GDEX FOR SLOVENE

Iztok Kosem

Trojina, Institute for Applied Slovene Studies &
Faculty of Arts, University of Ljubljana

WG3 Worskhop, Vienna, 12 February 2015

GDEX for Slovene

- Communication in Slovene project
  - 2008-2013
  - 3,2 million euro
  - **http://www.slovenscina.eu**
- Slovene Lexical Database (Krek & Gantar 2012)
- Corpora:
  - 620-million word FidaPLUS corpus (v1)
  - 1.2-billion word corpus of Slovene (Gigafida) (v2)

aktiven *(pridevnik)* **Fida PLUS 620m (SLD sketch grammar) freq = 43634** (59.1 per million)

osebek+biti	701	14.8
sekcija	21	31.1
društvo	25	22.09
član	20	19.6

>>

kakšen-g?	306	12.6
postati	175	38.75
ostati	73	30.28

>>

v_tožil-p	532	7.4
čas	75	29.91
vrsta	27	23.03
kot	19	18.9
dan	24	16.93
leto	36	16.64

>>

kdo-kaj?	32726	7.0
matrika	739	87.89
prebivalstvo	1063	63.52
preživljanje	358	62.93
oglje	190	57.45
politika	1819	56.96
učinkovina	203	53.65
snov	750	52.34
počitnice	518	51.51
oddih	180	49.66
član	1142	47.38
vloga	945	46.56
igranje	280	46.07
sodelovanje	809	44.54
sestavina	249	43.41
zglavnik	42	43.31
vzglavnik	63	42.84
oprema	618	42.83

v_rodil-p	688	6.4
leto	125	29.17
čas	30	17.9

>>

kako-kdaj?	12326	6.2
delovno	801	91.23
spolno	210	61.76
telesno	226	60.78
zelo	1781	59.29
fizično	157	54.16
biološko	97	53.11
najbolj	979	52.25
športno	173	51.84
politično	205	51.18
bolj	926	49.02
površinsko	44	48.96
trenutno	276	48.64
vedno	621	44.64
vsestransko	54	41.52
malo	458	39.89
potresno	27	38.42
izredno	140	38.36

Tickbox Lexicography - Select Examples

Lemma: aktiven

Gramrel: kdo-kaj?

Template: fidaplus_slovene Alternative GDEX configuration: None

prebivalstvo

- [ ] V večini držav v razvoju naj bi siva ekonomija vključevala med 30 - 70 odstotkov aktivnega prebivalstva.
- [ ] Leta 2001 naj bi bilo v EU brezposelnih 8 odstotkov aktivnega prebivalstva oziroma 15 milijonov oseb.
- [ ] Zakon določa, da je lahko le pet odstotkov aktivnega prebivalstva tujcev, torej približno 41.000 ljudi.
- [x] Brez službe je okoli 388.000 ljudi ali več kakor 22 odstotkov aktivnega prebivalstva.
- [ ] Do leta 2010 naj bi imelo 85 odstotkov aktivnega prebivalstva doseženo srednješolsko izobrazbo.
- [x] Brezposelnost na širšem območju Maribora je pred dobrima dvema letoma zajela že skoraj četrtino aktivnega prebivalstva.
- [ ] Rast zaposlenosti v ZDA je letos že presegla naravno povečanje aktivnega prebivalstva za skoraj pol odstotne točke.
- [x] Februarja 2001 je tako v Sloveniji internet uporabljalo okoli 19 % aktivnega prebivalstva.
- [ ] V črnomaljski in semiški občini je zaposlenih 5.634 ljudi ali 80,5 odst aktivnega prebivalstva.
- [ ] Tedaj bo razmerje med aktivnim prebivalstvom in upokojenci 10 proti 7.

