/instituut voor de
Nederlandse taal/

The Future of Academic Lexicography -- A White Paper

Kris Heylen & Vincent Vandeghinste

/instituut voor
de Nederlandse
taal/

Steurs F. , T. Schoonheim, K. Heylen and V. Vandeghinste (2020).
The Future of Academic Lexicography -- A White Paper.
Version 1.2. Available at www.ivdnt.org

In Loving Memory of
Tanneke Schoonheim (+ 25-08-2020)

Overview

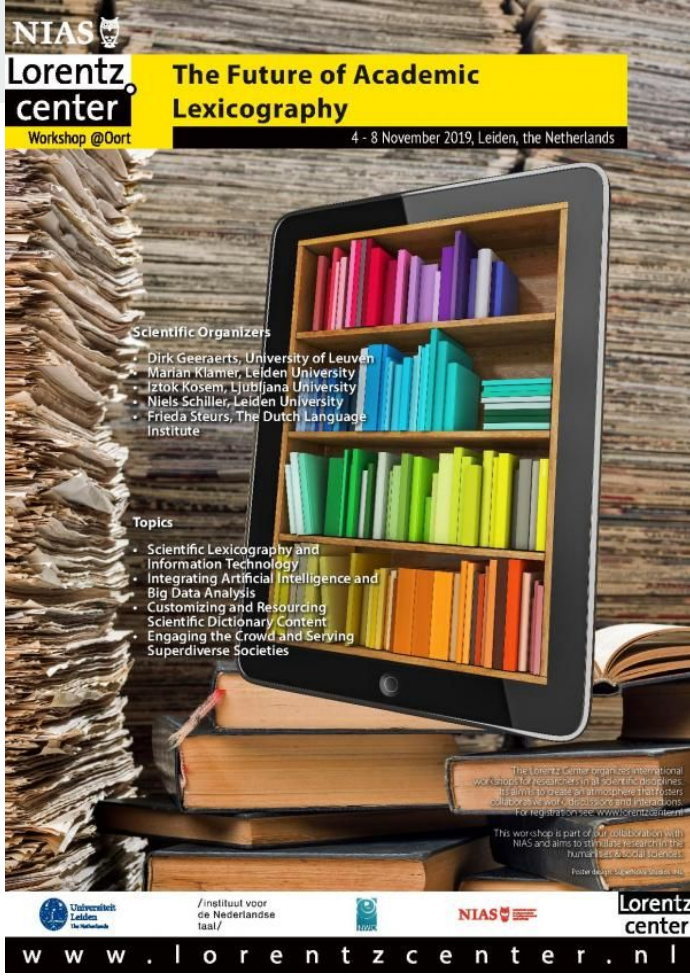
- Background of the White Paper
- Challenges for the Future of Academic Lexicography
  - Challenge 1. Role and Place in Society and Science
  - Challenge 2. Genericity vs. Specificity
  - Challenge 3. Scalability: Removing the bottlenecks
- White Paper: Conclusions and Recommendations
- Panel Session: The Future Starts with Post-editing Lexicography
- General Discussion

Overview

- **Background of the White Paper**
- Challenges for the Future of Academic Lexicography
  - Challenge 1. Role and Place in Society and Science
  - Challenge 2. Genericity vs. Specificity
  - Challenge 3. Scalability: Removing the bottlenecks
- White Paper: Conclusions and Recommendations
- Panel Session: The Future Starts with Post-editing Lexicography
- General Discussion

Background: the Workshop

- five-day NIAS-Lorentz workshop (Leiden, 4-8 November 2019)
- Organizers:
  - **Frieda Steurs** (Dutch Language Institute)
  - Dirk Geeraerts (University of Leuven)
  - Iztok Kosem (Lubljana University)
  - Niels Schiller, Marian Klamer (Leiden University)



Background: Academic Lexicography

- Academic lexicography = **evidence-based** lexicography
  - analysis of large amounts of language data with a theoretical underpinning
  - compilation of concise, yet high-quality knowledge about words
  - for the benefit of the entire language community
- Long tradition going back to the 19th century (WNT, OED, …)
- Early adopters of Information Technology and Data Processing
- … But what does the future hold?

