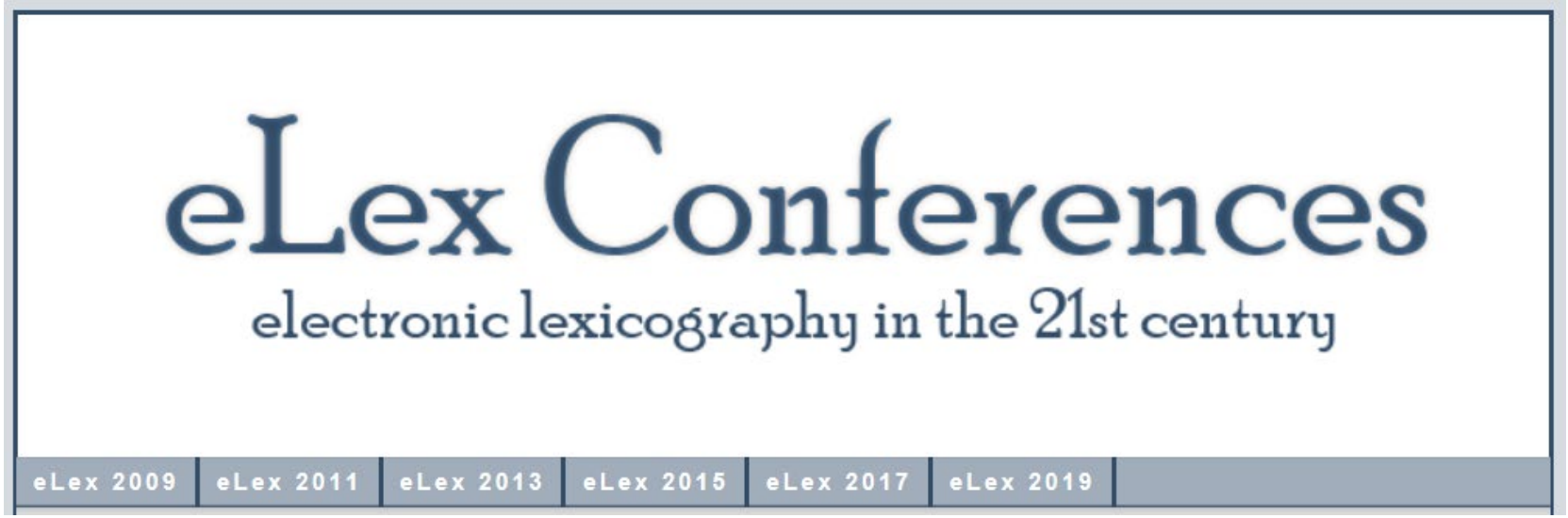
SMART LEXICOGRAPHY

Sintra, Portugal

1–3 Oktobar 2019.

https://elex.link/elex2019/



- eLex 2009, eLexicography in the 21st century: new challenges, new applications, Louvain-la-Neuve, Belgija.
- eLex 2011, Electronic lexicography in the 21st century: new applications for new users, Bled, Slovenia
- eLex 2013, electronic lexicography in the 21st century: thinking outside the paper, Talin, Estonija.
- eLex 2015, electronic lexicography in the 21st century: Linking lexical data in the digital age, Saseks, Engleska.
- eLex 2017: Lexicography from Scratch, Lajden, Holandija
- eLex 2019: Smart Lexicography, Sintra, Portugal.
- eLex 2021: ?????, Brno, Češka.

Smart Lexicography ~ Pametna leksikografija?

eLex 2019

- Bez papira! Mobilna app http://conference4me.eu/download/ .

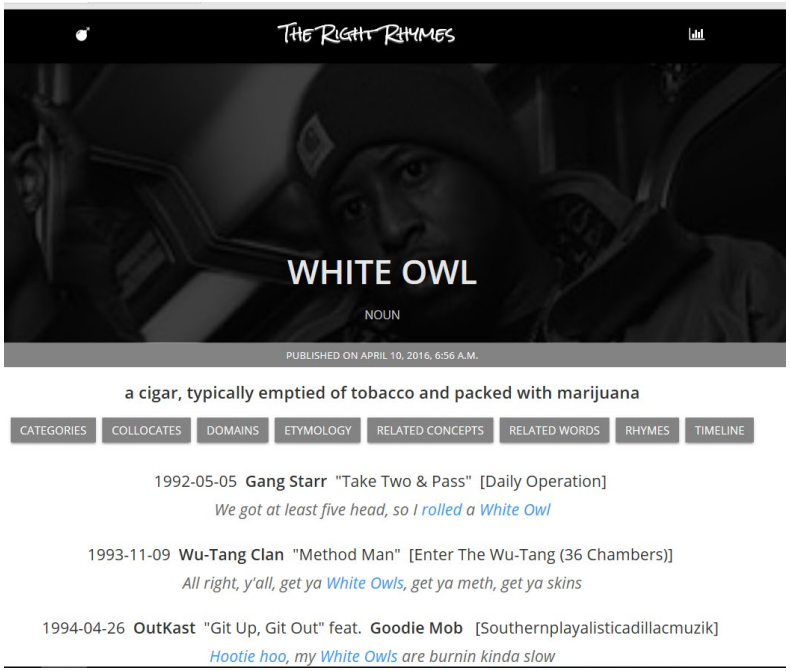
- Prenos uživo, ali i dalje dostupan na: https://www.youtube.com/user/VideoLecturesChannel/featured
- Svi radovi su dostupni pod CC licencom za sve godine, https://elex.link/
- Glavni organizator svih eLex-a **Iztok Kosem,** Centar za primenjenu lingvistiku, Trojina / Fakultet umetnosti, Univerzitet u Ljubljani

