A Distributional Multi-word Thesaurus in Sketch Engine

Miloš Jakubíček, Pavel Rychlý

Lexical Computing
jak@fi.muni.cz
milos.jakubicek@sketchengine.eu

NLP Centre, Masaryk University
pary@fi.muni.cz
pavel.rychlyk@sketchengine.eu

RASLAN 2019, 7. 12. 2019

Outline

Motivation

- word sketch and thesaurus in Sketch Engine
- word sketch: both single words and multi-words
- thesaurus: only single words so far $\Rightarrow$ extension to multi-words coming

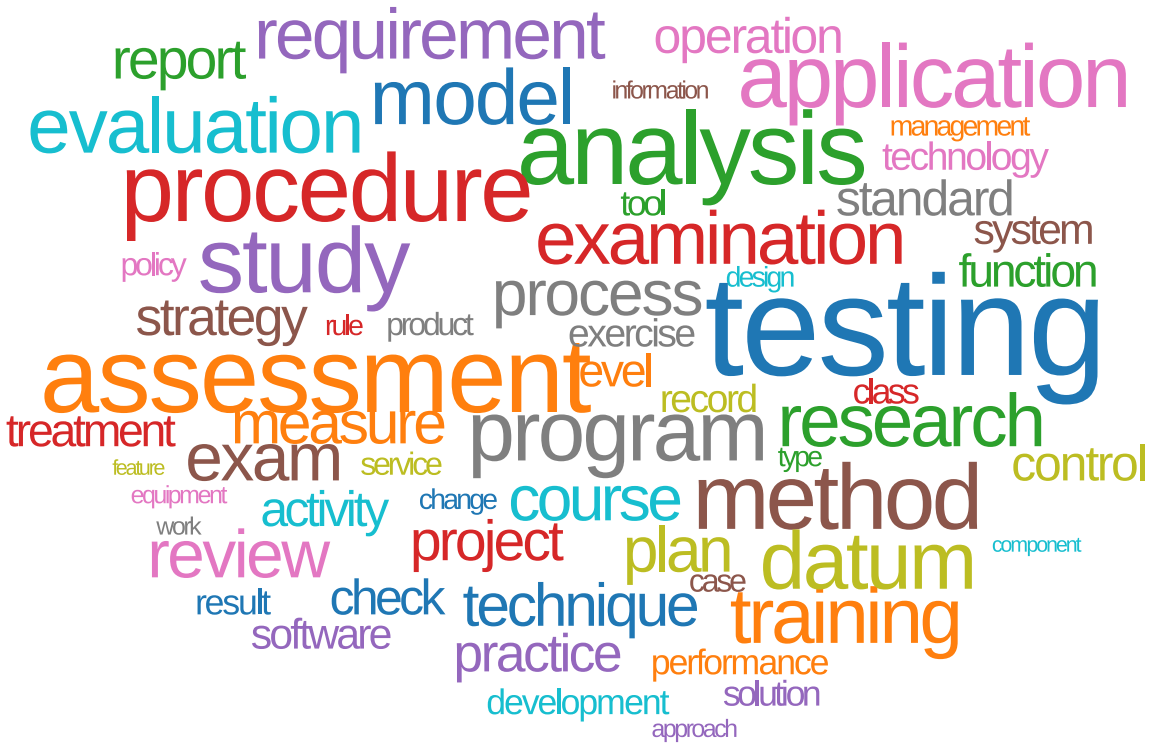
Word sketch

modifiers of "resource"	nouns modified by "resource"	verbs with "resource" as object	verbs with "resource" as subject
natural ••• natural resources	**management** ••• resource management	**allocate** •••	**include** ••• resources including
human ••• human resources	**allocation** ••• resource allocation	**manage** •••	**permit** ••• resources permit
water ••• water resources	**center** ••• Resource Center	**access** •••	**exist** ••• resources exist
limited ••• limited resources	**centre** ••• Resource Centre	**provide** •••	**provide** ••• resource provides
valuable ••• a valuable resource	**efficiency** ••• resource efficiency	**utilize** •••	**relate** ••• resources relating to
online ••• online resources	**conservation** ••• resource conservation	**conserve** •••	**cover** ••• resources covering
financial ••• financial resources	**guide** ••• Resource Guide	**share** •••	**allow** ••• resources allow
		pool •••	
		invest •••	
		need ••• resources needed to	