Background: Challenges for the Future

5-day discussion on **3 challenges** for the future of academic lexicography:

1. **Role and place** in society, science and the knowledge economy
2. **Scalability** of both the analysis and production process
3. Accessibility and **customization** for a diverse audience

Written up in a White Paper: Schoonheim, Steurs, Heylen, Vandeghinste (2020) The Future of Academic Lexicography --A White Paper

Background: a White Paper

Strategic **working document** about

1. **specific issues** with respect to the 3 challenges
2. **potential solutions,** their feasibility and impact, in short, middle and longer term;
3. types of **expertise** from different disciplines **necessary** to implement solutions;
4. **formats for collaboration** between experts from the different scientific disciplines

Circulated among workshop participants for feedback, integrated in version 1.2

Overview

- Background of the White Paper
- Challenges for the Future of Academic Lexicography
  - Challenge 1. Role and Place in Society and Science
  - Challenge 2. Genericity vs. Specificity
  - Challenge 3. Scalability: Removing the bottlenecks
- White Paper: Conclusions and Recommendations
- Panel Session: The Future Starts with Post-editing Lexicography
- General Discussion

Overview

- Background of the White Paper
- **Challenges for the Future of Academic Lexicography**
  - **Challenge 1. Role and Place in Society and Science**
  - Challenge 2. Genericity vs. Specificity
  - Challenge 3. Scalability: Removing the bottlenecks
- White Paper: Conclusions and Recommendations
- Panel Session: The Future Starts with Post-editing Lexicography
- General Discussion

Challenge 1. Role and Place in Science and Society

Original 19th c. projects conceived by **academics** for academics (philologists) as **monuments** to the nation's linguistic heritage

In the 21st century, dictionaries are part of the digital linguistic **infrastructure** serving an entire & diverse language **community.**

Language institutes as **public data providers** and **central repositories**:

- Scientific role and funding? multi-disciplinary nexus
- Societal role and funding? public policy target groups

see paper

Challenge 1. Role and Place in Science and Society

Project-based funding ⟺ Structural funding

- **building new** infrastructure and services
- typical funding in academia
- post-project sustainability

- **maintaining** infrastructure and services, esp. data
- guarantee continuity for users
- long-term strategy, stable funding

➔ difficult balance for many language institutes
➔ role description is often still in flux

/instituut voor de
Nederlandse taal/

INT example: Sign Language

INT is involved in Horizon 2020 project SignON

- automated sign language recognition and translation
- improved accessibility and annotation of
  - sign language corpora
  - sign language dictionary
- cooperation with the VGTC, KU Leuven, UGent, Taalunie, VRT...
- Role of INT
  - data collection
  - distribution of data within the project

Challenge 1. Role and Place in Science and Society

Need for **permanent** body/representation at the **European level**

- follow-up after the time-limited ELEXIS project ends
- structurally support joint, cross-disciplinary R&D
- lobby for the importance of lexicographic infrastructure for all languages
- cooperation through further Horizon Europe projects is nice, but not structural
- relation with European Federation of National Institutes of Language (EFNIL)?

Overview

- Background of the White Paper
- **Challenges for the Future of Academic Lexicography**
  - Challenge 1. Role and Place in Society and Science
  - **Challenge 2. Genericity vs. Specificity**
  - Challenge 3. Scalability: Removing the bottlenecks
- White Paper: Conclusions and Recommendations
- Panel Session: The Future Starts with Post-editing Lexicography
- General Discussion

Challenge 2: Genericity vs Specificity

Traditional dictionary projects delivered a **single end-product** with a **stand-alone** content for a specific **user group**

Now, language institutes are **public data providers**:

- **Diverse** end-users and usage scenarios
- Lexicography as linked **infrastructure** for external R&D

Challenge 2: Genericity vs Specificity

Diverse End-Users

Society has become much more diverse - Language users with different

- linguistic backgrounds (immigration, education)
- needs and expectations (professional, educational, leisure)

Ability to customise dictionaries to specific user needs is still limited

- Can academic lexicographic content be customized (automatically)?

