Automatic Extraction of Bilingual Slovak-English Equivalents

Radovan Garabík, Agáta Karčová

Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava
garabik@kassiopeia.juls.savba.sk, agatak@korpus.sk

COST ENeL WG3 meeting
Herstmonceux castle, UK, 2015-08-13

Parallel sk-en Corpus

- 11 million sentence pairs
- en 208 million tokens, sk 184 million tokens
- fiction: 5.1 million sentence pairs
- en 75.3 million tokens, sk 64.6 million tokens
- lemmatized, POS tagged, sentence aligned

Query

slovko Hľadať v korpuse par-sken-sk

áäčďéíĺľňóôöŕšťúüýžÁÄČĎÉÍĹĽŇÓÔÖŔŠŤÚÜÝŽ

[]^*„“"\=

O korpuse •  

#318

10517	Veril som , tak ako som veril odnepamäti , že sa nikdy nevyhýbal zodpovednosti . O svojom narodení som vedel , že to bol omyl , no ako som už spomínal , ani **slovkom** som nemohol spochybniť kvalitu života , ktorý mi zabezpečoval . V skutočnosti bola už polnoc , ked sa napokon začala rozchádzať aj hlavná časť zhromaždenia .	I still believed , as I always had , that he would never shirk responsibility for his actions . I knew my birth had been a mistake and , as I 'd often said , I had no quarrel at all with the quality of life he 'd given me since . It was indeed midnight by the time most of the crowd left to go home .
71993	Nie je v ňom ani **slovko** pravdy a mali by ste zaň čoskoro na krku žalobu za urážku na cti . “	∅
72687	Nečestný útok na nášho poslanca prostredníctvom očierňovania jeho syna . Na obvinení niet **slovka** pravdy . . . šokujúce . . . urážlivé . . . Popretie pôvodnej informácie a ospravedlnenie sa už pripravujú .	An unfair attack on our MP through his son . No truth in this allegation ... shocking ... libellous ... retractions and regrets are in the pipeline . ' ' The by - line in SHOUT ! says Usher Rudd . '
187441	„ Tak ako ? “ spýtala sa . Tón , akým vyriekla svoju otázku , dodal dvom malým , navonok bezvýznamným **slovkám** patričnú váhu a slečna Marplová hneď pochopila . „ Máš pravdu , “ prikývla .	Well ? " she said . It was a small insignificant word , but it acquired full significance from Mrs . McGillicuddy 's tone , and Miss Marple understood its meaning perfectly . " I know , " she said .
188464	„ Ty asi naozaj nemôžeš nič urobiť , “ prikývla zadumane slečna Marplová . (Keby bola pani McGillicuddyová venovala väčšiu pozornosť priatelkinmu tónu , azda by jej nebol ušiel ľahký dôraz na **slovku** ty .) „ Ohlásila si , čo si videla , železničným zriadencom aj polícii .	thoughtfully , " that there 's anything more you can do about it . " (If Mrs . McGillicuddy had been alert to the tones of her friend 's voice , she might have noticed a very faint stress laid on the you .) " You 've reported what you saw - - to the railway people and to the police .

Motivation

- make a good use of existing data
- replace/augment bilingual sk-en dictionaries
- (and sk-bg, sk-cs, too)
- while trying to reduce lexicographers' load

Motivation

- make a good use of existing data
- replace/augment bilingual sk-en dictionaries
- (and sk-bg, sk-cs, too)
- while trying to reduce lexicographers' load (to zero?)
- ... but still keep acceptable quality

