

GDEX FOR SLOVENE



Iztok Kosem, Milos Husak, Diana McCarthy

Outline

- (Good) examples in dictionaries
- GDEX tool
- GDEX for English
- GDEX for Slovene
 - ▣ Design
 - ▣ Evaluation
 - ▣ Findings
- Future plans

(Good) examples in dictionaries

- Examples in dictionaries
 - ▣ Illustrating usage - putting the words back in context
 - ▣ supporting definitions
 - ▣ Helping with navigation through longer entries
- Characteristics of good examples
 - ▣ Naturalness
 - ▣ Typicality
 - ▣ Informativeness
 - ▣ Intelligibility
- time-consuming task of searching good examples
- Potential for automatisisation?

časopis	mogoče že danes na tradicionalni zabavi po borzni	konferenci	, bodo borzniki lahko nazdravili tudi prvemu rojstnemu
revija	z doseženimi prodajnimi cenami, « je na novinarski	konferenci	pred dnevi poudaril Lenič. V veliki večini primerov
časopis	Ampak drugače enostavno ne gre, « je na novinarski	konferenci	pojasnjeval Brian Earley. V New Yorku se sicer ni
časopis	je včeraj prispel v Ljubljano, kjer bo sodeloval na	konferenci	IPI. (sta) Danes Pa kar brez hvale Vprašanje, kdaj
časopis	pobudo Rusije, da bi čimprej pripravili mednarodno	konferenco	o človekovih pravicah in mejah na Balkanu. Trajkovski
revija	načrti za prihodnost. To so dokazali tudi na tiskovni	konferenci	, sklicani v petek v večernih urah. Prisotni so bili
časopis	krst, « je povedal na četrtkovi slovenski novinarski	konferenci	v olimpijski vasi. O. G. Kdo bi se utegnil vmešati
revija	izbral kraljevsko držo in daje večer pred tiskovnimi	konferencami	razdeliti novinarjem vprašalnike. Na sami konferenci
revija	dva dni potekalo srečanje predstavnikov škofovskih	konferenc	JV Evrope. Tema srečanja cerkvenih dostojanstvenikov
časopis	optimalno hitrost širitve, zato je mora biti medvladna	konferenca	o reformi ustanov zaključena do konca letošnjega
neoznačeno	načela še niso upoštevala. O etnični meji smo na naši	konferenci	že govorili tudi s kritičnimi naglasi. Italijanska
časopis	indijski račun odpovedala sodelovanje. Novega roka za	konferenco	še niso določili. Ameriška podjetja so presenečena
revija	zelo pogosta oblika prireditve v podjetjih tiskovne	konferenca	. Seveda pa se pri Jezerškovih radi lotijo tudi priprave
časopis	prisluškoval nek drug državni organ. Na včerajšnji tiskovni	konferenci	pa je šel minister za gospodarstvo še dlje. Policijo
časopis	obilico padavin. Težko pričakovana uvodna tiskovna	konferenca	Michaela Schumacherja je že postavila tudi prvo uganko
časopis	je bila pod okriljem Združenih narodov organizirana	konferenca	o prepovedani trgovini z osebnim orožjem in lahko
časopis	sprejem Turčije. Turški predsednik pa je na tiskovni	konferenci	omenil, da sta z gostom razpravljala tudi o obrambnem
časopis	zakonodajno proceduro vloženi in na včerajšnji novinarski	konferenci	predstavljeni predlog zakona o zdravstvenem varstvu
časopis	Angelesu. Oracle Open World je velika, pravzaprav ogromna	konferenca	. Menda se je je letos udeležilo skoraj 20 tisoč ljudi
časopis	tem vprašanjem. Plaz pa je bil sprožen na tiskovni	konferenci	ob podpisu dodatka h kolektivni pogodbi za zdravnike

časopis izrazilo tudi vodstvo Nove Slovenije, ki je novinarsko **konferenco** sklicalo v zgodnjih popoldanskih urah. Predsednik

časopis pregona so uprizorili pravi spektakel. Tudi s tiskovno **konferenco** po jutranjih aretacijah. Čeprav mi ni čisto jasno