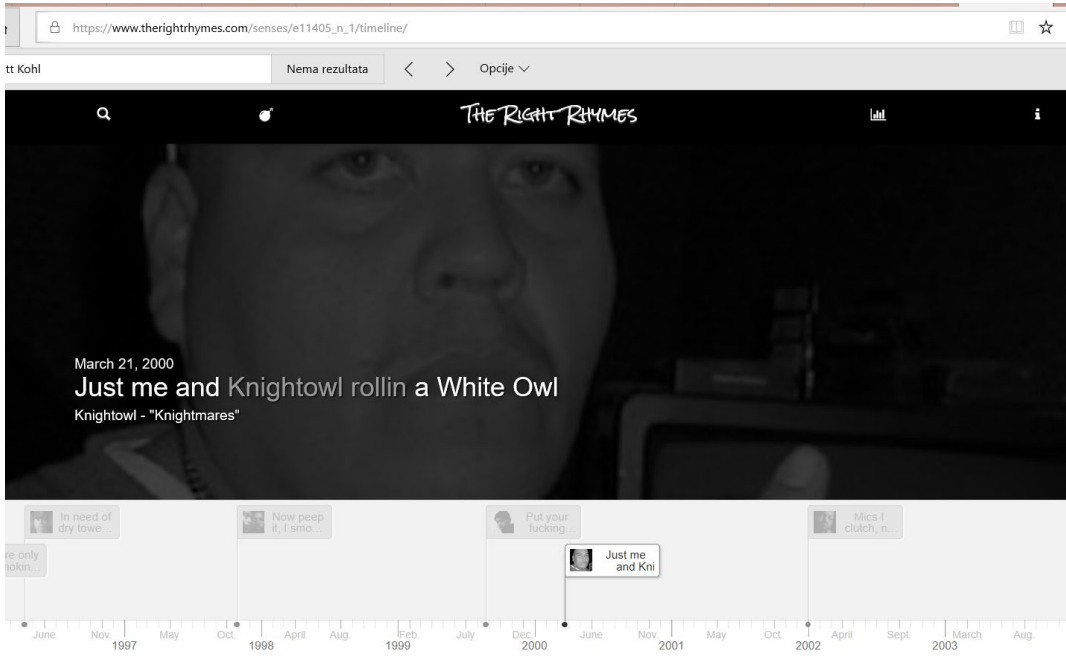
Keynote speakers

- Margarita Correia, CELGA-ILTEC, University of Coimbra / Univ. of Lisbon,
  VOC, a Spelling Dictionary for the Portuguese Language – role and characteristics
- Matt Kohl, GeoPhy, dobitnik nagrade Adam Kilgarriff
  The Right Rhymes: Smart Lexicography in Full Effect
- Alexander Geyken, Berlin-Brandenburg Academy of Sciences and the Humanities,
  The Center for digital lexicography of the German Language: new perspectives for smart lexicography
- David Baines, SIL International
  SIL's language data collection
- Maciej Piasecki, Wroclaw University of Technology
  Wordnet as a Relational Semantic Dictionary Built on Corpus Data

The Right Rhymes: Smart Lexicography in Full Effect

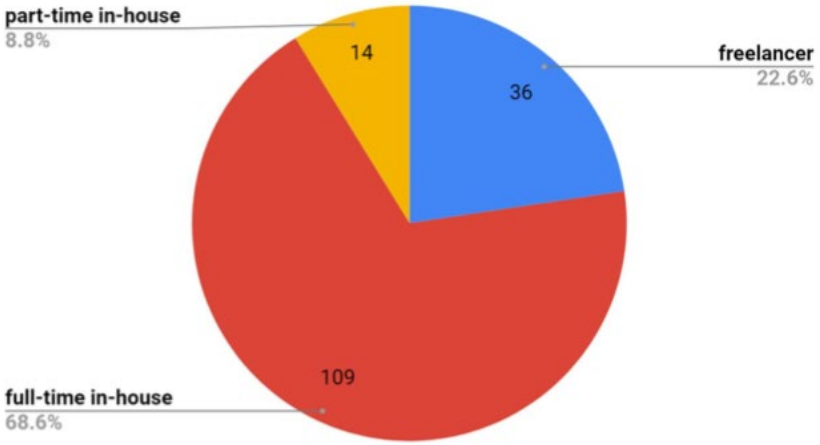
- Matt Kohl, GeoPhy (ranije OxfordPress)
- https://www.therightrhymes.com/

