Precomputed Word Embeddings for 15+ Languages

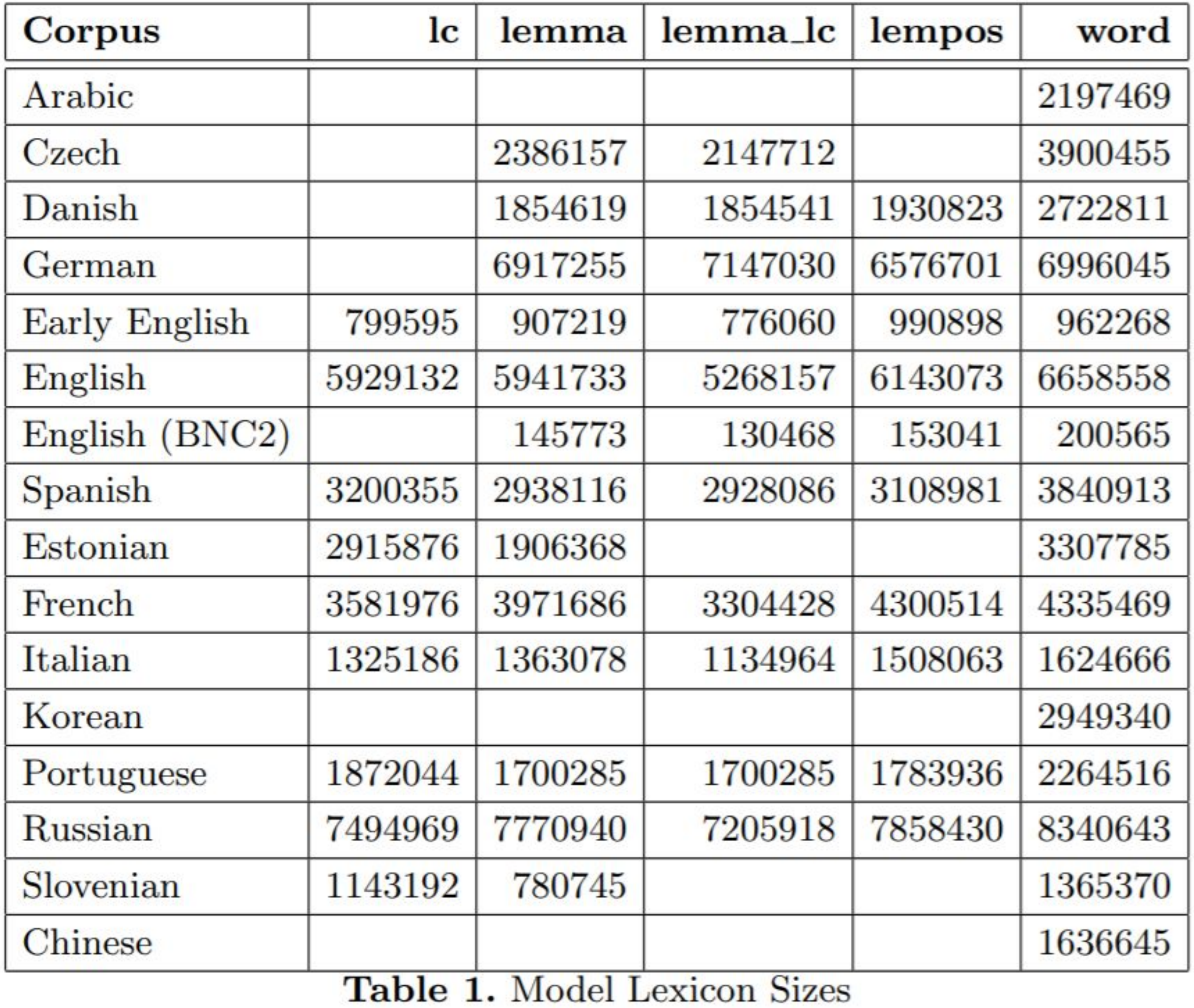
Ondřej Herman
FI MUNI, Lexical Computing

Word Embeddings

- Project words onto a vector space and keep interesting properties
    - Easy to process
- Useful for many downstream NLP tasks

Precomputed Word Embeddings

- Calculated using a modified version of fastText
  - Can read corpora indexed by manatee
- Available for 15 languages
  - Multiple attributes for every language
- We want to have a model for every language in Sketch Engine eventually



Corpus	lc	lemma	lemma_lc	lempos	word
Arabic					2197469
Czech		2386157	2147712		3900455
Danish		1854619	1854541	1930823	2722811
German		6917255	7147030	6576701	6996045
Early English	799595	907219	776060	990898	962268
English	5929132	5941733	5268157	6143073	6658558
English (BNC2)		145773	130468	153041	200565
Spanish	3200355	2938116	2928086	3108981	3840913
Estonian	2915876	1906368			3307785
French	3581976	3971686	3304428	4300514	4335469
Italian	1325186	1363078	1134964	1508063	1624666
Korean					2949340
Portuguese	1872044	1700285	1700285	1783936	2264516
Russian	7494969	7770940	7205918	7858430	8340643
Slovenian	1143192	780745			1365370
Chinese					1636645

