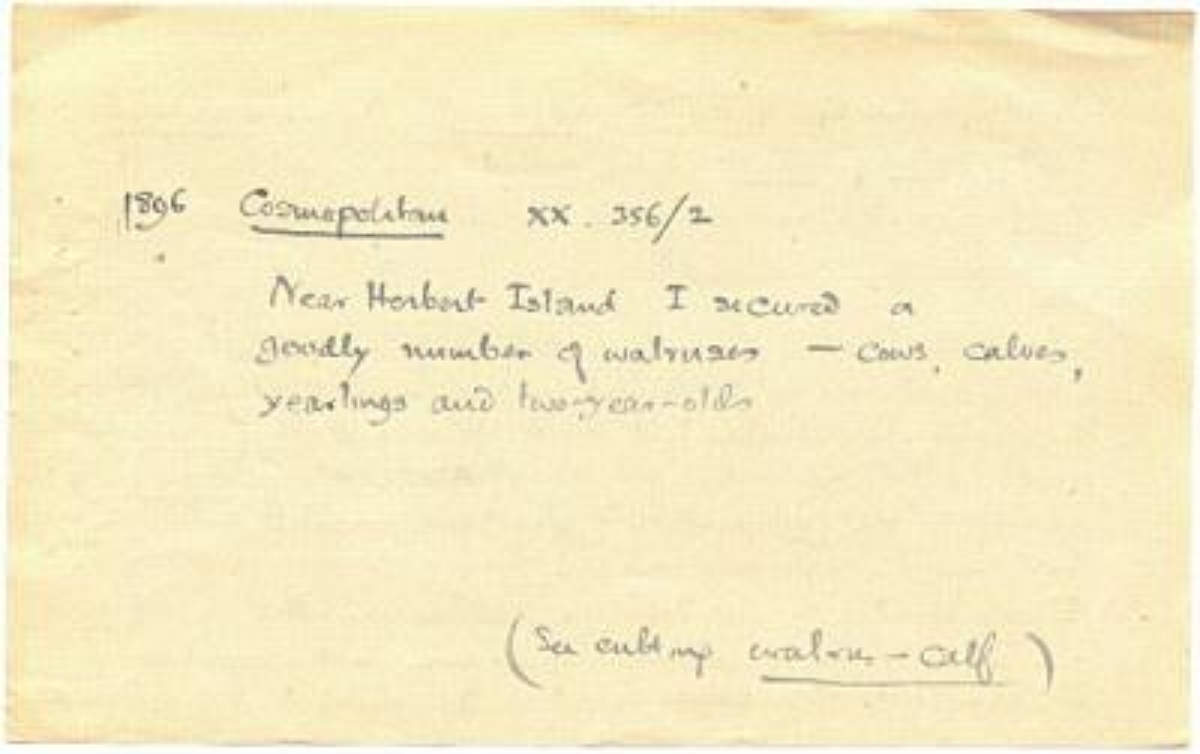
Lexicographers

- Data collection:
  - corpora and other forms of evidence
- Data analysis:
  - sifting the evidence to discover relevant facts
- Synthesis:
  - creating dictionary entries

Data analysis in 2010s

- Lexicographer's changing role
  - from scanning data, to identify lexicographically-relevant facts
  - to validating (or rejecting) decisions made by computer
- New role
  - Identify/describe what can be automated & expand set of automatable processes
  - Identify weaknesses in support software

Data analysis (OED 1910s)



1896 Cosmopolitan XX. 356/2

Near Herbert Island I secured a goodly number of walruses — cows, calves, yearlings and two-year-olds

(See cultivate walrus-calf)

Data analysis (COBUILD 1980s)

	br	l32	ook and cranny in the vessel where even a stray	seagull could be hiding and you can take my word
m	br	25	er. "every wave on the atlantic was like a dead	seagull drag- ing its driftwood artillery from h