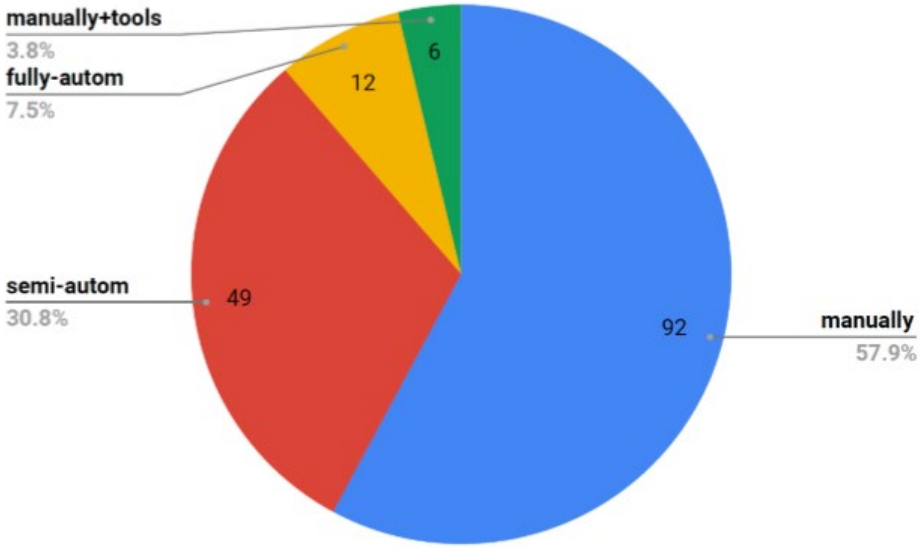




Lexicographic Practices in Europe: Results of the ELEXIS Survey on User Needs

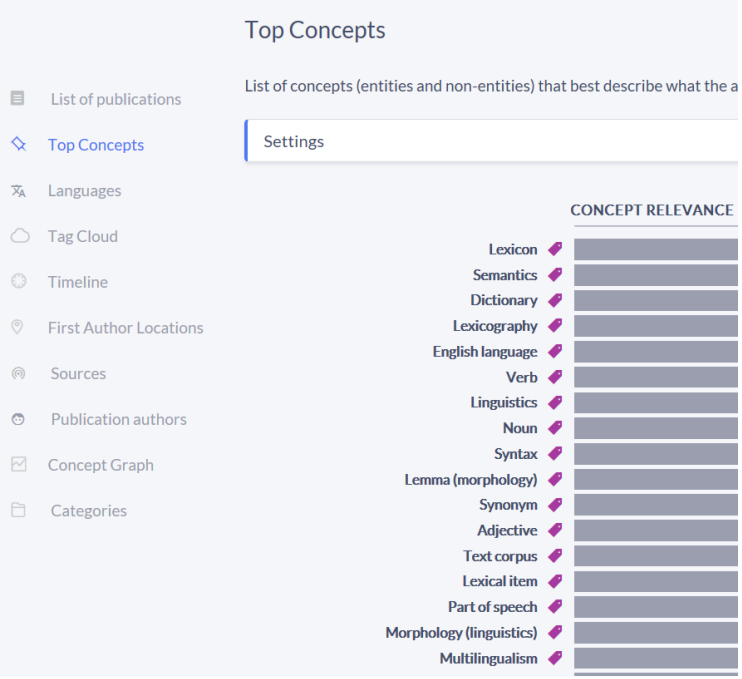
- Jelena Kallas, Svetla Koeva, Horizon 2020 project ELEXIS
- Najčešće korišćena kombinacija je sopstveni DWS i komercijalni CWS (npr. Ekilex, Sketch Engine) 55%
- CWS: Sketch Engine, IMS Open Corpus Workbench, CoRes , Korp, NoSketchEngine, AntConc, COSMAS II …
- Potreba za obučavanje i tesna saradnja sa IT i NLP

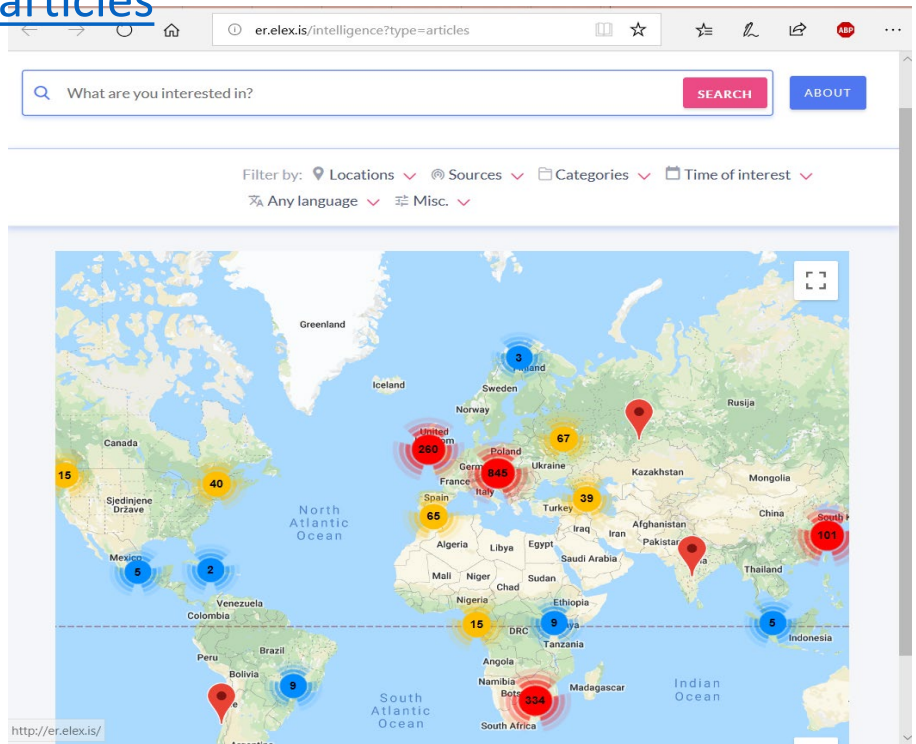




ELEXIFINDER: A Tool for Searching Lexicographic Scientific Output

- Iztok Kosem, Simon Krek
- Pretraga bibliografskih podataka iz oblasti leksikografije
- Zasnovan na Event Registry arhitekturi http://eventregistry.org/
- Elexifinder: 1,755 publikacija, 78 videa, 11 jezika
- http://er.elex.is/intelligence?type=articles