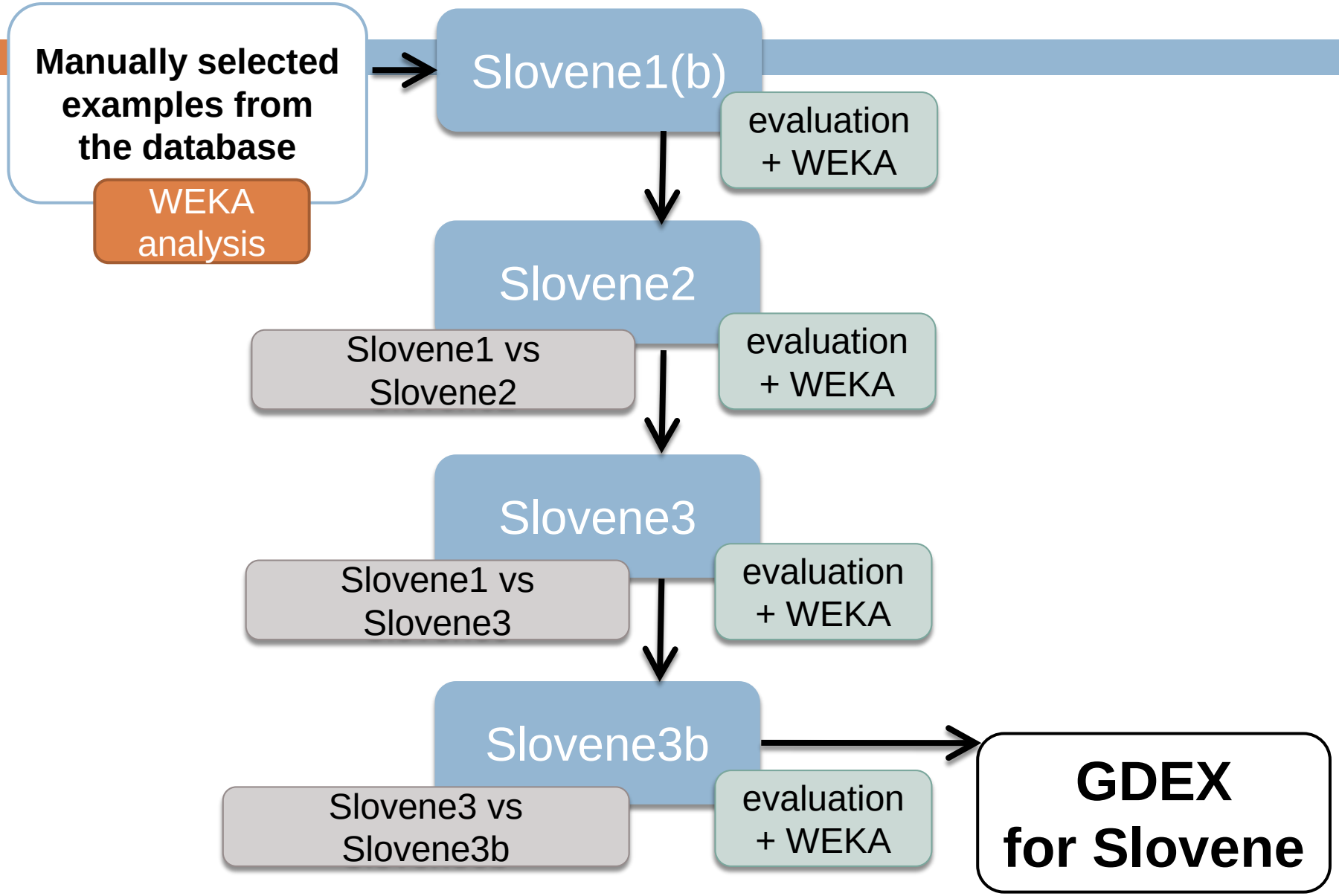
oglje

- [ ] Po zaužitju do treh jagod mora oseba vzeti aktivno oglje in nadomeščati z drisko izgubljeno tekočino.

