Development of HAMOD: a High Agreement Multi-lingual Outlier Detection dataset

Miloš Jakubíček, **Emma Romani**, Pavel Rychlý, Ondřej Herman

Natural Language Processing Centre, Faculty of Informatics, Masaryk University | Brno, Czechia
Lexical Computing | Brno, Czechia
Università degli Studi di Pavia, Faculty of Humanities | Pavia, Italy

RASLAN 2021: Recent Advances in Slavonic Natural Language Processing

Outline of the Presentation

- Introduction and motivation
- The word sketch-based thesaurus
- Thesaurus built from word embeddings
- Building HAMOD
- Evaluation
- Conclusions and future development

Introduction and motivation

HAMOD: High Agreement Multi-lingual Outlier Detection

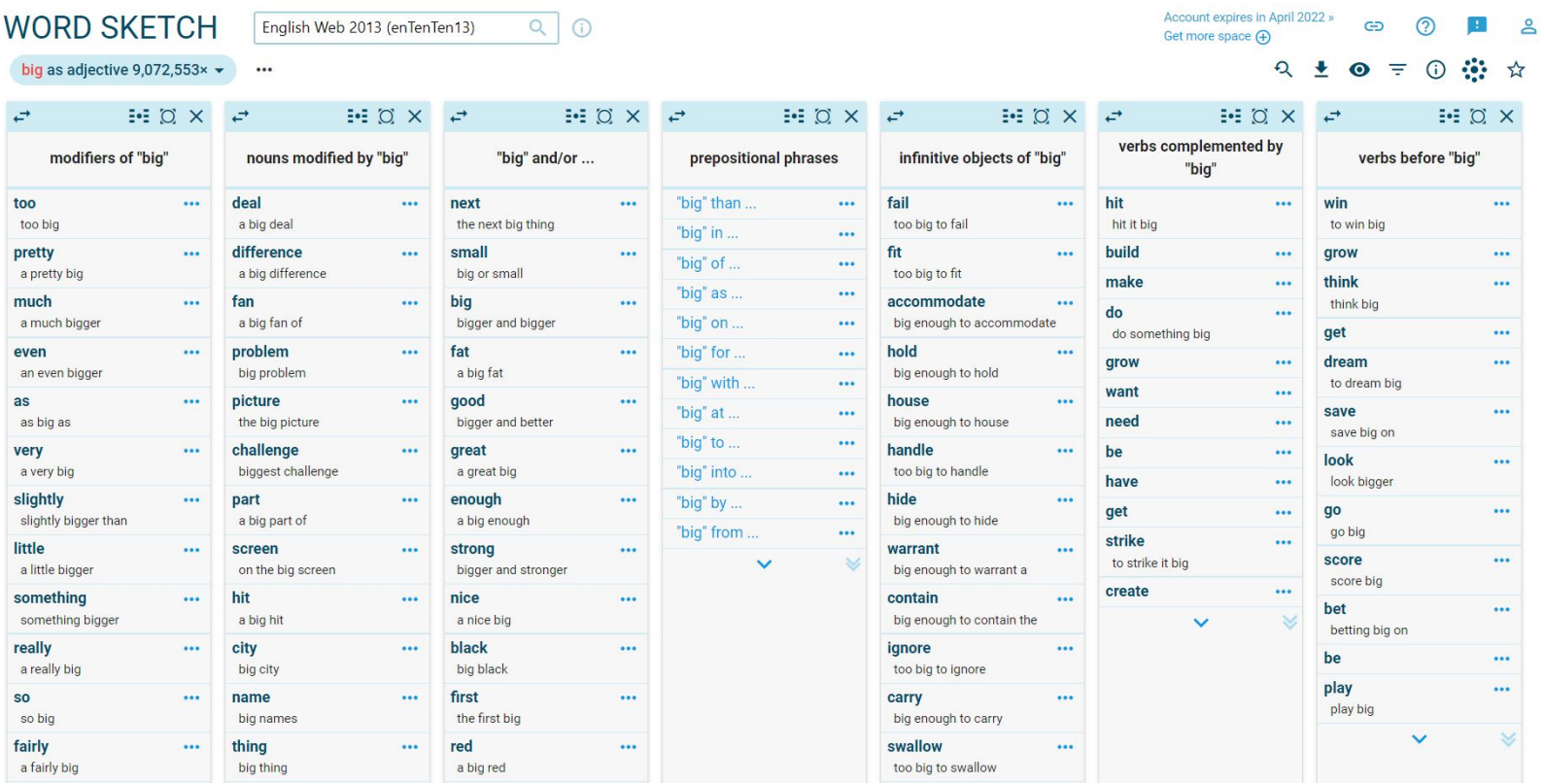
- creating of a dataset for exercising the **outlier detection task** that aims at **high inter-annotator agreement**
  - **outlier detection ➜** out of a set of words, which one is the word that “does not fit” to the others (= outlier)?
  - reliable method compared to other **intrinsic evaluation** methods (e.g., similarity judgments ➜ extremely low inter-annotator agreement)
- **evaluating** of **automatic distributional thesauri** with outlier detection
  - **thesaurus ➜** list of synonyms or words belonging to the same category (semantic field)
    - word sketch-based thesauri
    - word embeddings