Karst Exploration: Extracting Terms and Definitions from Karst Domain Corpus

- Senja Pollak, Andraž Repar, Matej Martinc, Vid Podpečan
- karstology; term extraction; term embeddings; term alignment; definition extraction; triplets; specialized corpora
- Ekstrakcija termina: statistički + prišireњe liste termina korišćenjem embedingsa
- Poravnanje sl-en termina
- Ekstrakcija kandidata za definicije
- Ekstrakcija relacija (na žalost bez detalja u ovom radu)

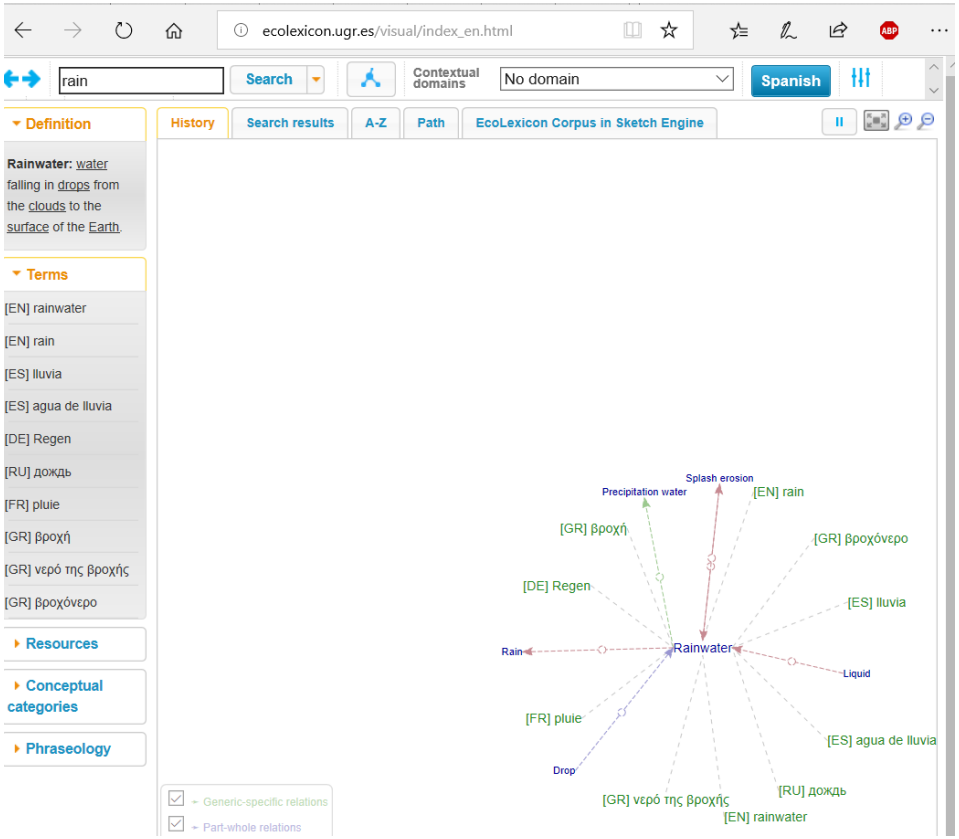
Enriching an Explanatory Dictionary with FrameNet and PropBank Corpus Examples

- Proširenje Online rečnika Tezaurs.lv (Litvanski) sa primerima iz semantički anotiranog korpusa
- Tezaurs.lv itegriše više d 300 rečnika (retrodigitalizovanih)
- Primeri potiču iz FrameNet-a i PropBank korpusa z alitvanski https://tezaurs.lv/#/sv/aizvērt

Figure 5: A corpus example ("as soon as Sophie had closed the garden gate behind her [she opened the envelope]") with parallel FrameNet and PropBank annotation, illustrating the sense and use of the headword "aizvērt" ('to close').