GDEX for Slovene v1

- GDEX for Slovene (Kosem, Husák and McCarthy, 2011)
- Initial GDEX configuration:
  - Non-language specific classifiers of English GDEX
  - analysis of manually selected examples in the database (using WEKA tool)
- Evaluation in TBL:
  - Comparing different GDEX configurations
  - Logging good (selected) and "bad" (unselected) examples
- Improving GDEX for Slovene based on:
  - Recorded observations
  - Analysis of good (and bad) examples
- Result: GDEX configuration Slovene3b

GDEX for Slovene – version 1



GDEX: Slovene3

prebivalstvo

- V večini držav v razvoju naj bi siva ekonomija vključevala med 30 - 70 odstotkov **aktivnega** prebivalstva.
- Leta 2001 naj bi bilo v EU brezposelnih 8 odstotkov **aktivnega** prebivalstva oziroma 15 milijonov oseb.
- Zakon določa, da je lahko le pet odstotkov **aktivnega** prebivalstva tujcev, torej približno 41.000 ljudi.
- Brez službe je okoli 388.000 ljudi ali več kakor 22 odstotkov **aktivnega** prebivalstva.
- Do leta 2010 naj bi imelo 85 odstotkov **aktivnega** prebivalstva doseženo srednješolsko izobrazbo.
- Brezposelnost na širšem območju Maribora je pred dobrima dvema letoma zajela že skoraj četrtino **aktivnega** prebivalstva.
- Rast zaposlenosti v ZDA je letos že presegla naravno povečanje **aktivnega** prebivalstva za skoraj pol odstotne točke.
- Februarja 2001 je tako v Sloveniji internet uporabljalo okoli 19 % **aktivnega** prebivalstva.
- V črnomaljski in semiški občini je zaposlenih 5.634 ljudi ali 80,5 odst **aktivnega** prebivalstva.
- Tedaj bo razmerje med **aktivnim** prebivalstvom in upokojenci 10 proti 7.

GDEX: Slovene2

prebivalstvo

- Zaradi bolečin v križu so največje težave pri **aktivnem** prebivalstvu.
- Tedaj bo razmerje med **aktivnim** prebivalstvom in upokojenci 10 proti 7.
- Zakaj je **aktivno** prebivalstvo na udaru?
- To pa je okrog 60 odstotkov **aktivnega** prebivalstva.
- Brez službe je okoli 388.000 ljudi ali več kakor 22 odstotkov **aktivnega** prebivalstva.
- Delovno aktivni in brezposelni sestavljajo skupaj **aktivno** prebivalstvo.
- V večini držav v razvoju naj bi siva ekonomija vključevala med 30 - 70 odstotkov **aktivnega** prebivalstva.
- Delež aktivnih žensk v skupnem številu **aktivnega** prebivalstva je 47 odstotkov.
- Do leta 2010 naj bi imelo 85 odstotkov **aktivnega** prebivalstva doseženo srednješolsko izobrazbo.
- Najbolj ogroža **aktivno** prebivalstvo, predvsem ljudi, stare od 35 do 45 let.