Change the project-specific approach towards

- **generic lexicographic databases** designed to be **customizable**
- integrate data analytics for **user and usage modelling** to allow (semi-automatic) customization.

Challenge 2: Genericity vs Specificity

Lexicographic data as an infrastructure

New goals and practices

- no longer building a specific dictionary
- building a central database and infrastructure
  - which is adaptable to specific use cases
  - which is useful for both human users and NLP applications
  - open to external R&D and integration in other applications
- lexicographic knowledge is used for different functions and in different shapes
  - rethinking the lexicographic process as a modular approach

Challenge 2: Genericity vs Specificity

Lexicographic data as an infrastructure

Making data available as an R&D resource

- European Network on e-Lexicography
- ELEXIS project linking dictionaries through Linked Open Data
- Nexus Linguarum European network for Web-centred linguistic data science
- Pret-a-LLOD project exploits combination of LOD and language technology to create ready-to-use multilingual data
- Lexicon Model for Ontologies (LeMOn) provides linguistic grounding for ontologies

/instituut voor de
Nederlandse taal/

INT example: The GiGANT lexicon

All the dictionaries and databases are linked in a central lexicon:

- computational lexicon of Dutch
- from sixth century till now
- words and word groups (incl. named entities)
- every possible variant of spelling and form
- accessible to API's as a 'service'

Challenge 2: Genericity vs Specificity

R&D challenges for **generic, yet customizable** lexicographic infrastructure:

- new lexicographic workflows for
  - generic content creation (independent of specific user groups/usage)
  - customization of the generic content for specific user groups
- user modelling of “distributed” and privacy-protected users
  - a new role for crowdsourcing /citizen science?
- linking heterogenous content (senses, definitions) created in different projects and for different users groups in one integrated database

Overview

- Background of the White Paper
- **Challenges for the Future of Academic Lexicography**
  - Challenge 1. Role and Place in Society and Science
  - Challenge 2. Genericity vs. Specificity
  - **Challenge 3. Scalability: Removing the bottlenecks**
- White Paper: Conclusions and Recommendations
- Panel Session: The Future Starts with Post-editing Lexicography
- General Discussion

Challenge 3. Scalability: Removing the bottlenecks

So far, e-lexicography focussed on integrating **“shallow” statistical corpus analyses** into **lexicographers’** toolset for easily identifying relevant linguistic facts for specific tasks (word sketches, examples).

We enter a **new phase** with **“deep” Artificial Intelligence** being integrated into the lexicographic workflow

for both **data processing** and **content creation** *see paper*

with a **new role for the lexicographer**

Challenge 3. Scalability: Removing the bottlenecks

Lexicographers' role in the **introduction of AI** to the workflow

- Determining the **bottlenecks** in current lexicographic workflow
- **Expert** to be "modelled" and emulated by the AI system
- **Input** to train the AI:
  - not just the end **product** (e.g. dictionary article)
  - the whole content creation **process** needs to made transparent
- Lexicographers as **users** that determine the **usability requirements** for the new AI-enabled workflow

/instituut voor de
Nederlandse taal/

INT example: Observing lexicographers at work

Technologists at INT have set out to perform a detailed observation of the daily work flow of lexicographers:

- how they work right now?
- what bothers them?
- which features they would like?

This led to a report that was discussed with a wider group of technologists and lexicographers:

- return of investment of each of the suggestions?
- highest return treated first

Challenge 3. Scalability: Removing the bottlenecks

Lexicographers' role **within an operational AI-workflow**

If partial automation through deep learning turns out to be possible, what tasks will lexicographers do?