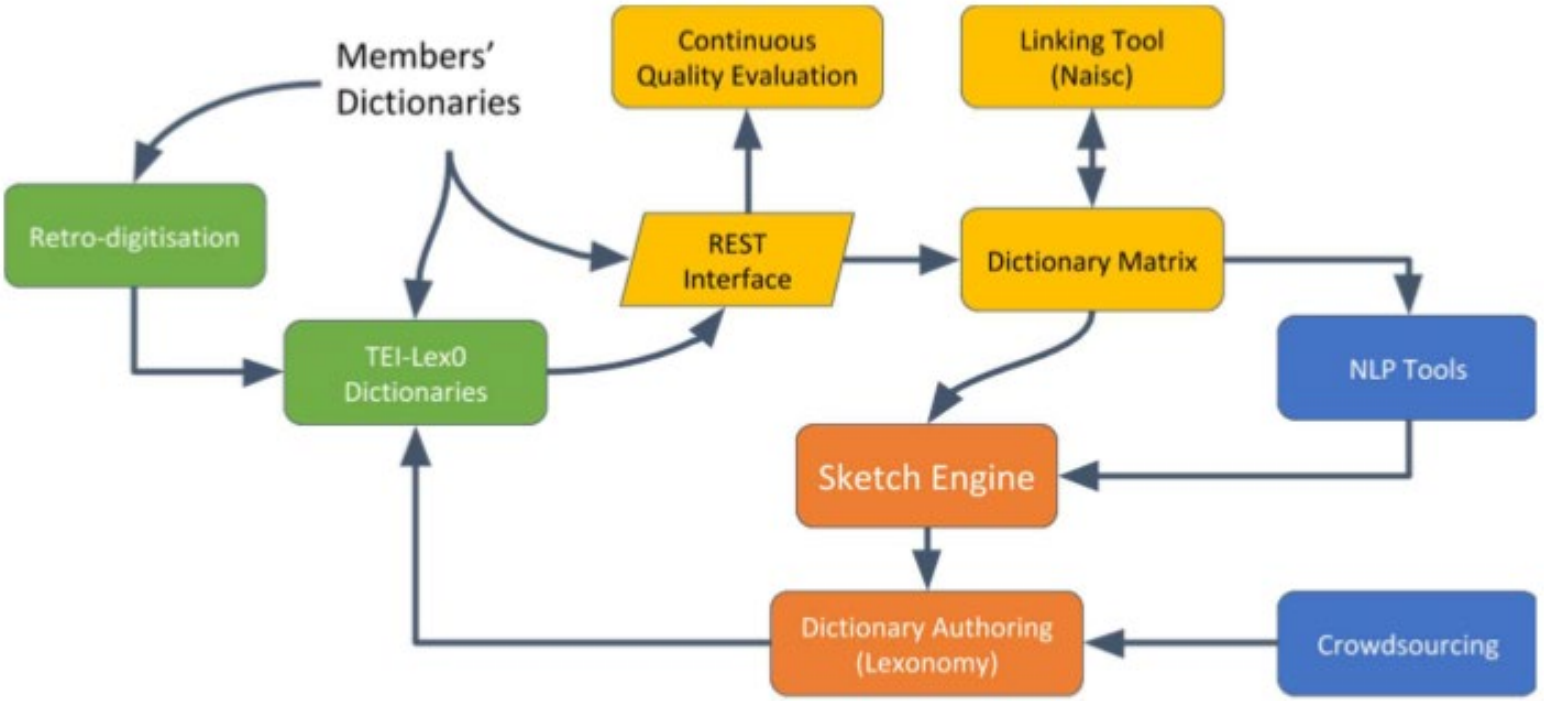
Ontological Knowledge Enhancement in EcoLexicon

- http://ecolexicon.ugr.es/visual/index_en.html
- Juan Carlos Gil-Berrozpe, Pilar León-Araúz, Pamela Faber



The ELEXIS Interface for Interoperable Lexical Resources

- John P. McCrae, Carole Tiberius, Anas Fahad Khan, Ilan Kernerman, Thierry Declerck, Simon Krek, Monica Monachini and Sina Ahmadi



Towards Electronic Lexicography for the Kurdish Language

Sina Ahmadi,
Hossein Hassani,
John P. McCrae

https://github.com/KurdishBLARK/KurdishLex

bend *f* bond; **li ~a** for the sake of, chained to, waiting for: *divê em êdî li benda sibehê ranewestin* we shouldn't stand around waiting for tomorrow; **~ kirin** *v.t.* to fetter, arrest; **man di ~a** to wait for

```
:lexicon a lime:Lexicon;
    lime:language <www.lexvo.org/page/iso639-3/kmr> ;
    lime:entry :lex_bend .

:lex_bend a ontolex:LexicalEntry, ontolex:Word ;
    ontolex:canonicalForm :form_bend ;
    rdfs:label "bend"@kmr-latn .

:form_bend a ontolex:Form ;
    dct:language <www.lexvo.org/page/iso639-3/kmr> ;
    ontolex:writtenRep "bend"@kmr-latn ;
    lexinfo:partOfSpeech lexinfo:noun ;
    lexinfo:gender lexinfo:feminine ;
    lexinfo:number lexinfo:singular ;
    ontolex:sense :bend_n_sense .

:en_bond a ontolex:LexicalEntry ;
    dct:language <http://lexvo.org/id/iso639-1/en> ;
    ontolex:sense :en_bond_sense .

:trans a vartrans:Translation ;
  vartrans:source :bend_n_sense ;
  vartrans:target :en_bond_sense .

:bend_n_sense a lexicog:UsageExample ;
  rdf:value "divê em êdî li benda sibehê
      ranewestin."@kmr-latn ;
  rdf:value "we shouldn't stand around waiting for
      tomorrow."@en .
```

SASA Dictionary as the Gold Standard for Good Dictionary Examples for Serbian

Ranka Stanković[1], Branislava Šandrih[1], Rada Stijović[2], Cvetana Krstev[1], Duško Vitas[1], Aleksandra Marković[2]

[1] University of Belgrade,
[2] Institute for Serbian Language SASA

eLex 2019

SMART LEXICOGRAPHY, Sintra, Portugal, 1–3 October 2019.

Overview

- Introduction
  - aim, role, SASA-Dataset
- SASA Dictionary
  - retro-digitization, modernisation, current practice
- Features of Dictionary Examples
  - role, extraction, APIs
- Feature analysis
  - gold and control dataset, distribution
- Preliminary Model for Identifying GDEX
- Future work and concluding remarks

The Current Practice of Dictionary Example

Interventions on examples

In the analysed dataset

- ~60,000 dictionary entries
- ~105,000 lexical units
- 70% of entries have examples

Summary of interventions:

- 66% examples were not modified;
- 20% had one shortening and no insertions;
- 6% more than one shortening and no insertions;
- 5% had an insertion but were not shortened;
- 2% had both insertions and shortenings

The Features of Dictionary Examples

The Role of Example Features

Initial set of features was

- inspired by (Kilgarriff et al., 2008; Kosem, 2017), and
- guided by recapitulation in (Kosem et al., 2019)

GDEX for Serbian ranks candidate sentences

Gold standard for the development of our method

- dictionary examples from 5 out of 20 volumes of DSA
- they were manually selected by experienced lexicographers
- multiple check-ups (through several phases)