8 seal

	br	127	e king's taster?' i looked at the unbroken lead	seal. "not uness you think some- one has brought
r	br	129	church until 1835. years later, galileo put his	seal on copernicus's discovery, wvas hauled up b
r	br	140	her lover to assuage her inner doubts, set the	seal on her femininity, provide her with psychic
m	br	172	brooding darkncss is lifted? could the seventh	seal or winter light have been conceived in anot
m	br	14	that never cleaned anything away, heavy thermal	seal over diesel fuel, mildew, garbage, excremen
a	br	34	s foot in it. lynn tried to be gracious but the	seal was set on her dislike of him. and somethin
	br	127	plant aboard.' "i've checked.' smithy broke the	seal. "we talked last night. at least, i did. yo
a	br	82	ce she discovered that, lynn thought, the final	seal would be set on jane's hatred and rebellion

10 sealed

r	br	133	ingenue. both their fates were, to some extent,	sealed. after "bunty" closed he went sady back t
r	br	129	place strips of the paper in a thin rubber tube	sealed at one end and connected to suction at th
a	br	151	d as superior and knowledgeable. a partnership.	sealed by why? so many exquisite little symmetri
a	br	138	g was led.' the europea party swept to dover in	sealed cars through back streets. "you were a lo
a	br	135	ss asked "thank you. i am not fond of salad.' a	sealed envelope passed to the prime minister wit
	br	86	c forms and filled them out. i put those in one	sealed envelope, the signed affidavit - i just
a	br	132	ottage he would flee to when all was signed and	sealed. he hadn't had a proper night's sleep for
	br	80	ed. "on a night like this? no fear. the gash is	sealed in polythene bags, then they're punctured
r	br	84	lions of years but his doom, paradoxically, was	sealed in the very fact that he became too perfe
r	br	199	ote out the telegram, put it into its envelope,	sealed it and handed it over to dolly. the four

2 sealing

m	br	48	gon stream thinning and trickling out: frontier	sealing, cencus grievance, black operations (pre
r	br	21	m each other. our once one-flesh divided again,	sealing me into me, him into him. he is now a te

3 seals

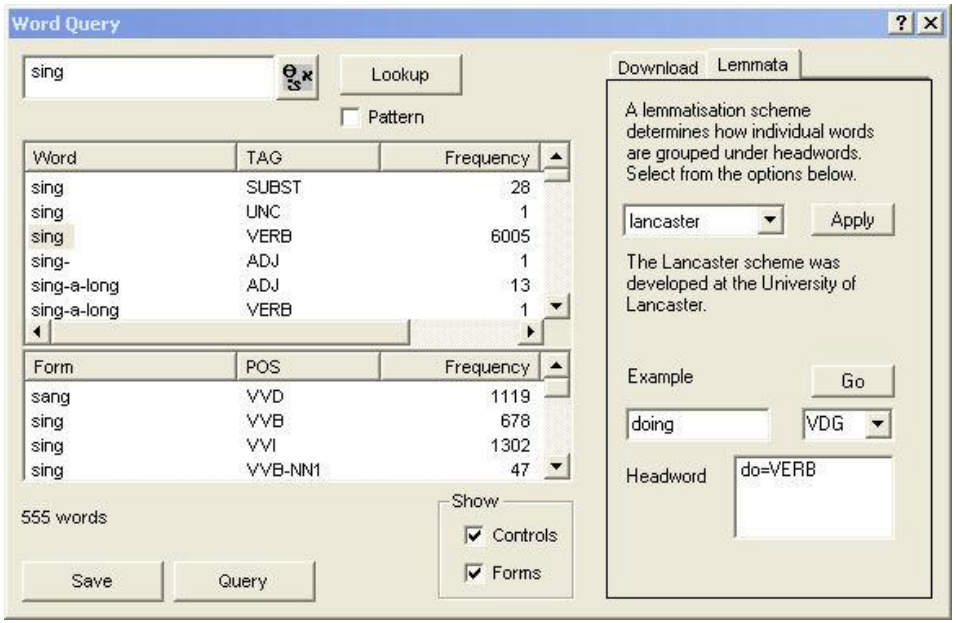
m	br	3	saw a row of old houses, huddled toge'her like	seals on a rock. then there was a long field tha
m	br	35	ang we'd get stone together and keep the lurps,	seals, recondos, green-beret bushmasters redunda
r	br	161	omen serves only their own artificial needs and	seals them off in their folie a deux from the re

Hanks on COBUILD

- "For the first time ever, lexicographers had evidence that enabled them to begin to see how words actually fitted together."
- "Opening a fresh concordance for all uses of a given word in a 7.3 million word corpus was like opening a window on a landscape of fresh snow on a sunny winter's morning."
- "Very often, a simple right-sort of concordances revealed patterns of usage that were at the same time unexpected and yet obvious (once seen)."