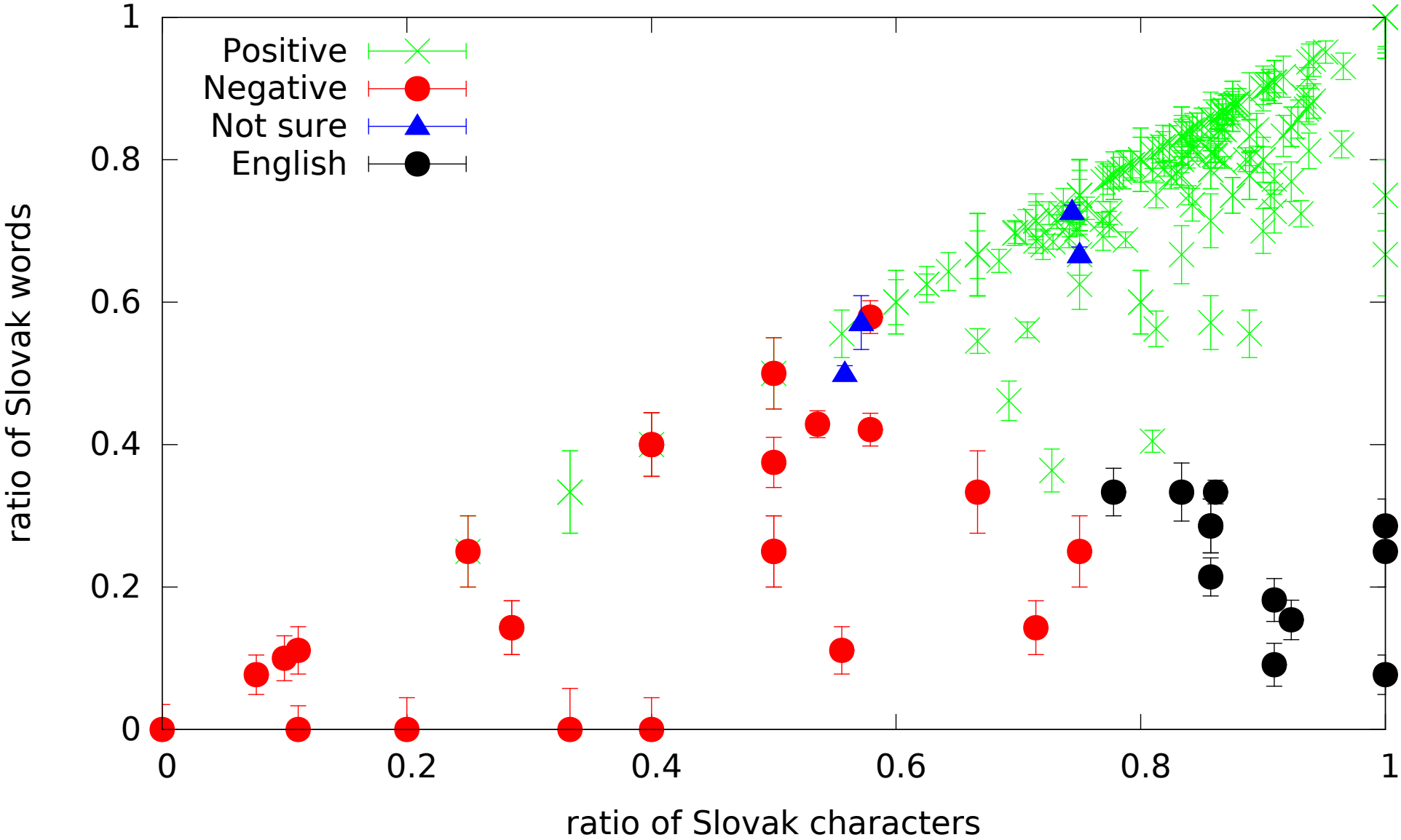
Phrase Extraction and Filtering

- acceptable quality of sentence alignment
- use GIZA++ (MOSES) to extract parallel equivalents (let's call them phrases)
- very disturbing noise (misaligned sentences, out of language texts)

```
as if they were throttling ||| akoby v nich škrtil ||| 1 4.0010
as if they were tired and ||| akoby boli unavené a ||| 1 0.0009
as if they were tired ||| akoby boli unavené ||| 1 0.00146031 1
as if they were true figures of ||| , akoby boli ozajstným stre
as if they were twigs . ||| , akoby to boli konáriky . ||| 1 7.3
as if they were twigs ||| , akoby to boli konáriky ||| 1 8.6126
as if they were watching television , ||| , akoby sledovali te
as if they were watching television ||| , akoby sledovali tele
as if they were watching ||| , akoby sledovali ||| 1 4.19488e-
as if they were weapons . ||| metajúci hromy - blesky . ||| 1 1
as if they were weapons ||| metajúci hromy - blesky ||| 1 1.709
as if they were well . ||| , ako keby boli zdraví . ||| 1 0.0002
as if they were well . ||| ako keby boli zdraví . ||| 1 0.00020
as if they were well ||| , ako keby boli zdraví ||| 1 0.0002390
as if they were well ||| ako keby boli zdraví ||| 1 0.00023908
as if they were yours ||| , akoby to boli tvoji ||| 1 1.2356e-0
as if they were ||| , ako keby boli ||| 0.25 0.0058746 0.018518
as if they were ||| , ako keby to boli ||| 1 0.00044178 0.03703
as if they were ||| , akoby boli ||| 0.192308 3.3213e-05 0.185
```

Cleaning up the Source Corpus