časopis prodaji krvni. V začetku novembra je bila v Pekingu prva **konferenca**, posvečena težavam, okuženim z virusom HIV in obolelim

časopis tem okviru v Ljubljani pripravil krajšo otvoritveno **konferenco**. Otvoritvenih dejavnosti se po besedah Milana Koritnika

časopis vse odprto. Vse je še mogoče - tudi polom mirovne **konference**. Od naših dopisnikov Ozračje v starodavnem rambouilletskem

časopis Janez Drnovšek in Viktor Orban. Kot je na tiskovni **konferenci** povedal slovenski premier, sta veliko pozornost namenila

časopis begal po štadionu, ker ni vedel, kje sploh je tiskovna **konferenca**. Ko je s pomočjo novinarskega kolega iz Prekmurja

revija Vodovodno 20 priromala uro pred začetkom tiskovne **konference**, torej je bil z njo Gavez vsekakor seznanjen. Če

knjiga S Hanssonom sta imela tudi improvizirano tiskovno **konferenco**. Sestanka se je udeležil tudi Ekholm. Še vedno je

časopis vse to izboljšala. Te dni ste se vrnilo z ministrske **konference** o bolonjskem procesu v Bergnu. Ko primerjate slovenski

časopis bodi dobra zaščita! Direktorju Tonetu Vogrincu je TV **konferenca** (TV Slovenija) prav prišla. Nabralo se je toliko

časopis o trendih na tem področju so govorili na tiskovni **konferenci** podjetja Nepremičnine Celje. Piše: Darja Veršec Po

časopis ugnala Sacramento, drugo najuspešnejše moštvo zahodne **konference**. Vidno vlogo pri uspehu moštva iz Minneapolisa je

časopis o tem bo mogoče zvedeti že na skorajšnji tiskovni **konferenci** finančnega ministra Dušana Mramorja v začetku prihodnjega

časopis Danes se začenja v Strasbourgu tridnevna evropska **konferenca** o položaju državnih tožilcev v 21. stoletju. Med

časopis SLS V občini Zavrč so se minulo soboto na prvi letni **konferenci** srečali člani in simpatizerji Slovenske ljudske stranke

časopis dni ne dobivamo le vabil na sto in eno novinarsko **konferenco** (tako strank kot posameznih kandidatov) na dan, temveč

časopis Economist Conferences organizira na leto 16 tovrstnih **konferenc** v srednji in vzhodni Evropi in približno 60 po vsem

časopis brigado, ki bo letos maja v Savinjskem gaju. Uradni del **konference** so zaključili s podelitvijo priznanj dolgoletnim

časopis - Na vabilu Konjeniške zveze Slovenije na tiskovno **konferenco** je bilo zapisano, da bodo na njej " ... podrobneje