BNC – BoE (1995-)

- SARA (SGML-Aware Retrieval Application)
- Bank of English (HarperCollins)
  - 50 million (http://titania.cobuild.collins.co.uk/form.html)



<w TO0>to <w VBI>be <w VVN>sung<c PUN>, <w AV0>thereby <
AV0>then <w AV0>half <w VVN>sung <w PNP>it<c PUN>:<poem>
T0>the <w NN2>ladies <w VVN>sung <w NN1-NP0>Erse <w NN2>
w AV0>only <w VBI>be <w VVN>sung <w CJC>but <w AV0>also <
X0>not <w AV0>often <w VVN>sung<c PUN>, <c PUQ>&bquo;<w
AJ0>novel <w CJC>and <w VVN>sung <w PRP>by <w AJ0>Junior <
<w NN1>solo<c PUN>, <w VVN>sung <w PRP>to <w DPS>his <w
)>vocal <w NN1>music <w VVN>sung <w PRP>by <w AT0>the <w
girlie <w CJT-DT0>that <w VVN>sung <w AT0>the <w NN1>song
w PNP>them <w CJS>as <w VVN>sung <w PNP>them<c PUN>.</u>

FIDA corpus (1997-2000)

- online (dial-up, 56 kbit/s modems)
- 100 million words
- POS-tagged (rule-based tagger)
- (complex) metadata
- consortium: **publisher**, LT company, university, technical institute
  - free for partners, restricted access
- purpose: lexicography, education, LT

Enovrstično iskanje
Razširjeno iskanje
Rezultati iskanj
Spiski (datoteke)
Pomoč
Nastavitve
Novice
Servisna stran

ODJAVA

Kontaktni naslov:
Korpus FIDA
DZS, Založništvo literature
Mestni trg 26
1538 Ljubljana
fida@dzs.si

Kratek opis točk

Enovrstično iskanje: Enostaven ali zahteven pogoj se napiše v eni vrstici. Uporabljajo se lahko logični in besedni operatorji. Postavijo se lahko tudi najbolj zapleteni pogoji, seveda pa morate biti vešči uporabe raznovrstnih operatorjev.

Razširjeno iskanje: Dodatni iskalni pogoji.

Rezultati iskanj: Shranjeno je do 10 prejšnjih iskanj.

Spiski (datoteke): Delo s spiski konkordanc in statistike (pregled, brisanje), ki smo jih shranili v besedilne datoteke.

Nastavitve: Uporabniške nastavitve.

Novice: Novice o uporabniškem vmesniku.

Enovrstično iskanje:

#1 zlikovec

Poišči Obriši

- kode MSD
- zapis posebnih znakov pri iskanju

Vrsta iskanja	Primer	Opis primera
enostavno iskanje besede	mize	poišče vse pojavitve besede "mize"
iskanje z nadomestnimi znaki (? nadomesti eno črko, * nadomesti poljubno zaporedje črk)	miz* miz?	poišče vse pojavitve besed, ki se začnejo z nizom "miz" poišče vse pojavitve štiričrkovnih besed, ki se začnejo z nizom "miz"
iskanje po kanalih (#1 lemma, #2 msd, #3 lemmas, #4 msds)	#1miza #2pkomein	poišče vse pojavitve besed z osnovno obliko "miza" poišče vse pojavitve pridevnikov (kakovostnih, nestopnjevanih...)
iskanje po frazah	#1okrogel_#1miza ki_*_je	poišče vse pojavitve, kjer sta zapored besedi z lemama "okrogel" in "miza" poišče vse pojavitve, kjer je med besedama "ki" in "je" še natanko ena beseda
iskanje po bližini (// za privzeto bližino ali /0 do /9 za število vmesnih besed)	#3stol//#3miza se/0je	poišče vse pojavitve, kjer sta lemi "stol" in "miza" blizu skupaj poišče vse pojavitve, kjer med "se" in "je" ni vmesnih besed (se_je ali je_se)
notranji in	#3voda&#3vod	poišče vse pojavitve besed, pri katerih je možna tako lema "voda" kot "vod"
notranji ne	#3vod&~#1vod	poišče vse pojavitve besed z možno lemo "vod", kjer je ta možnost pri analizi odpadla
in	se je	poišče vse pojavitve besed "se" in "je" v odstavkih, v katerih nastopata obe besedi