- remove unwanted elements
- find ‘improper’ language
- manual annotation of 219 random Slovak sentences
- $\rightarrow$ *positive*, *negative*, *not sure*
- 24 have been marked *negative* and 4 *not sure*.


Positive
Negative
Not sure
English
ratio of Slovak words
ratio of Slovak characters
0
0.2
0.4
0.6
0.8
1

- very good results
- final criterion: green trapezoid in the upper right corner

Frázy z paralelného slovensko-anglického korpusu.

informácií.

dodávateľa elektrickej energie ENEL ≈ the electricity supplier, ENEL
Dve energetické spoločnosti ENEL ≈ Two power companies, ENEL
ENEL bol v tom ≈ ENEL was at the
podnik Enel : výroba, rozvod ≈ for Enel : generation, transmission
Talianska, kde spoločnosť ENEL ≈ of Italy, where ENEL

Handbook of Slovak Nouns

- dynamic fusion of dictionary and corpus
- 933 (most frequent) nouns, most unique paradigms
- 2 main data sources:

Handbook of Slovak Nouns

- dynamic fusion of dictionary and corpus
- 933 (most frequent) nouns, most unique paradigms
- 2 main data sources:
  - corpus
  - morphological database

Handbook of Slovak Nouns

- dynamic fusion of dictionary and corpus
- 933 (most frequent) nouns, most unique paradigms
- 2 main data sources:
  - corpus
  - morphological database
- 2 main goals:

Handbook of Slovak Nouns

- dynamic fusion of dictionary and corpus
- 933 (most frequent) nouns, most unique paradigms
- 2 main data sources:
  - corpus
  - morphological database
- 2 main goals:
  - online
  - printed dictionary