- [časopis](#) s tem da bo v nedeljo organizirana tudi novinarska **konferenca** , na kateri bodo govorili: predsednik evropske Fibe
- [časopis](#) kar sami dočakali predstavnike medijev na novinarski **konferenci** . Odgovorni ne morejo priti, niti ko želijo. Včeraj
- [časopis](#) možnosti, da osvoji eno izmed medalj, " je na novinarski **konferenci** v Hamburgu menil nemški trener Dirk Bauermann. Nemci
- [časopis](#) kar sami dočakali predstavnike medijev na novinarski **konferenci** . Odgovorni ne morejo priti, niti ko želijo. Včeraj
- [časopis](#) informacij, ki jih je gorenjski klub na novinarski **konferenci** v Kranjski Gori predstavil javnosti, hkrati pa zanikal
- [časopis](#) Slavko Ažman in Pavle Rupar na včerajšnji novinarski **konferenci** v Kranjski Gori foto: Roman Šipič šNa Gorenjskem
- [časopis](#) šov. Najprej je dolgo zavračal prihod na novinarsko **konferenco** , med zbrane predstavnike sedme sile je pošiljal pomočnika
- [časopis](#) polno sobo in še pred uradnim začetkom novinarske **konference** ostro napadel poročevalce. Kot smo lahko razbrali
- [revija](#) ves strokovni potencial v Sloveniji. Prva delovna **konferenca** snovalcev strategije je bila na začetku decembra
- [revija](#) . Kot so menili tudi razpravljavci na prvi delovni **konferenci** , je zunanjeekonomske strategijo Slovenije težko ločevati
- [revija](#) dejavnost, promocijske aktivnosti (predvsem poslovne **konference** , sejmi doma in v tujini), informativni dnevi, svetovanje
- [časopis](#) prireditev, ki jih spremlja na televiziji ali tiskovni **konferenci** . Prej bi rekli, da ni sposoben konfrontacije z mediji
- [časopis](#) razne statute in pravne knjige. Solisti so na tiskovni **konferenci** stališča dokumentirali s pričevanji in pisnimi dokumenti
- [časopis](#) direktorja, vendar ne šest s strani udeležencev tiskovne **konference** . Ko direktor govori o » osnovnih vodilih, ki jih
- [časopis](#) listine «, kot pravi, ve samo on. Mi smo na tiskovni **konferenci** dobili fotokopijo dokumenta zdravniške komisije II.
- [časopis](#) ljudi tudi v revnejših območjih, opozarjajo udeleženci **konference** o debelosti, ki poteka te dni v Washingtonu. Socialno
- [časopis](#) Razburkana politična scena je tokrat v času pred letno **konferenco** postregla z izjavo prvaka politične stranke Lojzeta
- [časopis](#) primestne četrti Vičava - Orešje so te dni na tiskovni **konferenci** predstavili prva dva izmed sedmih programov ptujskega
- [časopis](#) Razburkana politična scena je tokrat v času pred letno **konferenco** postregla z izjavo prvaka politične stranke Lojzeta
- [časopis](#) zavarovalnih storitev. Kot je na včerajšnji tiskovni **konferenci** povedal generalni direktor SKB Generali dr. Christoph

GDEX – Good Dictionary EXamples

- Ranking tool
- Ranks examples according to their potential for dictionary purposes
- Automatic ranking, based on various syntactic and lexical features
- Which features can be ranked?

GDEX for English - heuristics

- Heuristics include classifiers which use:
 - punctuation
 - sentence length
 - word length
 - word frequency
 - keyword position
 - proper nouns, polysemy and pronouns
 - ...

GDEX for English - procedure

- Procedure:
 - ▣ Score on each classifier
 - ▣ Weight scores
 - ▣ Weighted average of classifier scores
 - ▣ Rank sentences
- First used in lexicography for Macmillan English Dictionary online (examples for collocations)
- Based on manually annotated good sentences in a set of concordances
- Classifiers useful for other languages?

GDEX for Slovene

- Communication in Slovene project
 - 2008-2013
 - 3,2 million euro
 - <http://www.slovenscina.eu>
- Lexical database for Slovene
- Sketch Engine
- Selecting examples in Word Sketch (using Tickbox Lexicography)

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
- preživetje [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:

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 četrtno **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.

ogljje

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

GDEX for Slovene – design

- Aim: configuration that offers 3 good examples out of 10 for each collocate
 - ▣ Lexicographers usually select from top 10-15
- Using non-language specific classifiers of English GDEX as a point of departure
- Using existing manually selected examples in the lexical database as a benchmark
- Using the WEKA tool to determine the values of classifiers
 - ▣ shows various statistics for each feature
 - ▣ 2D charts for determining efficiency of two measurements

Weka Explorer



- Preprocess
- Classify
- Cluster
- Associate
- Select attributes
- Visualize

- Open file...
- Open UR...
- Open DB...
- Generat...
- Undo
- Edit...
- Save...