Three potential **role descriptions** :

1. Quality assurance
2. Augmented / hybrid intelligence
3. Manager and mediator

Challenge 3. Scalability: Removing the bottlenecks

Role 1: Quality Assurance

The **AI creates** content and **lexicographers support** and correct:

- post-editing (like translators)
- sample-based quality checks for overall quality assurace
- creating gold standard data for training and testing AI

Challenge 3. Scalability: Removing the bottlenecks

Role 2: Augmented intelligence / hybrid intelligence

- Highly interactive **integration** of artificial intelligence and human expertise
- Lexicographer initiates content creation, and AI suggests an **autocompletion** (cf interactive translation)
- AI presents **difficult cases** to the human expert (active learning)

Challenge 3. Scalability: Removing the bottlenecks

Role 3: Manager and mediator

High level role for lexicographers overseeing the AI workflow

- managing the lexicographic project
- mediating between use cases and AI programming
- helping with translation of lexicographic tasks into a technical solution, (task/requirement specifications)
- managing the crowdsourcing of any post-editing

Challenge 3. Scalability: Removing the bottlenecks

Lexicographers' role **within an operational AI-workflow**

- Three **role** are not mutually exclusive
- New **skills** are required for lexicographers:
  - quality control in automated processes
  - interacting with AI
  - managing and (conceptual) design of AI systems

Overview

- Background of the White Paper
- Challenges for the Future of Academic Lexicography
  - Challenge 1. Role and Place in Society and Science
  - Challenge 2. Genericity vs. Specificity
  - Challenge 3. Scalability: Removing the bottlenecks
- White Paper: Conclusions and Recommendations
- Panel Session: The Future Starts with Post-editing Lexicography
- General Discussion

Overview

- Background of the White Paper
- Challenges for the Future of Academic Lexicography
  - Challenge 1. Role and Place in Society and Science
  - Challenge 2. Genericity vs. Specificity
  - Challenge 3. Scalability: Removing the bottlenecks
- White Paper: Conclusions and Recommendations
- Panel Session: The Future Starts with Post-editing Lexicography
- General Discussion

Conclusions and ...

Lexicography is changing

- from the old school dictionary writing handicraft
- into a more contemporary, efficient and **reusable infrastructure**
- scientific and societal **roles** have changed
- how artificial intelligence and natural language processing can **benefit from lexicographic resources**,
- how building lexicographic resources can **benefit from** the state-of-the-art models and tools developed in **AI and NLP**

Recommendations

1. Strengthening the ties between disciplines
    a. Communication between lexicographers and computational linguists
    b. Cooperation with university research groups in AI and NLP to set up applied lexicographic research
    c. Need for a permanent body at the European level to structurally support digital lexicography
2. Building a generic lexicographic infrastructure instead of writing specific dictionaries
    a. Link existing lexicographic resources through a generic lexicon
    b. Embed the generic lexicon in an international, multilingual context
    c. New, use-case-specific lexicographic resources should reuse information from the generic lexicon, improving consistency and productivity
    d. Quality assurance methodologies are needed for linking of lexicographic resources
3. Realism combined with ambition
    a. Small, realistic improvements can make a huge difference in practical lexicographic work
    b. Automation and crowdsourcing can remove bottlenecks, but require adequate quality assurance methodologies, involving the lexicographer

Overview

- Background of the White Paper
- Challenges for the Future of Academic Lexicography
  - Challenge 1. Role and Place in Society and Science
  - Challenge 2. Genericity vs. Specificity
  - Challenge 3. Scalability: Removing the bottlenecks
- White Paper: Conclusions and Recommendations
- Panel Session: The Future Starts with Post-editing Lexicography
- General Discussion

Overview

- Background of the White Paper
- Challenges for the Future of Academic Lexicography
  - Challenge 1. Role and Place in Society and Science
  - Challenge 2. Genericity vs. Specificity
  - Challenge 3. Scalability: Removing the bottlenecks
- White Paper: Conclusions and Recommendations
- Panel Session: The Future Starts with Post-editing Lexicography
- General Discussion

Panel Session

with
Annette Klosa-Kückelhaus (IDS Mannheim)
Vojtěch Kovář (Masaryk University / Lexical Computing)
Iztok Kosem (University of Ljubljana / Jožef Stefan Institute)

What will the future look like in the brave new world of post-editing lexicography?