The Features of Dictionary Examples

Feature Extraction (14 out of 41)

Character-based:

- sentence_length: Number of all characters
- no_digits: Number of digits
- no_weird_chars: Number of characters ("#$%&\'()*+-/:;<=>?@[\\]^_`{|}~’„”…)
- no_commas: Number of commas
- no_punctuation: Number of all punctuation marks

Token-based:

- No_all_tokens: Number of all tokens
- avg_token_len: Average token length
- max_token_len: Max token length
- no_all_words: Number of all words (contiguous sequence of letters)
- avg_word_len: Average word length
- no_capitalised_words: Number of words that begin with uppercase, which are not at the beginning of the sentence
- no_rare_tokens: Number of tokens with frequency ≤ threshold in the referent corpus
- avg_freq_in_corpus: Average word frequency in the referent corpus

Syntactic features:

- no_pronouns: Number of tokens tagged as pronouns

Feature analysis - Gold Dataset

Examples from SASA Dictionary supplied

- volume, headword, POS, linguistic labels
- type of editor's intervention, bibliographical source

Size of the gold corpus

- 133,904 examples comprising 1,711,231 words

Three types of partitioning were used:

- published volume (D01, D02, D18, D19 and D20),
- type of lexis/language (DSS standard, DNS non-standard)
- POS of the headword (N, V, A, ADV and X).

Feature analysis - Control Dataset

Digital library Biblisha

- collection with contemporary novels (CN) :
  - 7 novels written by contemporary Serbian writers and from
  - 7 novels written in German and translated to Serbian.
- domain knowledge two scientific journals (labeled SJ)
  - *he Journal for Digital Humanities Infotheca* and
  - *Underground Mining Engineering*.

SrpKor – Corpus of contemporary Serbian (DP)

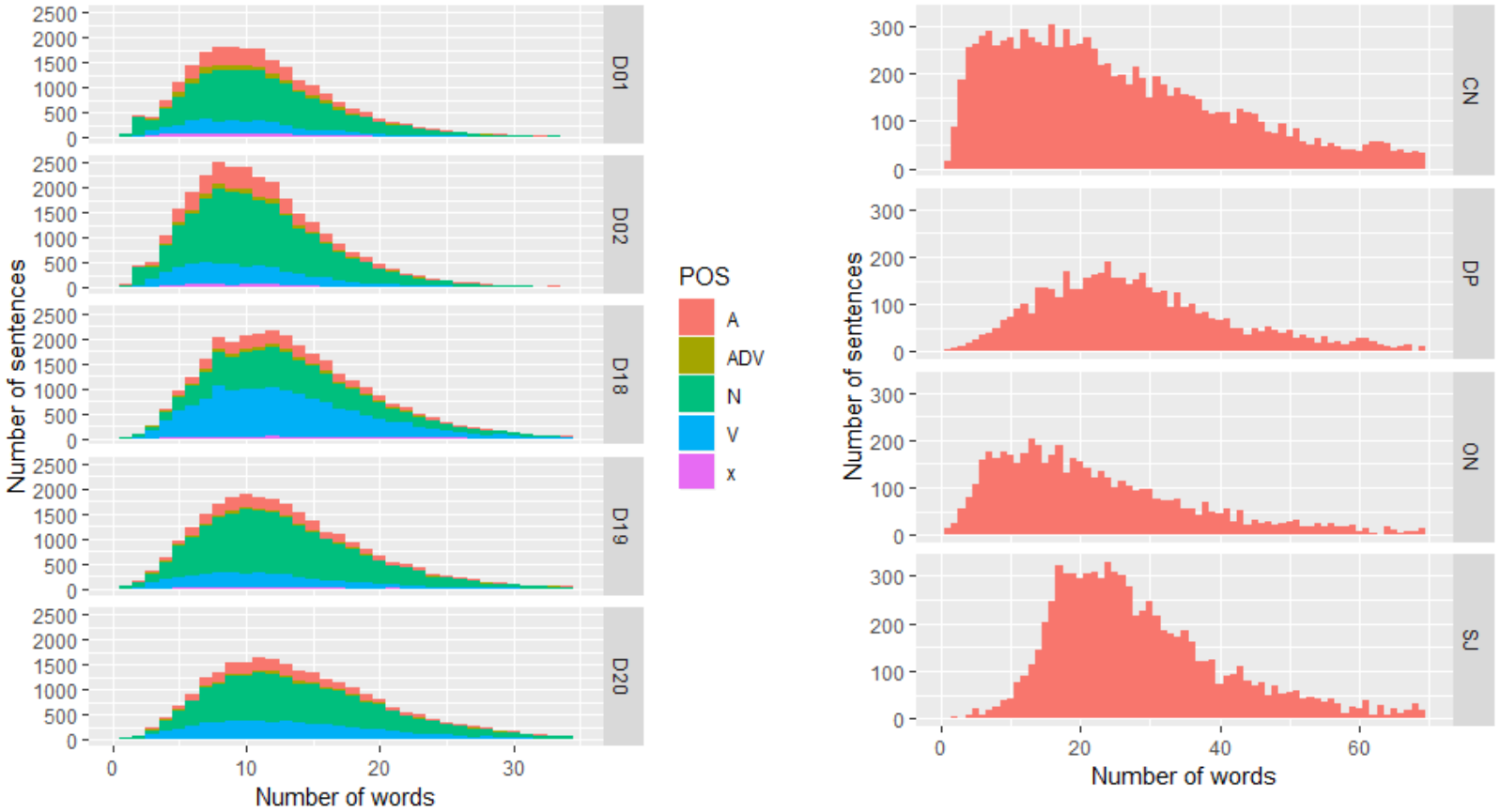
- 17 issues of the daily newspaper *Politika* published in 2001–2010