Filter

Choose Apply

Current relation
 Relation: Logged Sentences from TBLex-weka.filte...
 Instances: 18039 Attributes: 10

Attributes

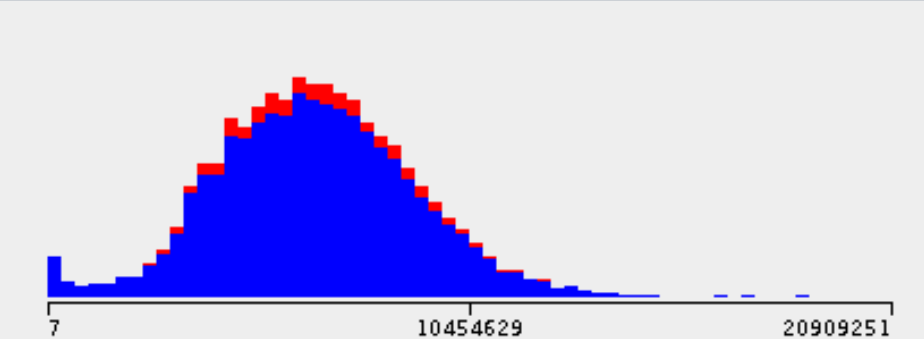
No.		Name
1	<input checked="" type="checkbox"/>	average freq
2	<input type="checkbox"/>	average_wordlen
3	<input type="checkbox"/>	keyword position
4	<input type="checkbox"/>	max freq
5	<input type="checkbox"/>	max_wordlen
6	<input type="checkbox"/>	min freq
7	<input type="checkbox"/>	min_wordlen
8	<input type="checkbox"/>	sentence length
9	<input type="checkbox"/>	whole sentence
10	<input type="checkbox"/>	good

Selected attribute

Name: average freq Type: Numeric
 Missi... 0 (0%) Distinct: 16859 Unique: 15905 (88%)

Statistic	Value
Minimum	7
Maximum	20909251
Mean	6658934.558
StdDev	2521620.151

Class: good (Nom)



Status
OK

x 0

GDEX for Slovene – design

- Duplicate or similar examples?
 - ▣ Levenshtein distance at least 30%
- Several configurations
- Differences in classifier values and weights
- Evaluation:
 - ▣ Word sketches
 - ▣ minimum frequency of a collocate = 15
 - ▣ Lemmas: nouns, verbs, adjectives, adverbs
 - ▣ Logging selected (and not selected) examples
- TickBox Lexicography feature that allowed the comparison of the results of two different configurations

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 četrtno **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.