Večvrstično iskanje - okvir - Microsoft Internet Explorer provided by Nova Ljubljanska banka d.d.
Razširjeno iskanje:
Pogoj:
- kode MSD
- zapis posebnih znakov pri iskanju
Od leta:
Do leta:
Prenosnik:
vsa besedila
Ft.P prenosnik
Ft.P.G govorni
Ft.P.E elektronski
Ft.P.P pisni
Ft.P.P.O objavljeno
Ft.P.P.O.K knjižno
Ft.P.P.O.P periodično
Ft.P.P.O.P.C časopisno
Ft.P.P.O.P.C.D dnevno
Ft.P.P.O.P.C.V večkrat tedensko
Ft.P.P.O.P.C.T tedensko
Ft.P.P.O.P.R revialno
Ft.P.P.O.P.R.T tedensko
Ft.P.P.O.P.R.S štirinajstdnevno
Ft.P.P.O.P.R.M mesečno
Ft.P.P.O.P.R.D redkeje kot na mesec
Ft.P.P.O.P.R.O občasno
Ft.P.P.N neobjavljeno
Zvrst:
vsa besedila
Ft.Z zvrst
Ft.Z.U umetnostna
Ft.Z.U.P pesniška
Ft.Z.U.R prozna
Ft.Z.U.D dramska
Ft.Z.N neumetnostna
Ft.Z.N.S strokovna
Ft.Z.N.S.H humanistična in družboslovna
Ft.Z.N.S.N naravoslovna in tehnična
Ft.Z.N.N nestrokovna
Lektorirano:
vsa besedila
Ft.L lektorirano
Ft.L.D da
Ft.L.N ne

IZVOR	ODSTAVEK	KONKORDANCA
0021271.0000004	toda če kje obstaja kdo, ki je zares pokvarjen	zlikovec , nevaren človek, izdajalec d
0021404.0000053	Postojna - Neznani	zlikovec je v Postojni vlomil v osebni
0021404.0000055	oškodoval za približno 8000 tolarjev. Nedaleč stran pa je	zlikovec pred neko stanovanjsko hišc
0021450.0000893	da si ga bodo za vedno zapomnili tako kolegi kot	zlikovci . Ima dar, da se med akcijam
0021450.0000895	Nekega dne dva	zlikovca oropata neko bergamsko ba
0021486.0000037	Ilirskobistriški policisti paiščejo	zlikovca , ki je med zadnjim vikendor
0021608.0000237	smrti ne gre nihče v pekel. V Duhu so	zlikovce kakopak čakale hude, peko
0021766.0000003	lastnino, odkrivajo mamila, vlečejo nam sani, preganjajo	zlikovce , lovcem prinašajo divjad, v
0021830.0000034	o dogodku, ko je z železniške postaje Štore neznani	zlikovec v lokomotivo potniškega vla
0021865.0000577	Meek je še en mojster borilnih veščin, ki mu	zlikovci ubijejo starše. Oh, ko bi prej
0013788.0000008	na prvo hišo? Naj si mislijo, da je	zlikovec pač nekje na tem območju,
0013788.0000042	-ja in ga bomo v zgodbi o policijskem iskanju	zlikovca
0014192.0000084	Lopati, last A. L. iz Kočevja.	Zlikovec je tudi odlomil pipo na večji
0014352.0000287	Nekakšen	zlikovec z melono in psom mi je izma
0011028.0000002	Sezona	zlikovcev je tu
0011028.0000004	polnem razmahu, temu primerno pa so se organizirali tudi	zlikovci . 'Ordinirajo' po plažah, vlam
0011028.0000005	bilo škode za 30 tisočakov. V začetku tedna je	zlikovec iz kioska na plaži v Portorož
0011319.0000013	Po dveh letih in pol je neznanega	zlikovca spet zapekla vest: šel je pon
0011319.0000016	Nekulturno dejanje neznanega	zlikovca je prijavljeno policiji v Šiški,
0011324.0001691	opisuje zaradi humane razneženosti zaslepljeno tretmansko oko. Gre za	zlikovce , hudodelce, žeparje, tatove
0011410.0000072	ima F. K. iz Novega mesta barako.	Zlikovec je razbil vse mize in klopi pr
0011662.0000262	ekscesni negativci, na smrt obsojene zverine, morilski krvoločni	zlikovci najhujše sorte, padejo v nem
0011662.0000262	negativcev, še hujših zverin, ja, morilskih krvoločnih	zlikovcev še hujše sorte, okej, neke v