Sketch Engine and the word sketch-based thesaurus

Word Sketch

- **word sketch** ➜ short summary of a **word's collocational behaviour** from the perspective of **individual grammatical relations** (noun's modifier, verb's subject etc.)
  - 1 word sketch: headword - grammatical relation - collocate
  - dependency syntax graph calculated using hybrid rule-based and statistical approach
  - **word sketch grammar** ➜ selects syntactically viable collocation candidates using CQL over morphological annotation
  - statistical scoring using word association score (**LogDice**: scalable association metric)

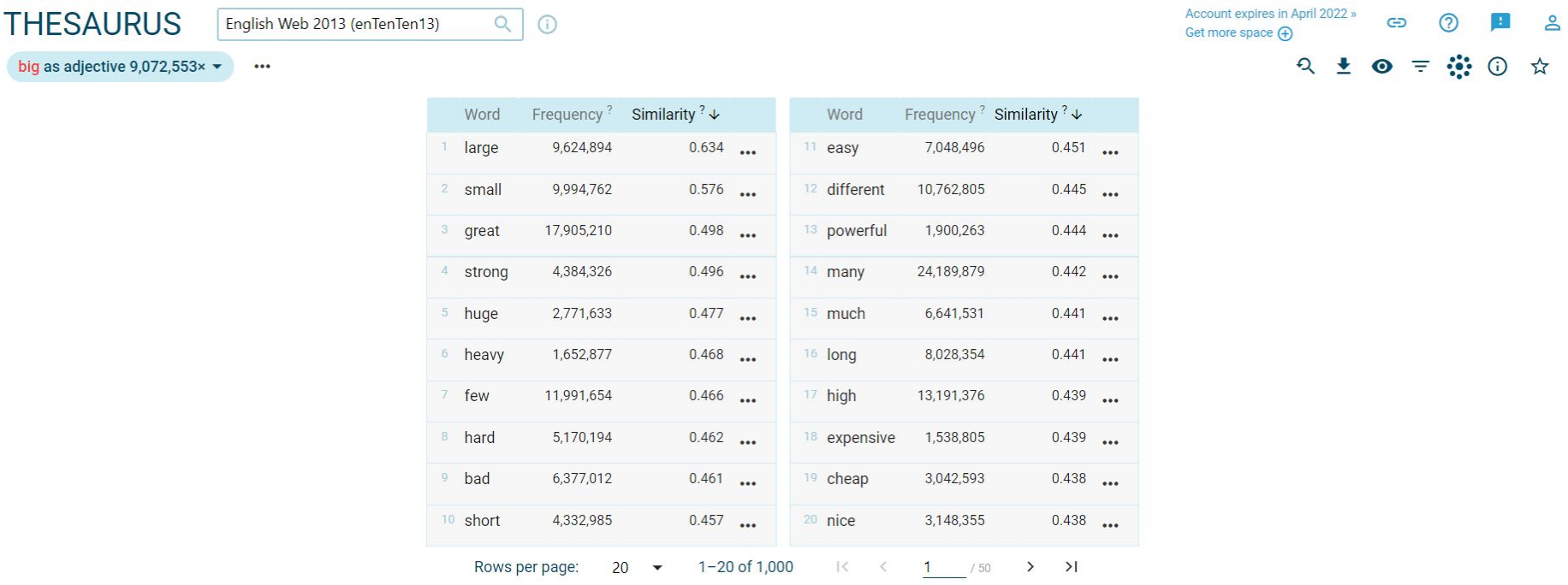
Word Sketch


WORD SKETCH
English Web 2013 (enTenTen13)
Account expires in April 2022 »
Get more space
big as adjective 9,072,553×
modifiers of "big"
too
too big
pretty
a pretty big
much
a much bigger
even
an even bigger
as
as big as
very
a very big
slightly
slightly bigger than
little
a little bigger
something
something bigger
really
a really big
so
so big
fairly
a fairly big
nouns modified by "big"
deal
a big deal
difference
a big difference
fan
a big fan of
problem
big problem
picture
the big picture
challenge
biggest challenge
part
a big part of
screen
on the big screen
hit
a big hit
city
big city
name
big names
thing
big thing
"big" and/or ...
next
the next big thing
small
big or small
big
bigger and bigger
fat
a big fat
good
bigger and better
great
a great big
enough
a big enough
strong
bigger and stronger
nice
a nice big
black
big black
first
the first big
red
a big red
prepositional phrases
"big" than ...
"big" in ...
"big" of ...
"big" as ...
"big" on ...
"big" for ...
"big" with ...
"big" at ...
"big" to ...
"big" into ...
"big" by ...
"big" from ...
infinitive objects of "big"
fail
too big to fail
fit
too big to fit
accommodate
big enough to accommodate
hold
big enough to hold
house
big enough to house
handle
too big to handle
hide
big enough to hide
warrant
big enough to warrant a
contain
big enough to contain the
ignore
too big to ignore
carry
big enough to carry
swallow
too big to swallow
verbs complemented by "big"
hit
hit it big
build
make
do
do something big
grow
want
need
be
have
get
strike
to strike it big
create
verbs before "big"
win
to win big
grow
think
think big
get
dream
to dream big
save
save big on
look
look bigger
go
go big
score
score big
bet
betting big on
be
play
play big

Word Sketch-based thesaurus

- **automatic derivation of distributional thesaurus** by calculating similarity of word sketch contexts ➜ for each word, which words share most collocates in the same grammatical relation

Word Sketch-based thesaurus



	Word	Frequency	Similarity		Word	Frequency	Similarity	
1	large	9,624,894	0.634	11	easy	7,048,496	0.451	
2	small	9,994,762	0.576	12	different	10,762,805	0.445	
3	great	17,905,210	0.498	13	powerful	1,900,263	0.444	
4	strong	4,384,326	0.496	14	many	24,189,879	0.442	
5	huge	2,771,633	0.477	15	much	6,641,531	0.441	
6	heavy	1,652,877	0.468	16	long	8,028,354	0.441	
7	few	11,991,654	0.466	17	high	13,191,376	0.439	
8	hard	5,170,194	0.462	18	expensive	1,538,805	0.439	
9	bad	6,377,012	0.461	19	cheap	3,042,593	0.438	
10	short	4,332,985	0.457	20	nice	3,148,355	0.438	

