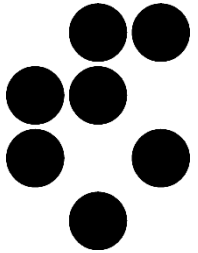


Monitor corpus of Slovene and automatic text categorization

IZTOK KOSEM

JOŽEF STEFAN INSTITUTE & UNIVERSITY OF LJUBLJANA

16. 6. 2022

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA KULTURO

Structure

- Monitor corpora
- The SLED project
- Monitor corpus of Slovene Trendi
- Thematic categorisation of texts

Monitor corpora

- Growing needs for resources that help with monitoring language use
- Why?
  - Detection of new words (lexical neologisms)
  - Detection of new senses (semantic neologisms)
  - Detection of new uses of existing words or multiword units
  - Detection of drops or cessations in word or sense usage
- For whom?
  - Lexicographers (e.g. dictionaries)
  - Linguists, sociolinguists etc. (research)
  - Teachers and learners (education materials)
  - Others

Monitor corpora

- Case studies
  - Bank of English (650+ mio words): since 1991, no info on updates
  - ONLINE (Czech National Corpus): since 2017, last update 2021 (ONLINE-NOW; ONLINE_ARCHIVE)
  - Timestamped JSI web corpora (Sketch Engine, 18 languages): 2014 - Apr 2021
  - NOW corpus (News on the web; Davies 2016-); 15 billion words, from 2010-now
  - Coronavirus corpus: 1.3 billion words, 2020 – November 2021
  - Corpus of Contemporary American English (COCA): over 1 billion words 1990 - March 2020

Monitor corpora

- Slovene resources
  - Jezikovni sledilnik (Language Monitor), Centre for Language Resources and Technologies, University of Ljubljana
    - https://viri.cjvt.si/sledilnik/eng/
    - Monthly calculations of most trending words
    - JSI Newsfeed used
    - Many steps (data filtering, calculations) still manually run, no pipeline
  - Growing Dictionary of the Slovenian Language (Fran Ramovš Institute of the Slovenian Language)
    - source of info unclear

SLED project

- Call by the Ministry of Culture:
  - Funding of projects aimed at the development and upgrading of infrastructure for Slovene in the digital environment
- Funding period: Sept. 2021 – Oct. 2022
- Amount: 58.300 EUR
- Project website: https://sled.ijs.si

- Core team:
  - Iztok Kosem (project leader)
  - Simon Krek
  - Jaka Čibej
  - Kaja Dobrovoljc
  - Luka Krsnik
  - Nikola Ljubešič
  - Primož Ponikvar
  - Janez Brank







Activities and aims

- **Activity 1: Monitor corpus of Slovene**
- **Activity 2: Data collections**
  - Survey into user needs
  - Data containing various calculations on word usage (e.g. words or phrases with highest surge in recent month(s) or year(s), word of the day etc.)
  - Regular uploads of data into CLARIN.SI repository
- **Activity 3: Automatic text categorization (by topic)**

Trendi - monitor corpus of Slovene (1)

- **Contents of Trendi**
  - to complement Gigafida, a reference corpus of written standard Slovene
  - 110 sources (selected from 243 sources in JSI Newsfeed)
  - Version 2022-05 just published!
    - 565 million tokens, nearly 20 million tokens per month
    - News sources regularly added (e.g. necenzurirano.si)

Spremljevalni korpus Trendi 2022-05 // Monitor corpus Trendi 2022-05

Counts	
Tokens	565,308,991
Words	473,161,507
Sentences	25,186,942
Paragraphs	8,326,466
Documents	1,436,548

General info	
Corpus description	Document
Language	Slovenian
Encoding	UTF-8
Compiled	06/10/2022 18:51:54
Tagset	Description

Lexicon sizes	
word	1,842,693
lempos	1,087,105
tag_en	1,303
tag	1,303
ud_pos	17
ud_feats	2,269
id	9,770
ud_dep	33
ud_head_lemma	656,488
ud_head_tag_en	1,244
ud_head_tag	1,244

Word list

Corpus: **Trendi (spremljevalni)**
Total number of items: **142**
Total frequency: **1,436,548**

text.publisher	document frequency
Slovenska tiskovna agencija STA	228,456
MMC RTV Slovenija	121,357
STA d.o.o.	97,916
Delo	81,789
24ur.com	76,333
Siol.net Novice	57,555
Dnevnik	55,335
vecer.com	53,538
Svet24	34,714
Slovenske novice	31,356
Žurnal24	30,465
Športni Dnevnik Ekipa	30,039
Tednik Demokracija	28,902
Vestnik	28,760
Gorenjski Glas	27,964