Statistika

Iskanje po: osnovnih oblikah besed

Besed pred zadetkom: 1

Besed za zadetkom: 0 ✓izvedi

Urejanje

Opozorilo: urejanje daljših seznamov lahko traja precej dolgo. Ne priporočano urejanje seznamov daljših od 5000 besed.

Urejanje po: besedah spredaj

Upoštevaj označeno besedo ☐

✓izvedi

Sito

KONKORDANCA

pa vem le to, da nič ne vem. Bojim se Boga kot hudiča . Ali v vas lahko zaupam?
, je po papeževi želji leto Očeta. Pa tudi hudiča se ni treba bati , ker ga ima Bog v oblasti in nam ne
od antidepresivov začneš rediti. Vse od takrat se jih bojim kot hudič križa.)
Tako slovenstvo nekoč in danes kot vsak človek sta stalen boj med Bogom in hudičem .
prastari so tudi ljudje, ki se trave še vedno bojijo kot hudič križa.
izstopiti iz koalicije s SDS (tega se Alojz Peterle boji kot hudič križa, zato je že napovedal možnost, da se
. Razlog? Izjemno visoke kazni, katerih se vsi bojijo kot hudič križa.
emnoge pustolovščine: junaška druščina se na svojih vandranjih ni bala ne boga ne hudiča , možje so trepetali samo pred meglo, iz
roke za sedeče kinoušesno-drugogodbeno občinstvo, ki se boji dratarjev in tehnologov kot hudič križa?
noben ni upal iti čez, ker so se vsi bali hudiča .
prodajalci, na primer tile iz ALG, se javnosti bojijo kot hudič križa. Najprej, globoko v osemdesetih, je bila
! Le kako sem zašel v ta pekel." Hudič mu je začel v predsmrtni boj prišepetavati besede, ki jih j
bi zame poiskali knjige o vedeževanju. Tarota sem se bala kot hudiča . Na začetku sem vedeževala iz običajnih kart.
reprezentance BiH, velemojster, ki se za šahovnico ne boji niti hudiča , igralec, ki že od mladostnih nog ne spoštuje
vrsto industrijske politike, ki se je sicer ameriška ideologija boji kot hudič križa. Ne glede na motiv: če je država
, ki se, če smo malce ironični, kot hudič križa boji javne diskusije o človekovi prvinskosti. Strip je
in je z njo naravnost obseden. Jaše ga "hudič, nečista žival", boji se glasu krvi. Čim bolj ta glas duši,
Rim - Italijani se kot hudič križa bojijo množičnega prihoda beguncev s Kosova. Za v
tekstom, a tudi v takih primerih se avtorji kakor hudič križa bojijo pridodati denimo glagole, s katerimi je često v
-reprezentanco ali mlade, predvsem Primoža pa so se bali "kot hudič križa". Njegova velika kristalna globusa sta bila n
To je krik sprevržene vere, vere hudičev in brezbožnežev: Boga se bojijo .
v dveh tekmah. Morebitnega tretjega srečanja v Ljubljani se boji kot hudič križa, tudi zato, ker bo sodil Litvanec Brazauskas
Bolnišnice in dohtarjev se bojim kot hudič križa, takrat sem se jih veselil v čudni vznesenos