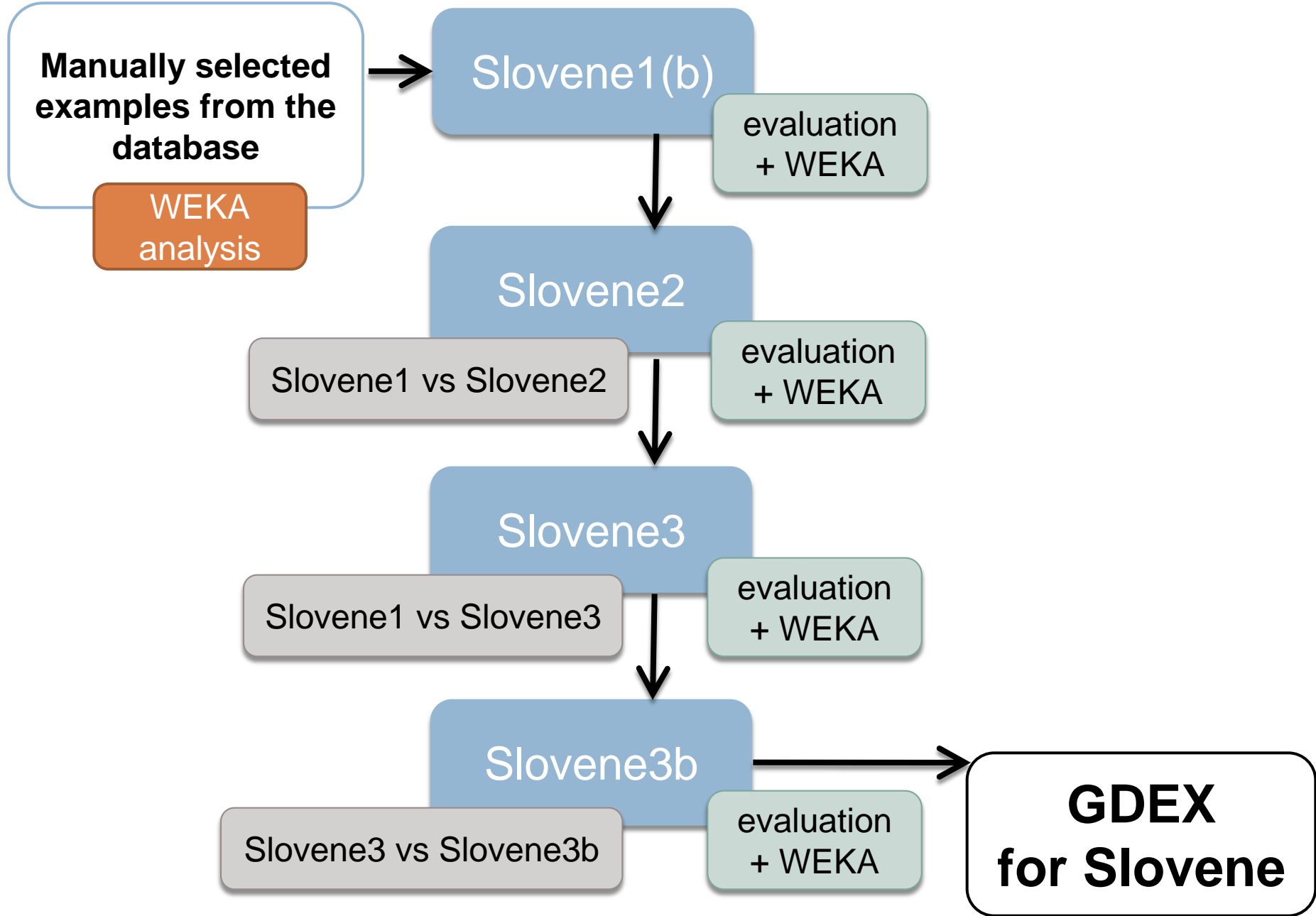
GDEX: Slovene2

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.

To modify or not to modify

- Authentic corpus examples vs modified corpus examples
- Lexical database example vs dictionary example
- Modified lexical database example \Rightarrow good dictionary example



GDEX: Slovene 1

bolečina

Bukovčeva **muči** bolečina na zadnjem delu stegna na zamašni nogi.

Ločanko **mučijo** bolečine v hrbtu, tako da ne more trenirati.

Vas **mučijo** bolečine v nogah ali hrbtenici?

Vrhničanko spet **mučijo** bolečine pod kolenom, Šentjernejšanko pa boli gleženj.

Občasno vas **mučijo** bolečine v hrbtu in kronične težave s kožo.

Martina Hvastija **mučijo** bolečine v želodcu, Andreja Hauptmana rahla viroza.

Vas **mučijo** bolečine v križu?

Pogosto jih **mučijo** revma in bolečine v sklepih.

Vas **mučijo** bolečine v hrbtenici, čeprav redno delate vaje za hrbet?

Včasih jih **mučijo** tudi krčevite bolečine in kislo izpahovanje iz želodca, po driskah so zaprti in spet obratno.

Vas **mučijo** bolečine v nogah ali hrbtenici?

Ločanko **mučijo** bolečine v hrbtu, tako da ne more trenirati.

Občasno vas **mučijo** bolečine v hrbtu in kronične težave s kožo.

Vas **mučijo** bolečine v hrbtenici, čeprav redno delate vaje za hrbet?

Bukovčeva **muči** bolečina na zadnjem delu stegna na zamašni nogi.

Pogosto jih **mučijo** revma in bolečine v sklepih.

Tedaj so me **mučile** močne bolečine, zato sem šla v bolnišnico.

Zadnje čase jo namreč **mučijo** bolečine v kolkih in težko stoji.

Vrhničanko spet **mučijo** bolečine pod kolenom, Šentjernejšanko pa boli gleženj.

Vas **mučijo** bolečine v križu?

GDEX: Slovene 1b

GDEX: Slovene2

bolečina

Zadnje čase jo namreč **mučijo** bolečine v kolkih in težko stoji.

Vas **mučijo** bolečine v nogah ali hrbtenici?