Thesaurus

test *(noun)* Alternative PoS: verb (freq: 941,372)
enTenTen [2012] freq = 1,915,482 (147.70 per million)

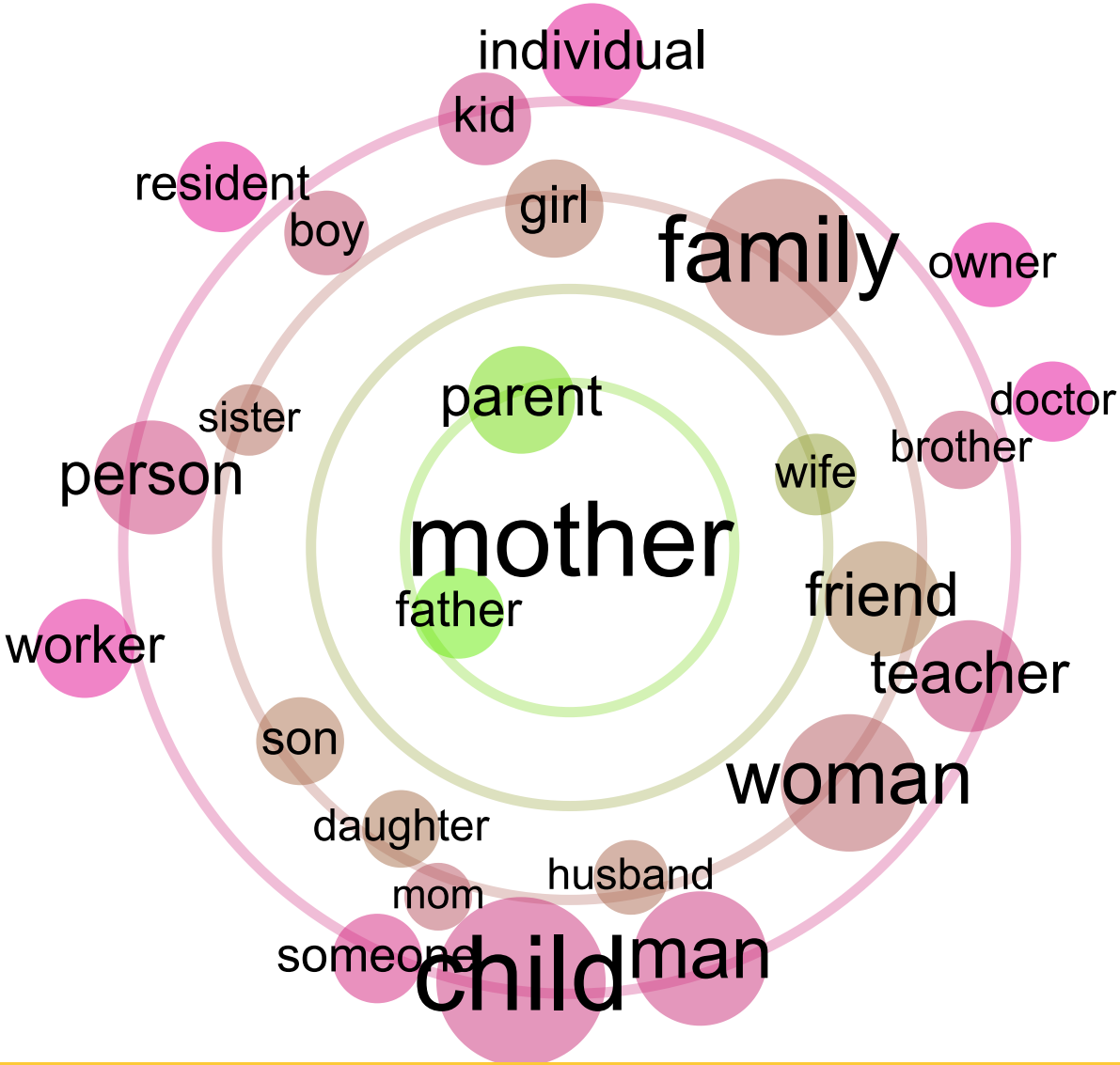
Lemma	Score	Freq
testing	0.520	558,727
assessment	0.410	640,347
analysis	0.399	1,196,660
procedure	0.382	1,311,372
study	0.380	3,090,402
method	0.373	2,760,051
application	0.366	3,171,582
program	0.365	6,442,955
datum	0.362	3,165,540
evaluation	0.360	468,130
model	0.357	2,557,538
training	0.354	2,486,409
research	0.354	3,171,715
examination	0.352	375,991
requirement	0.349	1,734,482
exam	0.349	373,769
review	0.348	1,803,362