FidaPLUS (2003-2006)

- online (slowly to broadband)
- 620 million words
- POS-tagged (rule-based tagger, upgraded)
- metadata still complex (same taxonomy)
- consortium: publisher, LT company, **university**, technical institute
  - freely available online, authentication
- purpose: lexicography, education, LT

Aims

- From lexicographers to students
  - analysis, survey
  - booklet
  - simple query window
  - avoid flood of intimidating information
- BUT!
  - limited by the existing technology
  - DEMO

First glimpse of WS – 2002

- Adam Kilgarriff and Michael Rundell
- *Lexical Profiling Software and its Lexicographic Applications – a Case Study*
- Euralex 2002, Copenhagen

- Just for English
- Just a combination of collocation patterns

PP_about: ratio = 12.07 : 1, counts = 86

PP_with 665		**PP_between** 135		**PP_at** 61		**object_of** 1637		**before_prepn** 2651	
18.65 :1		13.2 :1		1.74 :1		1.69 :1		1.33 :1	
friend	19	princess	8	table	4	overhear	33	topic	93
stranger	5	Charles	4	dinner	3	steer	25	snatch	16
passenger	5	woman	6	time	7	record	46	lull	14
people	22	people	8	party	4	tape	11	deep	34
man	17	officer	3	moment	3	tap	16	listen	57
artist	5	child	4	school	3	resume	14	eavesdrop	9
pupil	5					hold	63	engage	30
girl	7					interrupt	14	hum	11
woman	11					continue	42	buzz	10
mother	7					prolong	8	babble	7
visitor	4					make	120	transcript	11
colleague	4					hear	38	deep	26
speaker	4					conduct	13	gist	6
lady	4					finish	19	recording	17
wife	5					end	20	murmur	7
minister	6					start	34	participant	13
husband	4					remember	27	fragment	11
queen	3					recall	12	indulge	7
adult	3					dominate	10	extract	8
doctor	4					keep	34	course	31

The real deal – 2004

- Adam Kilgarriff, Pavel Rychly, Pavel Smrz, David Tugwell
- *The Sketch Engine*
- Euralex 2004, Lorient

- Sketch Grammar (cross-lingual)
- Sketch Difference, Thesaurus

Slovene Word Sketches (2006)

- Simon Krek and Kilgarriff
- *Slovene Word Sketches*
- Proceedings of the 5th Slovenian/First Intnational Lanugage Technology Conference, Ljubljana, Slovenia

- First version in 2005

čas	glava

coord	7334	0.5
prostor	837	44.63
datum	144	37.89
trud	75	35.92
kraj	247	34.76
Vera	33	31.35
energija	155	30.85
denar	206	27.91
se	192	26.5
bit	35	25.96
zaslonka	13	25.71
potrpljenje	21	24.85
trimesečen	17	23.36
da	97	23.18
on	131	21.66
svoj	26	20.95
ne	63	19.83
napor	25	19.0
tudi	43	18.33
kraja	24	18.02
njegov	35	17.97

is_obj4_of	2506	3.9
skloniti	98	54.21
beliti	83	49.81
dvigniti	218	49.15
razbijati	71	44.74
odsekati	60	43.2
pomoliti	45	42.62
nagniti	59	40.97
tiščati	46	40.92
stakniti	39	39.75
obrniti	90	34.92
odrezati	44	33.54
nasloniti	29	32.21
sploščiti	20	32.18

a_modifier	19068	1.4
sklonjen	157	59.54
obrit	66	47.71
kronan	35	39.24
dvignjen	62	36.12
zeljnat	23	35.75
odsekan	33	35.5
Hermanov	34	35.11
razgret	40	34.63
mrtvaški	34	31.12
trezen	41	30.65
bikov	22	29.33
ovnov	16	29.12
video	106	29.03
koničast	34	28.96
pobrit	13	27.94
zeljen	18	27.92
bister	35	27.26

prec_po	1090	9.0
rojiti	91	65.67
udariti	142	55.22
motati	49	51.25
popraskati	24	42.95
treščiti	38	39.83
poditi	32	38.99
tolči	34	37.54
lopniti	12	32.97
tepsti	20	32.27
praskati	13	28.88
bloditi	12	28.06
plesti	14	27.2
trepljati	9	26.25
poškodovati	23	25.79
čohati	6	25.61
pobriti	5	24.37
srati	8	23.96