Some “provocative” statements.

/instituut voor de
Nederlandse taal/

The Future of the **Lexicographic Workflow**

The gradual **replacement of lexicographers** by an AI will not only make the lexicographic process more **time and cost efficient**, it will also improve the quality of the lexicographic content by **removing the inconsistencies** between individual lexicographers, and by vastly **expanding the amount of empirical data** that can be considered for each entry.

REPLY STATEMENT BY: Annette Klosa-Kückelhaus

REPLY by Annette Klosa-Kückelhaus

IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

The gradual replacement of lexicographers by AI will not only make the lexicographic process more time and cost efficient, it will also improve the quality of the lexicographic content by removing the inconsistencies between individual lexicographers, and by vastly expanding the amount of empirical data that can be considered for each entry.

Can lexicographers be replaced?

- No, as core steps in the lexicographic process can only be done by humans, cf. subdivision of senses, choosing examples illustrating different senses, catching and adequately describing pragmatic restrictions and nuances.
- But intelligent software should support lexicographers (e.g. one-click lexicography).

Will AI make the lexicographic process more time and cost efficient?

- Maybe not: the more automatically compiled data, the more time for proofing is needed to ensure quality of content.
- Expectations of dictionary users regarding reliability of dictionary content should not be neglected.

REPLY by Annette Klosa-Kückelhaus

IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

The gradual replacement of lexicographers by AI will not only make the lexicographic process more time and cost efficient, it will also improve the quality of the lexicographic content by removing the inconsistencies between individual lexicographers, and by vastly expanding the amount of empirical data that can be considered for each entry.

Will AI improve the quality of lexicographic content by removing inconsistencies?

- No, as the lexicon itself is full of inconsistencies (e.g., there is hardly any true synonymy).
- But intelligent programs should support lexicographers in writing the entries (e.g., cross-referencing between entries).

Will AI improve the quality of lexicographic content by vastly expanding the amount of empirical data?

- Not necessarily: Experienced lexicographers can quickly disambiguate senses, understand and define meanings and notice whether there are usage restrictions etc. based on reasonably small numbers of corpus citations (ideally pre-selected automatically).

The Future of **Lexicographic Crowdsourcing**

The generation of high-quality lexicographic content by an Artificial Intelligence will make it possible to **crowdsource the post-editing of all tasks** in the lexicographic process, even definition writing, with **minimal oversight** by lexicographers. Additionally, input from crowdsourcing can teach the AI to highly **customize lexicographic content** to specific user groups in a way that lexicographers cannot.

REPLY STATEMENT BY: Iztok Kosem

Reply by: Iztok Kosem

eLex 2021

Can lexicography even survive without crowdsourcing?

Reply by: Iztok Kosem

eLex 2021

Crowdsourcing myth

experts vs non-experts

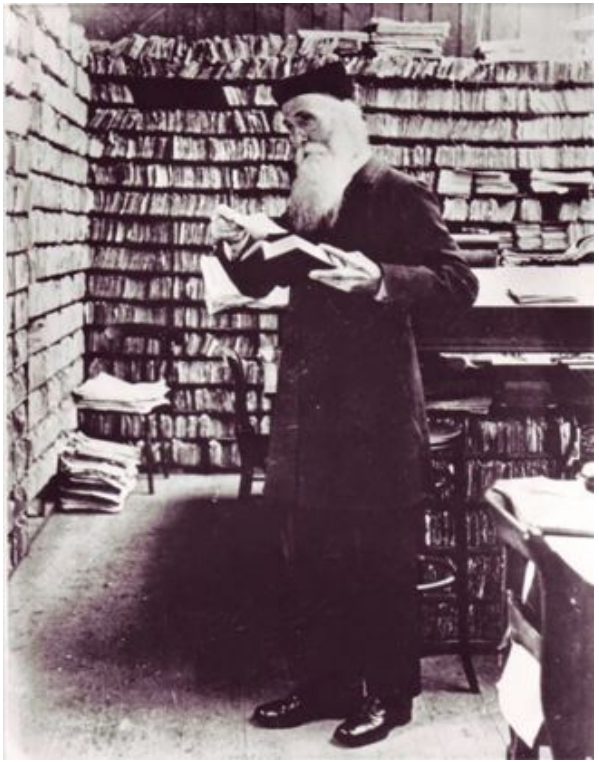
(lexicographers) (everybody else?)