Keywords (Trendi vs Gigafida 2.0)

	Trendi (spremljevalni)		*Gigafida v2.0 (referenčni, dedupliciran)*		
lempos	frequency	frequency/mill	frequency	frequency/mill	Score
quot-s	601,486	1064.0	12	0.0	107.3
nbsp-s	384,144	679.5	119	0.1	68.3
koronavirus-s	296,948	525.3	172	0.1	52.8
covid-s	228,868	404.9	0	0.0	41.5
pandemija-s	136,041	240.6	1,573	1.2	22.4
epidemija-s	211,646	374.4	9,936	7.5	22.0
cepivo-s	159,051	281.4	12,703	9.5	14.9
okužba-s	327,586	579.5	40,839	30.6	14.5
sta-n	172,811	305.7	17,203	12.9	13.8
karantena-s	74,140	131.1	2,574	1.9	11.8
piškotek-s	69,415	122.8	1,797	1.3	11.7
cepljenje-s	139,038	246.0	19,089	14.3	10.5
Nijz-s	44,790	79.2	76	0.1	8.9
coviden-p	41,216	72.9	0	0.0	8.3
cepljen-p	56,378	99.7	4,340	3.3	8.3
epidemiološki-p	46,091	81.5	1,564	1.2	8.2
href=-s	40,361	71.4	0	0.0	8.1
Šarec-s	48,100	85.1	2,345	1.8	8.1
okužen-p	101,315	179.2	20,149	15.1	7.5
covid-k	36,131	63.9	0	0.0	7.4
LMŠ-s	41,197	72.9	1,877	1.4	7.3
PCR-s	34,326	60.7	142	0.1	7.0
Levica-s	32,515	57.5	101	0.1	6.7
PCT-s	28,408	50.3	57	0.0	6.0
cepiti-g	44,880	79.4	7,026	5.3	5.9
odmerek-s	76,880	136.0	20,024	15.0	5.8
zajezitev-s	32,542	57.6	2,592	1.9	5.7
Dončić-s	47,801	84.6	9,402	7.1	5.5
testiranje-s	89,671	158.6	27,429	20.6	5.5

Trendi - monitor corpus of Slovene (2)

- **Article collection and annotation pipeline**
  - Articles downloaded daily, in JSON format; JSI Newsfeed service used for now
  - Deduplication at URL level
  - Annotation:
    - Classla-stanza (Ljubešić and Dobrovoljc 2019): sentence segmentation, tokenisation, morphosyntactic annotation and lematisation
    - parsing using Universal Dependencies
    - Named-entity recognition

Trendi - monitor corpus of Slovene (3)

- **Article collection and annotation pipeline**
  - Conversions:
    - CONNL-U → TEI XML (daily)
    - TEI-XML → VERT (monthly)
- **Availability**
  - In corpus tools KonText CLARIN.SI and NoSketchEngine CLARIN.SI
  - ccTRendi version (selected paragraphs from each text) in CLARIN.SI repository under CC 4.0 BY-SA

Automatic text categorization (1)

- **Developing a classification of categories**
  - Three sources:
    - Six popular read news portals
    -
    -
  -
  -
  -

Aktualno

Odločitev vlade v sredo / Prvi sveženj ukrepov proti draginji naj bi zajemal pogonska goriva

Energetiki napovedujejo dodatne podražitve električne energije. Vlada medtem pripravlja ukrepe za blažitev posledic energetske in prehranske draginje. Prvi sveženj naj bi bil že pripravljen za sprejemanje v sredo na seji vlade.

Dirka po Sloveniji / Pogačar: Vesel sem, da sem na startu ene najlepših dirk na svetu

Na 28. Dirki po Sloveniji bo lansko zmago branil Tadej Pogačar, prvi kolesar svetovne lestvice, ki je tudi letos glavni favorit, da v Novo mesto prikolesari v zeleni majici.