Handbook of Slovak Nouns

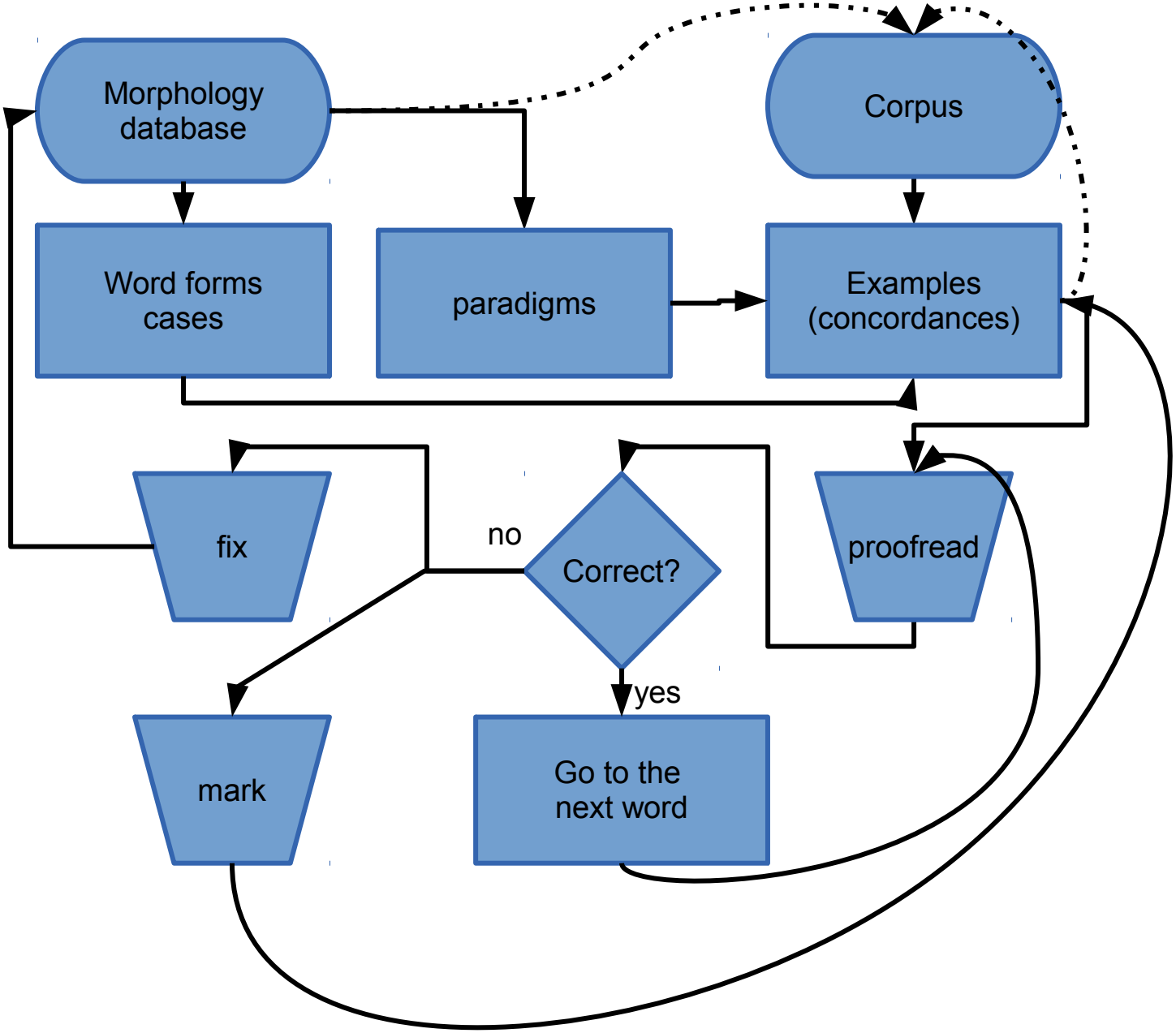
- dynamic fusion of dictionary and corpus
- 933 (most frequent) nouns, most unique paradigms
- 2 main data sources:
  - corpus
  - morphological database
- 2 main goals:
  - online
  - printed dictionary
- “Release early, release often” –
  `http://slovniky.korpus.sk/?d=noundb`

Handbook of Slovak Nouns

- accuracy of morphological analysis (and lemmatization): 94%
- very acceptable for a corpus
- completely unacceptable for a (printed) dictionary
- proofread!

Dictionary Compilation

1. get lemmas from frequency list
2. get paradigms from morphological database
3. for each case & inflected form, get a corpus example
4. proofread the examples (and the paradigm)
5. error in morphology: fix morphological database
6. error in corpus example (disambiguation): mark proper case
7. ... feed this later back into manually annotated corpus
8. re-train the tagger (points 5 & 7)
9. repeat (from point 3)


Morphology database
Corpus
Word forms cases
paradigms
Examples (concordances)
fix
no
Correct?
proofread
mark
yes
Go to the next word

The Dictionary

- online access – paradigms with (live, not proofread) corpus examples
- planned printed version – those 933 entries (each for one paradigm, represented by one noun)
- entry structure:
  - the paradigm
  - examples
  - nouns from the same class

Look at the Dictionary

Thank you for your attention