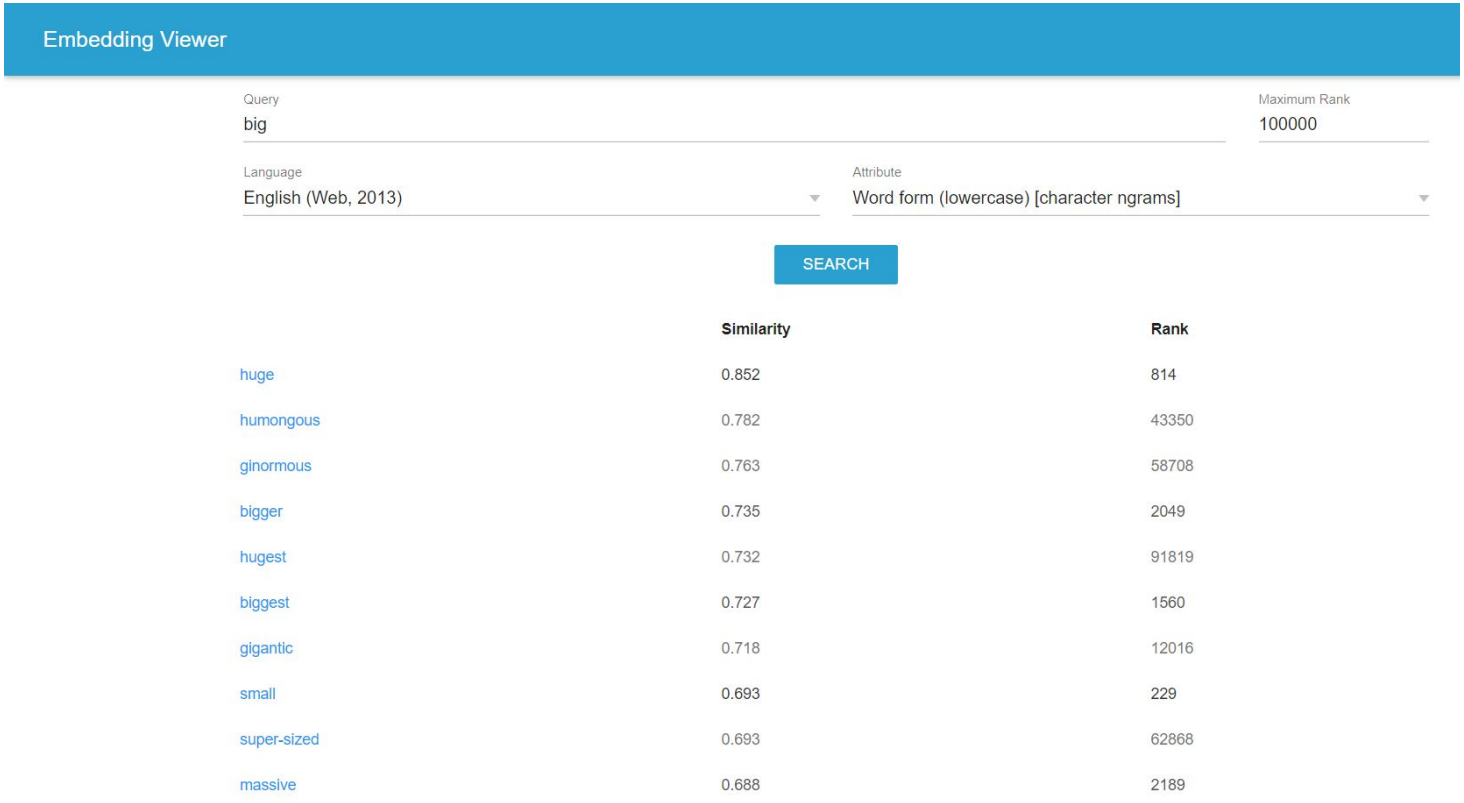
Rows per page: 20 1–20 of 1,000 1 / 50

Thesaurus built from word embeddings

Distributional thesaurus

- calculating the **vector representation** for each word in a corpus (= word embedding)
- using distances between individual vectors as **measure of words' (dis)similarity**
  - FastText
  - Word2vec
- based on corpora in Sketch Engine ➜ no need for part-of-speech tagging and lemmatization

Distributional thesaurus



	Similarity	Rank
huge	0.852	814
humongous	0.782	43350
ginormous	0.763	58708
bigger	0.735	2049
hugest	0.732	91819
biggest	0.727	1560
gigantic	0.718	12016
small	0.693	229
super-sized	0.693	62868
massive	0.688	2189

Building HAMOD

Dataset construction

- current **languages**: English, Czech, Slovak, Estonian, French, German, Italian
- source dataset: English ➜ **translation/adaptation** of the dataset to the other languages
  - avoid ambiguities
  - comparable (not parallel) datasets
- 1 set:
  - 8 **inliers** ➜ words that are part of a **semantic category** or together define a **topic**
    - examples: musical instruments, means of transport, fruit trees, parts of head, sport verbs
  - 8 **outliers** ➜ words that do not belong to the category because they **lack some relevant properties**

Outlier detection exercise

- each human **evaluator** goes through all the sets (only once) for their **native language**
- 1 exercise: **8 inliers** + **1 outlier** (randomly chosen from the list of outliers for each set)
- in each turn, the evaluator selects the **outlier**