Non-experts to the rescue!

19th century

OED Reading programme:

public collected quotations on word usage



Reply by: Iztok Kosem

eLex 2021

Updating the dictionary + promotion (with control)

Collins Dictionary Thesaurus

Latest New Word Suggestions

The latest words suggested for inclusion in the Collins English Dictionary.
Suggest your word.

Search word suggestions

Search

wanderwort

Submitted bydadge on 4 Jun 2021 Pending Investigation
"(linguistics) a word that has spread through several cultures, often in different forms"

n-gram

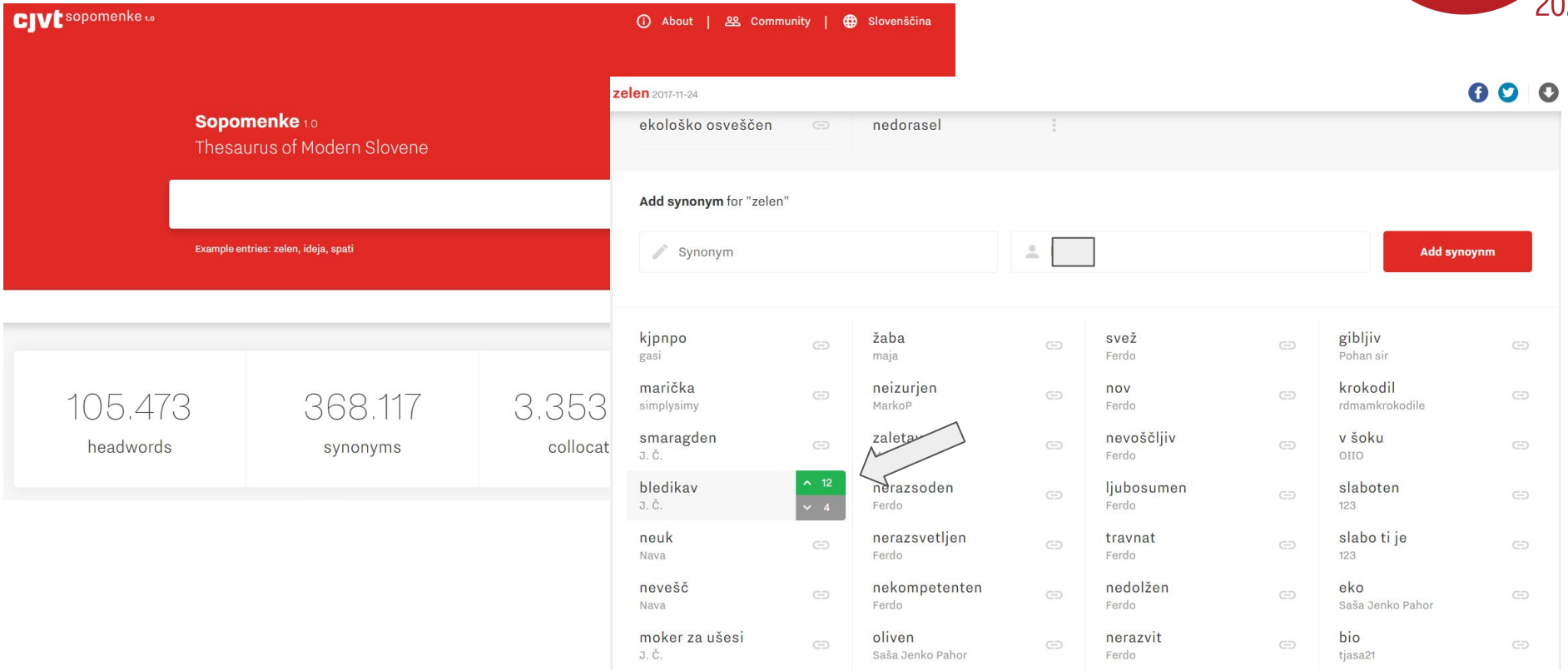
Submitted byplainname on 4 Jun 2021 Pending Investigation
"(in computational linguistics) a sequence of words, or a model of the probability of a word occurring in a sequence"