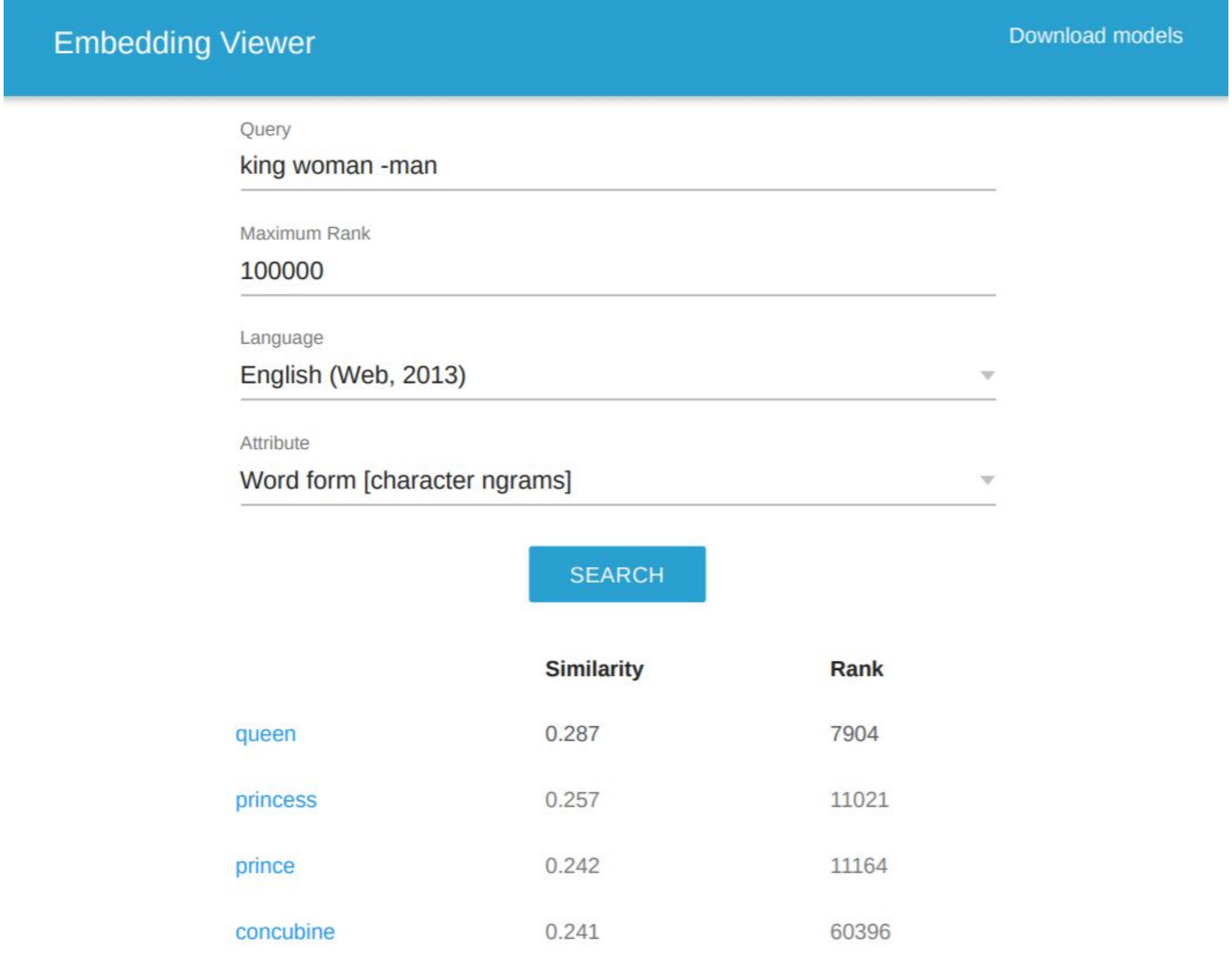
Table 1. Model Lexicon Sizes

Precomputed Word Embeddings

- Available for download at https://embeddings.sketchengine.eu/
- Licensed under the terms of the *Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License*
    - Mention the source
    - For non-commercial uses
    - Derivative works can be shared under the same terms

Embedding Viewer

- The models can be queried online at https://embeddings.sketchengine.eu/
- Similarity queries and vector algebra are supported
- ~5 pages of Python, depends only on NumPy



Query
king woman -man

Maximum Rank
100000

Language
English (Web, 2013)

Attribute
Word form [character ngrams]

	Similarity	Rank
queen	0.287	7904
princess	0.257	11021
prince	0.242	11164
concubine	0.241	60396

Embedding Viewer API

- The results can be obtained through a REST API
  - As JSON
  - As tab-separated columnar data

```
$ curl ’https://embeddings.sketchengine.eu/?q=dog&lim=100000&n=5&
       model=English+%28Web%2C+2013%29%7CLemma&raw’
```

```
    puppy  0.8980982303619385 4139
    cat    0.8976492285728455 1678
    canine 0.8802799582481384 8694
    pup    0.8700659275054932 9166
    pet    0.8562509417533875 1622
```

```
$ curl ’https://embeddings.sketchengine.eu/?q=cat&lim=100000&n=5&
       model=English+%28Web%2C+2013%29%7CLemma&json’
```

```
{"w":[
    ["dog", 0.8976492881774902, 685],
    ["kitten", 0.8868610858917236, 8330],
    ["feline", 0.8669211864471436, 15259],
    ["pet", 0.8627837896347046, 1622],
    ["chinchilla", 0.847865283489227З, 51731]]
}
```

Parameter	Description
q=QUERY	a complete query formatted as described above
pos=WORD	a single query word, can be specified multiple times
neg=WORD	a single query word complement, can be specified multiple times
pos_vec=VEC	same as pos, but interpreted as a comma-separated vector
neg_vec=VEC	same as neg, but interpreted as a comma-separated vector
n=N	the amount of rows to be returned
lim=N	maximum rank of the result entries
model=NAME	name of the embedding model
json	format the result as JSON
raw	format the result as TSV (tab-separated columnar format)
vec	include the word vectors in the result

Table 2. Embedding API Query Parameters

Conclusion

- Multiple embedding models are available through https://embeddings.sketchengine.eu/
    - For download
    - Through API
    - Through Web UI