Thesaurus


brother
employee
soldier
adult
worker
group
anyone
father
lady
leader
player
citizen
owner
writer
kid
consumer
other
baby
female
person
one
woman
dad
member
student
guy
patient
officer
man
husband
mom
girl
son
victim
someone
driver
client
's
spouse
parent
people
dog
child
sister
boy
resident
customer
family
friend
wife
doctor
couple
teacher
individual
daughter
folk
everyone
partner
lover
neighbor

Thesaurus


individual
kid
resident
girl
boy
family
owner
parent
doctor
sister
brother
person
wife
mother
friend
father
worker
teacher
son
woman
daughter
husband
mom
someone
child
man

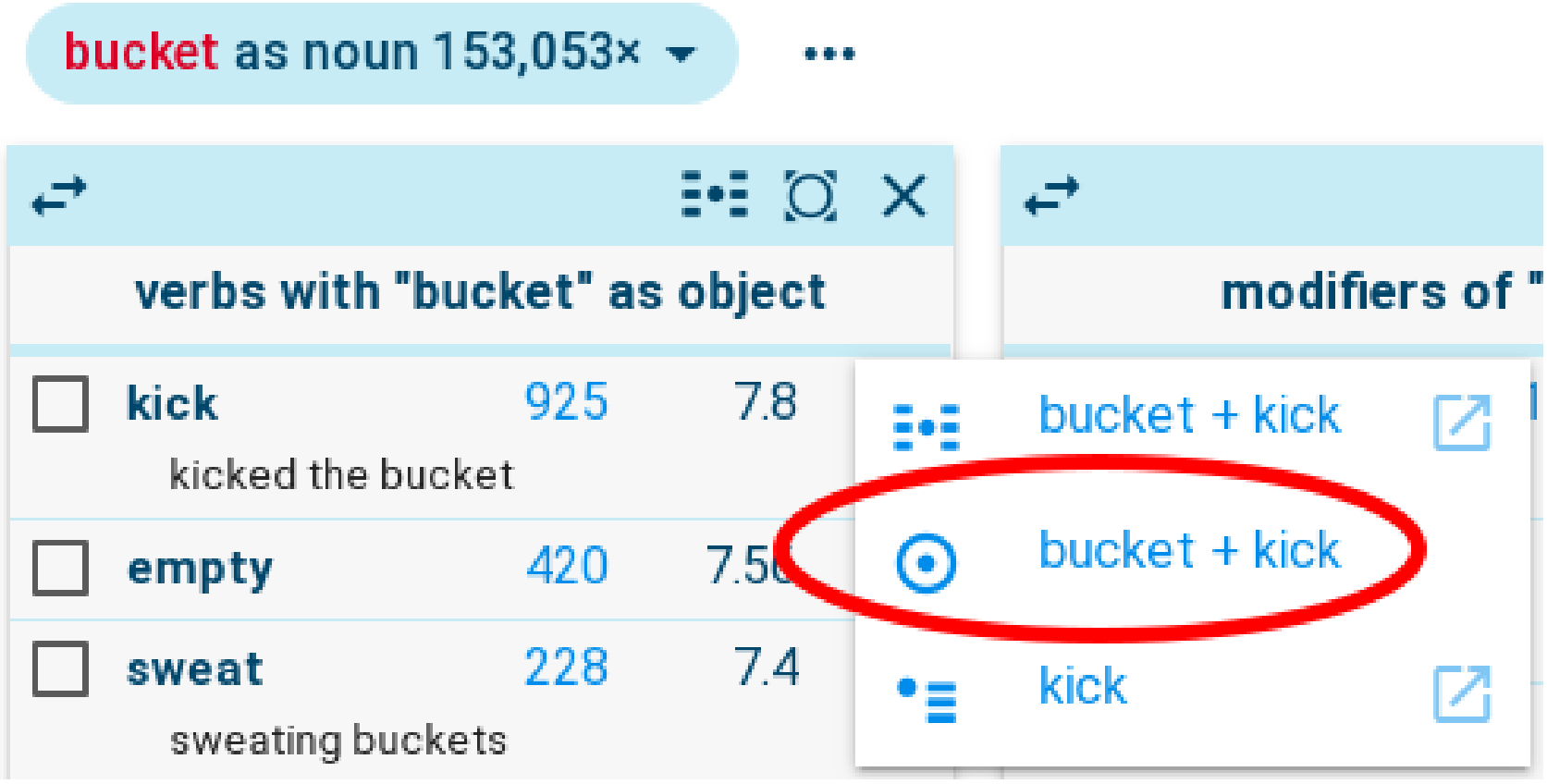
$$Dist(w_1, w_2) = \frac{\sum_{(r,c) \in ctx(w_1) \cap ctx(w_2)} AS_{(w_1,r,c)} + AS_{(w_2,r,c)} - (AS_{(w_1,r,c)} - AS_{(w_2,r,c)})^2/50}{\sum_{i \in ws_1} AS_i + \sum_{i \in ws_2} AS_i}$$