Findings

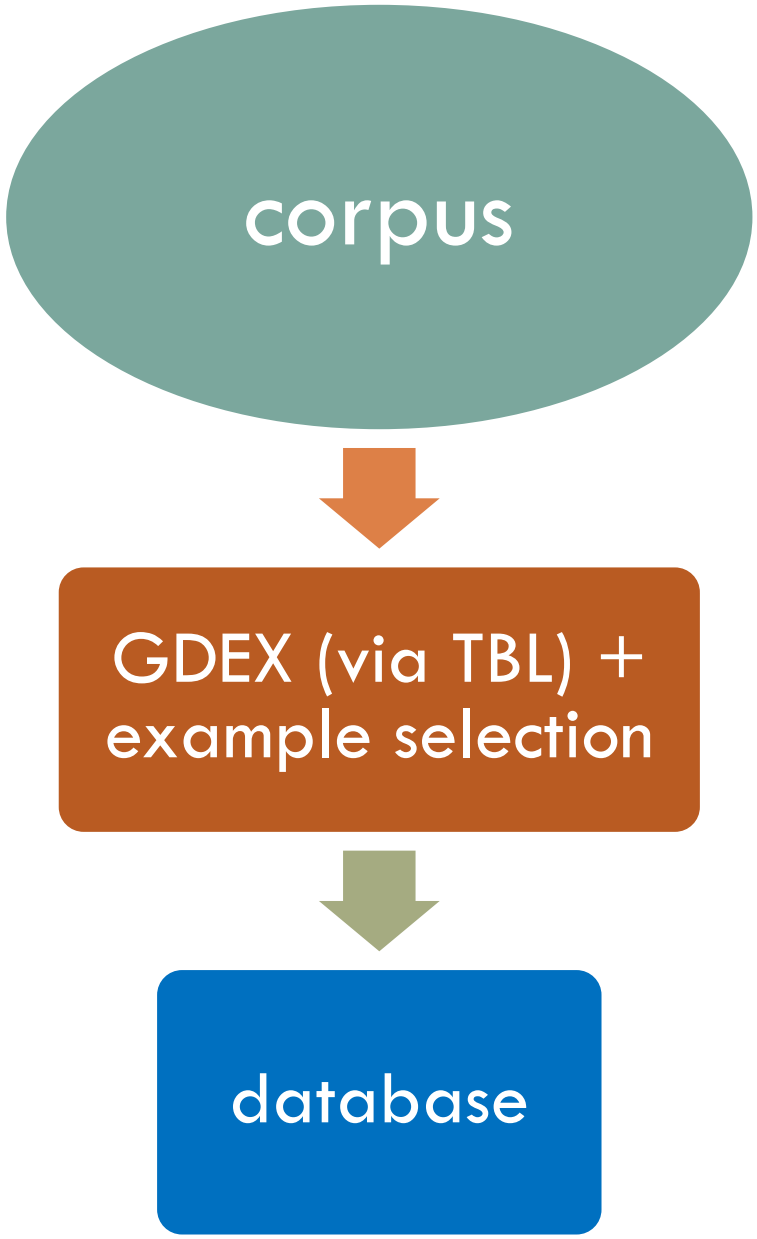
- Sentence length
  - from 8-30 to 15-35 → considerable improvement
- Keyword position
  - English – beginning of the sentence (0-20%)
  - Slovene – middle to end of the sentence (40-100%)
- Penalizing repetitions of the word in the same example
- Sentence length (max 60)
- Word length (>18 characters)