Sketch Difference

cona = zone

območje = area

green = cona

red = območje

a_modifier	4941	36173	2.6	1.9
erogen	75	13	58.6	21.4
obrten	288	9	58.3	5.6
ekonomski	411	19	54.5	5.2
carinski	271	214	53.4	33.9
industrijski	284	76	52.2	19.4
prost	317	79	50.1	16.7
prostocarinski	61	21	47.3	21.8
obmejen	7	313	11.5	46.7
moder	139	20	41.4	8.2
okupacijski	35	10	38.6	14.6
nekdanji	41	710	16.6	37.3
tamponski	17	12	35.8	23.1
demilitariziran	9	30	25.3	33.8
zaprt	5	162	7.6	32.4
brezcarinski	26	17	30.3	16.5
konvergenčen	12	6	29.7	15.0
operativen	34	47	28.5	21.8
bivši	15	199	11.9	27.4
posamezen	9	338	5.1	26.8
svoboden	42	43	25.5	14.9
turističen	16	183	11.0	23.6
visok	8	353	3.1	23.5
koprski	12	110	11.8	23.2
mesten	11	190	7.2	21.7
siv	23	14	20.9	8.1

19 spor 0.273 spopad 0.267 vojna 0.266 nasilje 0.202 boj 0.17

03 težava 0.288 razmera 0.279 situacija 0.264 dogajanje 0.228 dogode problematika 0.18 potreba 0.172 stvar 0.172

25 katastrofa 0.243 padec 0.198 zlom 0.185 razpad 0.179 izbruh 0.17

24 posledica 0.21 pomanjkanje 0.209 nevarnost 0.188 pritisk 0.182 vp

29 politika 0.196 proces 0.189 razvoj 0.184 sila 0.182 gibanje 0.182 o

23 recesija 0.211 revščina 0.178

21

07 nesreča 0.196 bolečina 0.169

05 reforma 0.203 napad 0.196 volitve 0.178 poseg 0.176 akcija 0.171

93 napetost 0.19

87

82 poraz 0.175

Thesaurus

kriza = crisis

Gigafida (2012)

- online (broadband)
- 1,2 billion words
- POS-tagged (statistical tagger, better!)
- metadata simplified
- consortium: ~~publisher,~~ LT company, university, technical institute, language institute
  - freely available online, ~~authentication~~
- purpose: education, lexicography / LT

The split

- Communication in Slovene project
  - 2008-2013
- Gigafida corpus
  - Slovene Lexical Database (dictionary)
    - Sketch Engine
    - for lexicographers
  - online concordancer(s)
    - for education and general public

Web concordancers

- Log analysis of FidaPLUS concordancer
- FidaPLUS web survey
- Analysis of existing corpus tools
- Analysis of popular web tools (Google etc.)

Survey – findings

- Simple search – regularly used by 72% users
- Advanced search – rarely used (only 8% use it regularly)
- Lack of intuitiveness
- the manual is almost key to learning how to use a corpus tool
  - "…if you are not using the interface for a while, you forget what the search commands are, and you don't (want to) bother with looking into the manual"
  - "…the interface should have **a modern design, it** should be more **user-friendly, and its use should be clear and transparent"**

Main design principles

- similarity to the well-known non-linguistic tools (e.g. Google)
- No registration
- Minimum navigation
- No redundant functions (less is more)
- Simplicity of searches
- Help and tips in pop-up windows
- Simple descriptions of functionality (no terminology)

Concordancers

- For lexicographers
  - NoSketchEngine
    - http://nl.ijs.si/noske/index-en.html
  - Sketch Engine
    - http://ske.amebis.si/
- For education / general use
  - Gigafida
    - http://www.gigafida.net/
  - Kres
    - http://www.korpus-kres.net/
- DEMO