Ločanko **mučijo** bolečine v hrbtu, tako da ne more trenirati.

Občasno vas **mučijo** bolečine v hrbtu in kronične težave s kožo.

Pogosto jih **mučijo** revma in bolečine v sklepih.

To ga je bolj **mučilo** kot pa ostra bolečina v prsih.

V hrbtenici so ga **mučile** kronične bolečine.

Vas **mučijo** bolečine v hrbtenici, čeprav redno delate vaje za hrbet?

Ste zdravi in redno telovadite, a vas vseeno **muči** bolečina?

Tedaj so me **mučile** močne bolečine, zato sem šla v bolnišnico.

Bukovčevo **muči** bolečina na zadnjem delu stegna na zamašni nogi.

Ločanko **mučijo** bolečine v hrbtu, tako da ne more trenirati.

Vas **mučijo** bolečine v nogah ali hrbtenici?

Vrhničanko spet **mučijo** bolečine pod kolenom, Šentjernejčanko pa boli gleženj.

Občasno vas **mučijo** bolečine v hrbtu in kronične težave s kožo.

Martina Hvastija **mučijo** bolečine v želodcu, Andreja Hauptmana rahla viroza.

Vas **mučijo** bolečine v križu?

Pogosto jih **mučijo** revma in bolečine v sklepih.

Vas **mučijo** bolečine v hrbtenici, čeprav redno delate vaje za hrbet?

Včasih jih **mučijo** tudi krčevite bolečine in kisló izpahovanje iz želodca, po driskah so zaprti in spet obratno.

GDEX: Slovene1

GDEX: Slovene1

bolečina

Bukovičeva **muči** bolečina na zadnjem delu stegna na zamašni nogi.

Ločanko **mučijo** bolečine v hrbtu, tako da ne more trenirati.

Vas **mučijo** bolečine v nogah ali hrbtenici?

Vrhničanko spet **mučijo** bolečine pod kolenom, Šentjernejčanko pa boli gleženj.

Občasno vas **mučijo** bolečine v hrbtu in kronične težave s kožo.

Martina Hvastija **mučijo** bolečine v želodcu, Andreja Hauptmana rahla viroza.

Vas **mučijo** bolečine v križu?

Pogosto jih **mučijo** revma in bolečine v sklepkih.

Vas **mučijo** bolečine v hrbtenici, čeprav redno delate vaje za hrbet?

Včasih jih **mučijo** tudi krčevite bolečine in kislo izpahovanje iz želodca, po driskah so zaprti in spet obratno.

GDEX: Slovene3

Janica si močno želi nastopiti vsaj na slalomu v Aspnu, toda še vedno jo **mučijo** bolečine.

Znano je, da jo **mučijo** bolečine v hrbtenici in da jemlje močne analgetike.

Zadnje čase jo namreč **mučijo** bolečine v kolkih in težko stoji.

Obenem po podatkih zdravstvenih statistik SZO številne bolnike s kroničnimi obolenji **mučijo** tudi bolečine v hrbtenici.

Prav tako z največjo težavo vstanete iz postelje, saj vas **mučijo** bolečine.

Po dveh do štirih dneh ga **mučijo** zaspanost, depresija in bolečine v zgornjem delu trebuha.

Poskrbeli bodo, da vas ne bodo več **mučile** bolečine v hrbtenici.

Na poti sta imela kar nekaj težav, saj je imel Leon angino, Marka so **mučile** bolečine v kolenih, oba pa sta dobila tudi žulje.

Matere ponavadi hitro ugotovijo, zakaj otroček joka: ali je lačen ali žejen, ali ga morda **mučijo** bolečine.

Ste zdravi in redno telovadite, a vas vseeno **muči** bolečina?

Janica si močno želi nastopiti vsaj na slalomu v Aspnu, toda še vedno jo **mučijo** bolečine.

Obenem po podatkih zdravstvenih statistik SZO številne bolnike s kroničnimi obolenji **mučijo** tudi bolečine v hrbtenici.

Znano je, da jo **mučijo** bolečine v hrbtenici in da jemlje močne analgetike.