Serbian ELTeC Collection old novels (ON)

- 10 novels and excerpts from 15 novels published 100 or more years ago

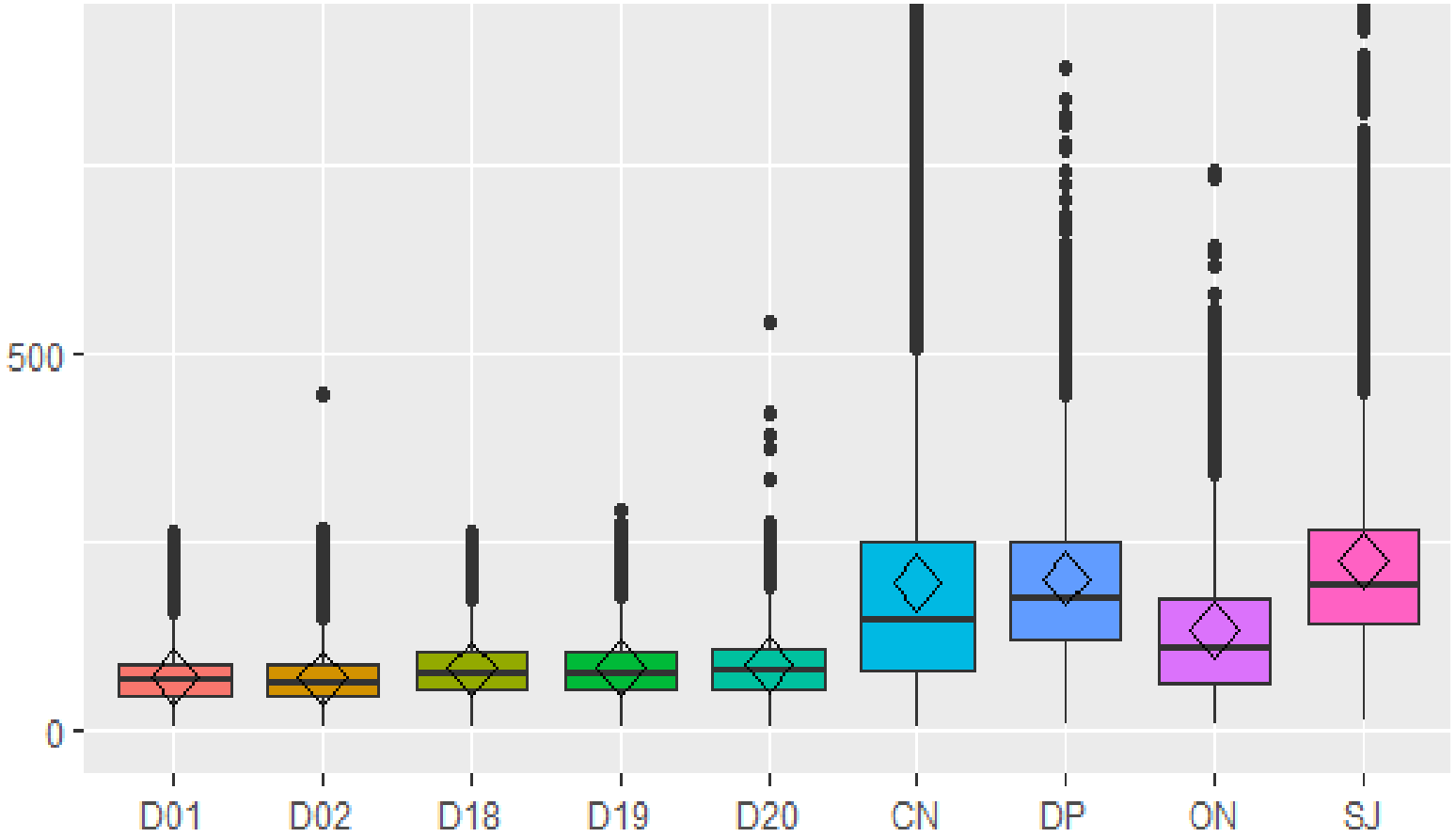
Feature analysis

- Feature distribution in the gold dataset of good examples
  - The histogram of the number of words in examples