The term $(AS_i - AS_j)^2/50$ is subtracted in order to give less weight to shared triples, where the triple is far more salient with w_1 than w_2 or vice versa. We find that this contributes to more readily interpretable results, where words of similar frequency are more often identified as near neighbours of each other.

Extensions to multi-words

The most difficult thing is to say what a multi-word is.
The rest is basically straighforward.

Multi-word sketches



Multi-word sketches

<s> I think all of Drogheda was listening to the radio when the draw was being made and I thought **somebody** was **pulling** my **leg** when we were drawn against them. </s>

<s> "I thought **somebody** was **pulling** my **leg** , but one morning, just after it snowed, I saw footprints going to the tower and I thought it was one of the guys, but I went in there to do some work and nobody was there. </s>

<s> "Both Andy and I were taken aback when we first received the initial email from him, to the extent that we thought it was **somebody** **pulling** our **leg** . </s>

<s> She has no problem if **somebody** **pulls** her **leg** and she will give right back to you. </s>

<s> Ok, **somebody** 's **pulling** my **leg** , right?! </s>

Multi-word sketches

WORD SKETCH

English Web 2015 (enTenTen15)

	WS collocations	Frequency	Score
1	pull-v particles after "%w" with object out-x	127,030	9.25
2	pull-v particles after "%w" out-x	65,648	7.66
3	pull-v particles after "%w" with object off-x	57,617	9.57
4	pull-v pronominal subjects of "%w" he-d	48,229	5.29
5	pull-v pronominal objects of "%w" it-d	48,039	5.93
6	pull-v pronominal subjects of "%w" i-d	47,720	4.32
7	pull-v particles after "%w" with object up-x	42,505	7.6

Multi-word thesaurus

- salient multi-word sketches == multi-words
- same scoring

Table: Thesaurus items for the phrase "kohl helmut chancellor" on the BNC.

score	frequency	item
1.00	755	kohl chancellor helmut
0.88	1790	kohl helmut
0.75	7307	kohl chancellor
0.34	606	kohl
0.20	140	mitterrand president
0.18	536	chancellor kohl
0.17	340	bush president us
0.17	153	bush us president
0.17	20	re-unification
0.17	116	bush us
0.16	370	bush george president
0.16	283	bush president george
0.16	116	clinton president

Conclusions

- initial approach complete, to be implemented within Sketch Engine
- need to disregard ordering of multi-word items
- evaluation and comparison (as in preceding talk)