stewardship

Submitted byplainname on 4 Jun 2021 Published
"the act or role of looking after something (wider sense, not just management or administration)"

Reply by: Iztok Kosem

Updating the dictionary (community control)



Reply by: Iztok Kosem

eLex 2021

Under-resourced languages:

- No lexical resources without crowdsourcing

Living dictionaries: An Electronic Lexicography Tool for Community Activists

(Gregory D. S. Anderson, Anna Luisa Daigneault)

(eLex 2021 panel on Tools)

...to crowdsource the post-editing of all tasks in the lexicographic process, **even definition writing**, with minimal oversight by lexicographers.

- this is already happening
- two methodological variables:
  - level of automation of lexicographic content
  - innovativeness and ability in preparing tasks for crowdsourcers

Reply by: Iztok Kosem

eLex 2021

...input from crowdsourcing can learn the AI to **highly customize lexicographic content** to specific user groups in a way that lexicographers cannot.

- Customisation is not possible without knowing the needs and preferences of the users!
- In such cases in e-lexicography, lexicographers are non-experts!

/instituut voor de
Nederlandse taal/

The Future of **Post-Editing Lexicography**

Whereas translation is a linear process that only *applies* linguistic knowledge to one data-point at the time, lexicography is a synthetic process that *creates* new, high-level linguistic knowledge out of many data-points. The post-editing framework from translation is therefore not applicable in lexicography, and a real one-click dictionary is impossible.

REPLY STATEMENT BY: Vojtěch Kovář

Reply by Vojtěch Kovář

- I partly agree with the statement, but
- the "one-point vs. many-points" and "applies vs. creates linguistic information" descriptions are over-simplified description of the situation on both sides -- translation is in fact a really complex process, vs. creating a dictionary entry can be decomposed into more simpler tasks
- however, the two tasks are really different, so the translation framework cannot be just transferred without significant changes (this is where I agree with the statement)
- but that does not mean at all that the post-editing lexicography is not possible, just that the process needs to be carefully thought-out and tested to fit well into the process of dictionary creation

Overview

- Background of the White Paper
- Challenges for the Future of Academic Lexicography
  - Challenge 1. Role and Place in Society and Science
  - Challenge 2. Genericity vs. Specificity
  - Challenge 3. Scalability: Removing the bottlenecks
- White Paper: Conclusions and Recommendations
- Panel Session: The Future Starts with Post-editing Lexicography
- General Discussion

The Future of **Lexicographic Data**

The ability of Deep Learning systems to capture semantics are advancing so quickly that **lexicographic data are becoming increasingly irrelevant** as a source of semantic knowledge for R&D in NLP. The relation between NLP and Lexicography has become a **one-way street**: Deep Learning NLP can provide lexicography with automatically created content for post-editing, but lexicographic data has nothing to offer to NLP.

/instituut voor de
Nederlandse taal/

The Future of **Lexicographic Institutes**

The lexicographic institute of the future will **only consist of IT staff and lexicographic project managers** that mediate between content generated by a lexicographic AI and post-editing done by the crowd.

/instituut voor de Nederlandse taal/

Thank you!

/instituut voor de Nederlandse taal/

Steurs F. , T. Schoonheim, K. Heylen and V. Vandeghinste (2020). The Future of Academic Lexicography -- A White Paper. Version 1.2. Available at www.ivdnt.org