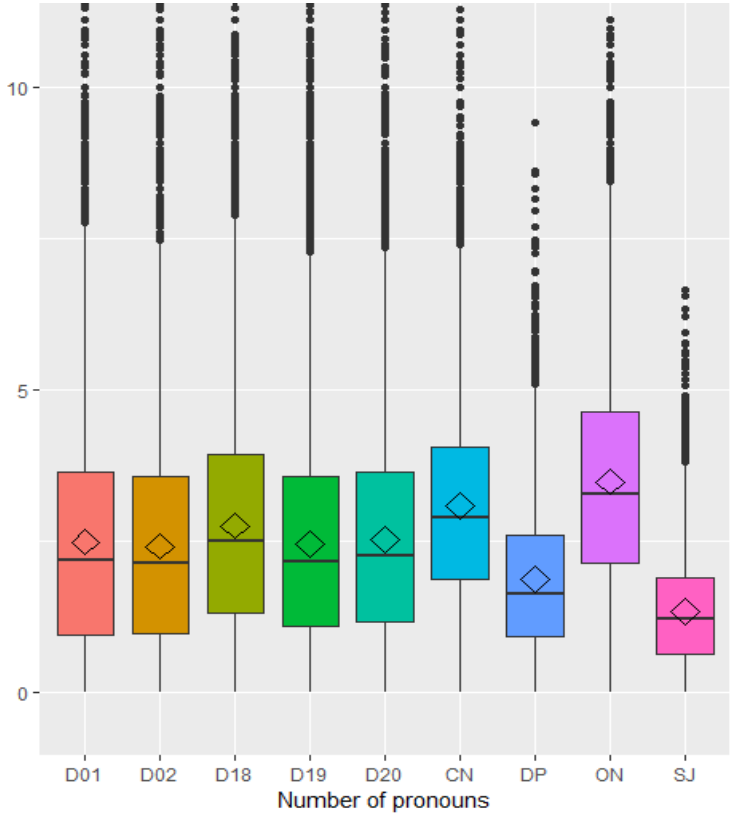
Feature analysis

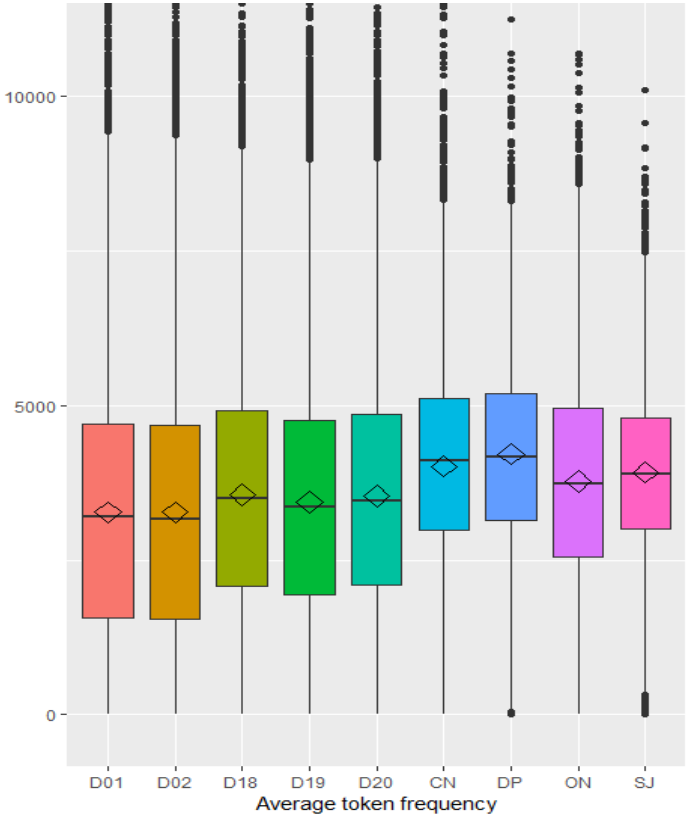
- Feature distribution on both corpora
  - Boxplot of sentence (example) length (in number of characters) per partition



Feature analysis

- Feature distribution on both corpora
  - Boxplot of number of pronouns and token frequency per partition





Preliminary Model for Identifying Good Dictionary Examples

Filtering and ranking

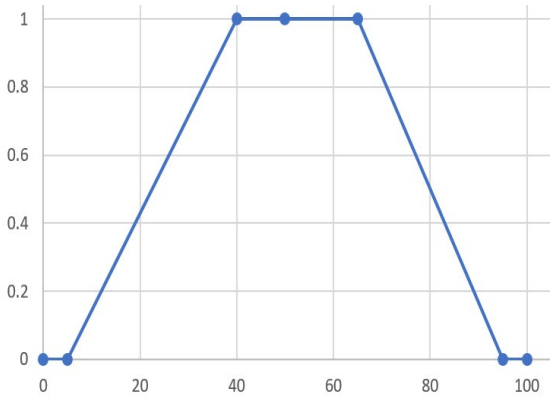
- rules obtained from analysed data (feature vectors)
- combined into a single score.

GDEX function

- inspired by the state of the art functions:
- *blacklist(), greylist() and optimal_interval()*.

optimal_ interval function

- four key percentiles from the gold SASA dataset
- feature values <= first and > last are assigned 0.01,
- in the middle interval scores are 1, and
- between them a linear interpolation function is used.



40th and 65th percentile of SASA dictionary for number of words are the same as values in the example given to the Sketch engine...

greylist function

- two key values are used (5th and 95th percentiles)
- values lower of the 5th are assigned a score of 1,
- higher than 95h a score of 0, and
- between them linear interpolation is used.

Preliminary Model for Identifying Good Dictionary Examples

Gold dataset: training (80%) and a validation set (20%)
(NO non-standard; OK standard Serbian)

Results of the Logistic Regression binary classifier

		Precision	Recall	F1-score	Number of samples
NO	(NS)	0.84	0.68	0.75	11,056
OK	(SS)	0.73	0.87	0.79	11,180
ALL		0.78	0.77	0.77	22,236

- The future system for semi-automatic identification of good dictionary examples implies development of more modules
  - user interface for feature extraction and
  - fine tuning for GDEX parameters,
  - Integration with corpus
- Evaluation of first results of the developed core components is encouraging.

eLex 2021 Brno, Češka

Ali pre toga (srodni):

- LREC 2020 (25. novembar apstrakt) Marsej maj 2020 https://lrec2020.lrec-conf.org/en/
  Tema: Less Resourced and Endangered Languages

- EURALEX, Alexandroupolis, (3. februar apstrakt) Grčka, 8-12 Septembar 2020 https://euralex2020.gr/
  Tema: *Leksikografija za inkluziju*

obrigado pela atenção

hvala na pažnji

хвала на пажњи