- simple web interface for the exercise

Evaluation

Inter-Annotator Agreement

- currently computed for **Czech** and **Estonian**: **< 90%** of absolute raw agreement
- **successful run**: an exercise where all sets were correctly fulfilled by an evaluator

Language	Success runs	All runs	Agreement
Czech	2,082	2,150	**0.97**
Estonian	3,285	3,525	**0.93**

Evaluation of distributional thesauri

- **overall Accuracy** (Acc: the outlier was correctly identified?)
- **Outlier Position Percentage** (OPP: average percentage of the right answer)

Corpus	Corpus size	Dataset size	SkE Acc	SkE OPP	Word2Vec Acc	Word2Vec OPP
czTenTen12	5G	232	0.573	0.898	0.655	0.871
enTenTen13	22G	296	0.456	0.847	0.655	0.873
EstonianNC 17	1.3G	296	0.564	0.832	0.547	0.784
deTenTen13	19G	232	0.349	0.798	0.323	0.764
frTenTen12	6.8G	232	0.276	0.744	0.427	0.768
skTenTen11	0.6G	296	0.389	0.777	0.591	0.851
itTenTen16	5.8G	296	0.453	0.856	0.581	0.869

Conclusions and future development

Future development

- improvement of the dataset for further **development**, **evaluation** and **comparison** of distributional thesauri
  - **extension** of the dataset: **100 exercise dataset**
  - covering of **more languages** (EU at first)
- **monitoring of IAA** and **adjustment** of the dataset accordingly ➜ **high IAA**
- maintaining the discriminative power of the dataset ➜ **ability to discover differences between individual thesaurus types** (to be revisited in case it is lost)
- optimizing distributional thesauri

thank you for listening!