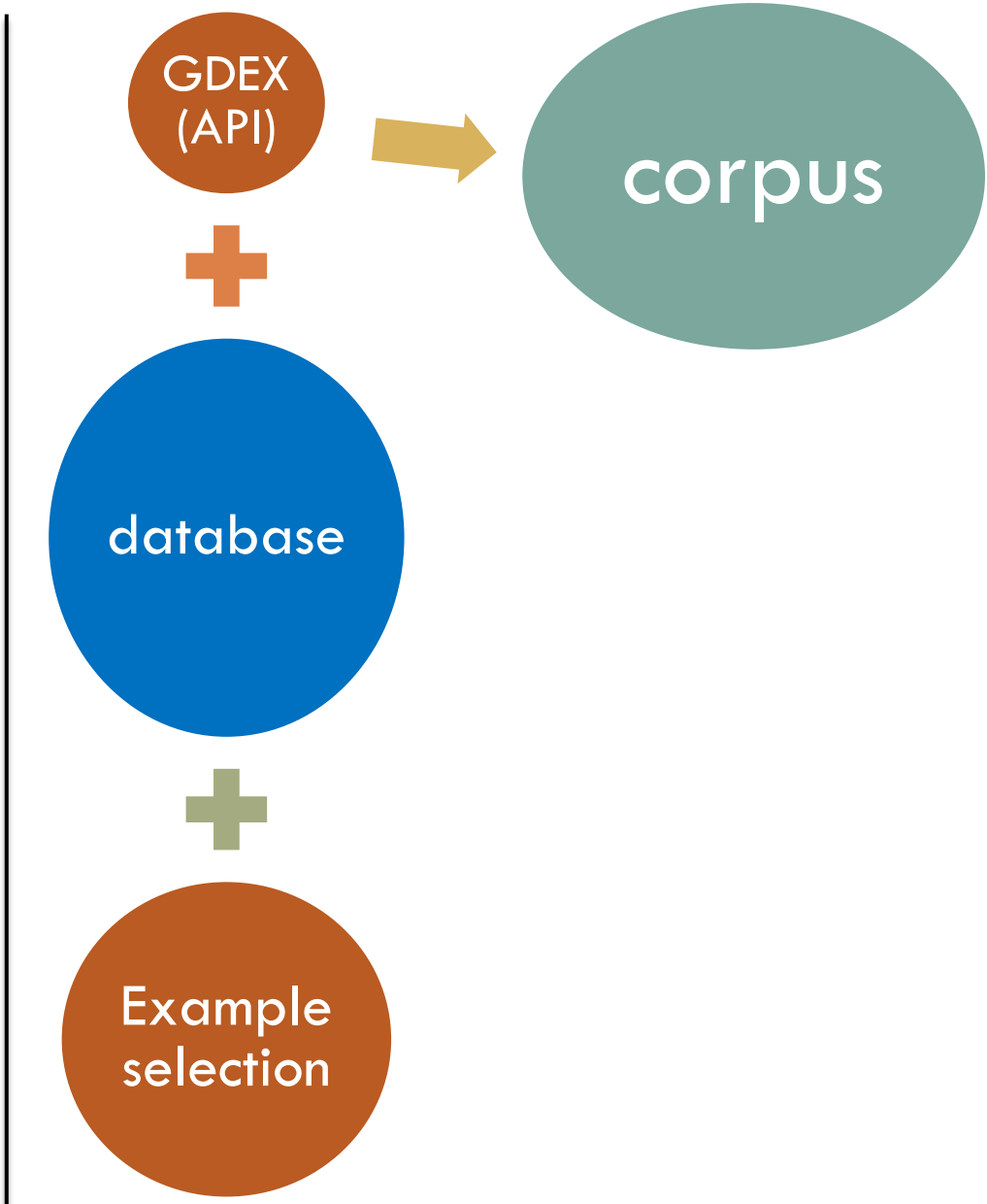
GDEX for Slovene – from v1 to v2

- Automatic extraction:
  point of departure → GDEX for Slovene v1
-
-

GDEX for Slovene – from v1 to v2

- Automatic extraction:
  point of departure → GDEX for Slovene v1
- Aim: separate GDEX configurations for nouns, verbs, adjectives, adverbs
- Different task: **first 3 examples of each collocate need to be good** (not any 3 out of 10 examples)


corpus
GDEX (via TBL) + example selection
database


GDEX (API)
corpus
database
Example selection

Classifiers – no change

- Boolean classifier group (binary) (weight = 100)
  - Whole sentence
  - Classifier matching regexp ([<|\][>/\\])
  - Any token frequency < 3
- "Penalty" classifiers
  - Proper nouns (weight = 2): -0.2 deduction for each proper noun
- Example diversity: Levenshtein distance > 30%

Fine-tuning of classifiers

- Removed classifiers:
  - Boolean: maximum token length
  - Percentage of tokens with frequency above 104
- Classifiers moved under boolean:
  - classifier penalizing web addresses, emails
  - keyword repetition (matching lemma, not token)
- Changed classifiers:
  - Token length (originally 6 – from English GDEX → 8)
  - maximum sentence length = 60 → 35-40 tokens
- Changed weights:
  - Sentence length (2 → 10)
  - Capital letters (2 → 4)
  - Symbols (1 → 5)
  - Punctuation (1 → 5)

New classifiers

- Blacklist of sentence-initial words:
  - *sledi, zatorej, torej, nato, vendar, gre, oboji, dotlej, zato, tovrsten, to, ta, slednji, tak, takšen, potekati*
  - *both, it follows, thus, therefore, then, but, this is, till then, because, this type of, this, that, latter, it takes place*
- Blacklist of sentence-initial phrases
- Penalty for lemmas with frequency below 600 or 1000
- Separate classifier for commas (penalty for multi-clause sentences)
- **Third-collocate classifier!** (e.g. *take a **long** walk*)

Summary

- Slovenian experience:
  - Good results
  - Particularly good at **helping** to identify good **database** examples
  - More useful when used at collocational (under gramrels) than at lemma level
- GDEX already used in various projects
  - Lexicographic (Slovene lexical database)
  - Terminological (TERMIS)
  - Pedagogical (Pedagogic corpus-based grammar)