Prav tako z največjo težavo vstanete iz postelje, saj vas **mučijo** bolečine.

Poskrbeli bodo, da vas ne bodo več **mučile** bolečine v hrbtenici.

Zadnje čase jo namreč **mučijo** bolečine v kolkih in težko stoji.

Na poti sta imela kar nekaj težav, saj je imel Leon angino, Marka so **mučile** bolečine v kolenih, oba pa sta dobila tudi žulje.

Matere ponavadi hitro ugotovijo, zakaj otroček joka: ali je lačen ali žejen, ali ga morda **mučijo** bolečine.

Ste zdravi in redno telovadite, a vas vseeno **muči** bolečina?

Po dveh do štirih dneh ga **mučijo** zaspanost, depresija in bolečine v zgornjem delu trebuha.

Janica si močno želi nastopiti vsaj na slalomu v Aspnu, toda še vedno jo **mučijo** bolečine.

Znano je, da jo **mučijo** bolečine v hrbtenici in da jemlje močne analgetike.

Zadnje čase jo namreč **mučijo** bolečine v kolkih in težko stoji.

Obenem po podatkih zdravstvenih statistik SZO številne bolnike s kroničnimi obolenji **mučijo** tudi bolečine v hrbtenici.

Prav tako z največjo težavo vstanete iz postelje, saj vas **mučijo** bolečine.

Po dveh do štirih dneh ga **mučijo** zaspanost, depresija in bolečine v zgornjem delu trebuha.

Poskrbeli bodo, da vas ne bodo več **mučile** bolečine v hrbtenici.

Na poti sta imela kar nekaj težav, saj je imel Leon angino, Marka so **mučile** bolečine v kolenih, oba pa sta dobila tudi žulje.

Matere ponavadi hitro ugotovijo, zakaj otroček joka: ali je lačen ali žejen, ali ga morda **mučijo** bolečine.

Ste zdravi in redno telovadite, a vas vseeno **muči** bolečina?

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)

www.barbarabrezigar.info	24
www.chucknorrisfacts.com	24
www.diabetictraveler.com	24
www.galerijaambienta.com	24
www.godsendinstitute.org	24
www.gsmworldcongress.com	24
www.ibm.com/sidejavnost:	24
www.indexprohibitorum.si	24
www.missioncleopatre.com	24
www.studentskazalozba.si	24
www.telekomunikacije.org	24
www.thezalavillabali.com	24
www.uselessknowledge.com	24
www.vzplamtiteznova.info	24
www1.k9webprotection.com	24
xxxxxxxxxxxxxxxxxxxxxxxxxxxx	24
zlatakartica@mercator.si	24

neznanstvenomorskiti	20
normálníti	20
northumberlandLˇkiti	20
novinarjinovinarjiti	20
obdavčeno predlagati	20
obdolžencema odvzeti	20
oblačííti	20
obtožnice predlagati	20
obvladovanjeöti	20
ocena97886989999zati	20
ocenjujetetrenerjati	20
odprofesionalizirati	20
odškodninoupokojeni	20
oficirjainjetavideti	20
ohranjajovzdrževati	20
oksibendazoldelovati	20
omogočauveljavljati	20

Remaining issues

- Sentence initial adverbs often found in “bad” examples
 - ▣ *nato, tako, torej, potem, poleg tega, zaradi tega*
 - ▣ after that, so, then, in addition, because (of that)
- Sentence ending with something other than full stop, exclamation mark or question mark
- Different configurations working better with different types of lemma:
 - ▣ Slovene3 and Slovene3b worked better for nouns and adjectives
 - ▣ Keyword position 40-100% did not work for some verbs
 - ▣ Evidence that more than one GDEX configuration for a language is needed

Future plans

- Further improvement of GDEX for Slovene
 - ▣ Adding blacklists
 - ▣ Adding multisense classifier (sloWNET synsets, Fišer, 2009)
- Using GDEX in automatic extraction of entry information (grammatical relations, collocates, examples)
 - ▣ Designing and testing different configurations for different types of lemmas
 - ▣ **Top** three examples always good