Automatic text categorization (1)

- **Developing a classification of categories**
  - Three sources:
    - Six popular read news portals
    - Thematic codes by IPTC (International Press Telecommunications Council)
    - Categories from corpora, esp. SYN_2015 and Estonian National Corpus
  - Aim was to have a relatively small number of categories
  - To compare: IPTC 17 top-level (out of 1400); SYN_2015 13; ENC 25
  - Final set: 13 categories

Trendi category	Slovene	SYN_2015	ENC	IPTC
Arts and culture	5	culture	culture & entertainment	arts, culture and entertainment
Crime and accidents	6	crime	/	disaster and accident
Economy	6	economy	economy, finance & business + 2	economy, business and finance; labour
Environment	2	/	nature & environment	environmental issue
Health	3	/	health	health
Leisure	4	leisure	beauty + 7	lifestyle and leisure
Politics and Law	1	politics	politics & government	Politics + 2
Science and Technology	5	/	science, technology & IT	science and technology
Society	1	social life	society; religion; sex; women	social issue + 2
Sports	6	sports	sports	sport
Weather	4	/	/	weather
Entertainment	4	/	culture & entertainment*	arts, culture and entertainment*
Education	1	/	education	education

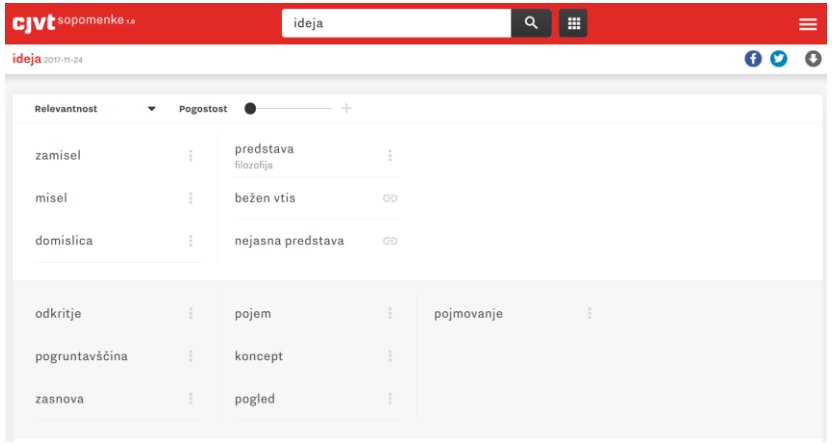
Automatic text categorization (2)

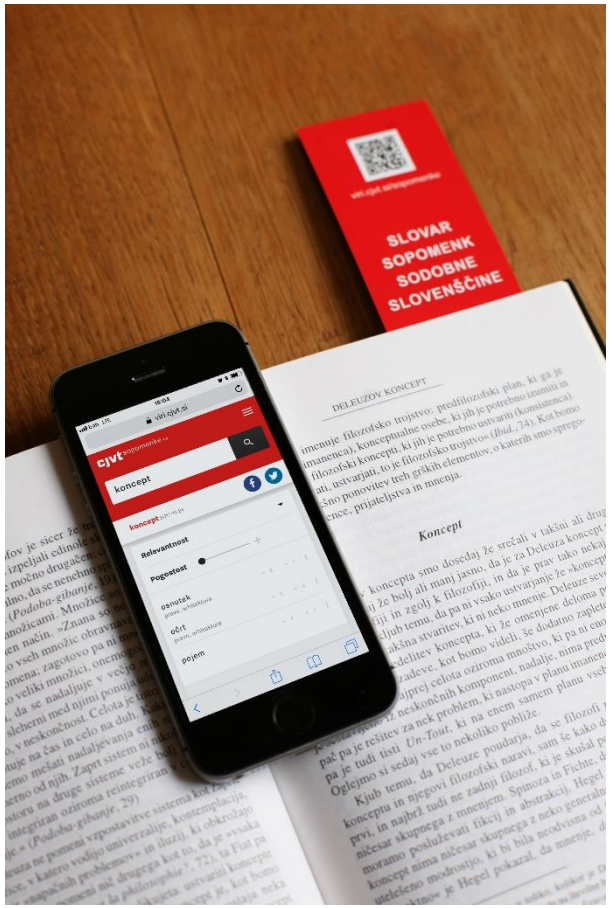
- **Preparing training datasets**
  - class balance: each category to contain an equal number of examples
  - source balance: equal or similar number of examples from various resources
  - Two datasets: 1000 instances per category; 3000 instances per category (to see if larger means better)
  - Evaluation datasets: 100 per category for development, 100 for evaluation
  - Problems:
    - Some categories have subpages on only certain portals (e.g. education, weather)
    - Some categories even do not have subpages (e.g. politics, society)

Future work

- Regular monthly updates of the Trendi corpus
- Regular updates of statistical calculations
- Finishing the development of automatic text categorisation tool
- Finishing the development of a new newsfeed collection pipeline
- And last but not least…

Implementation into lexicographic practice!





Thank you!

Hvala!

Aitäh!