Automated Semantic Analysis of Danish

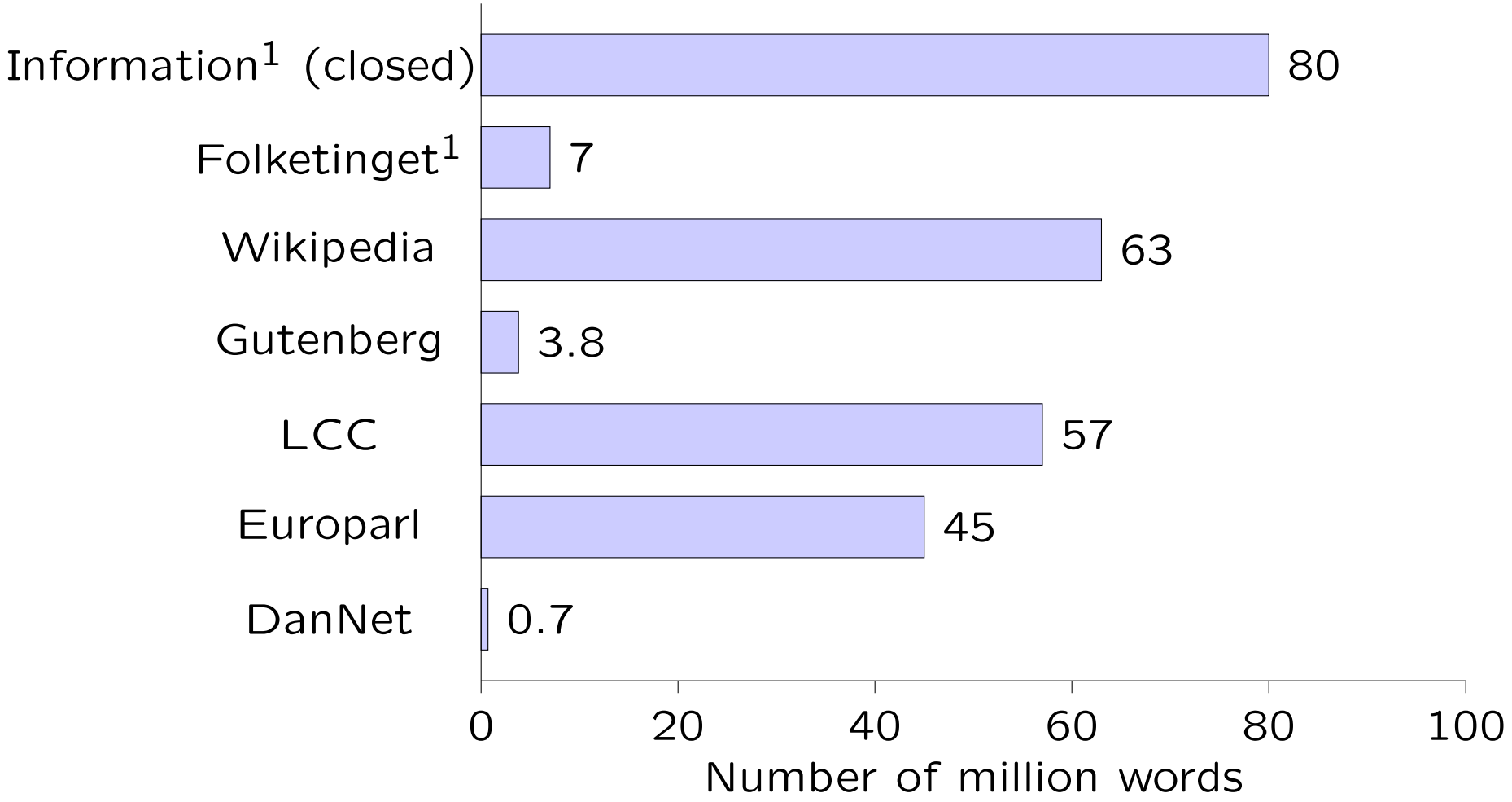
Finn Årup Nielsen

Cognitive Systems, DTU Compute, Technical University of Denmark

31 August 2018

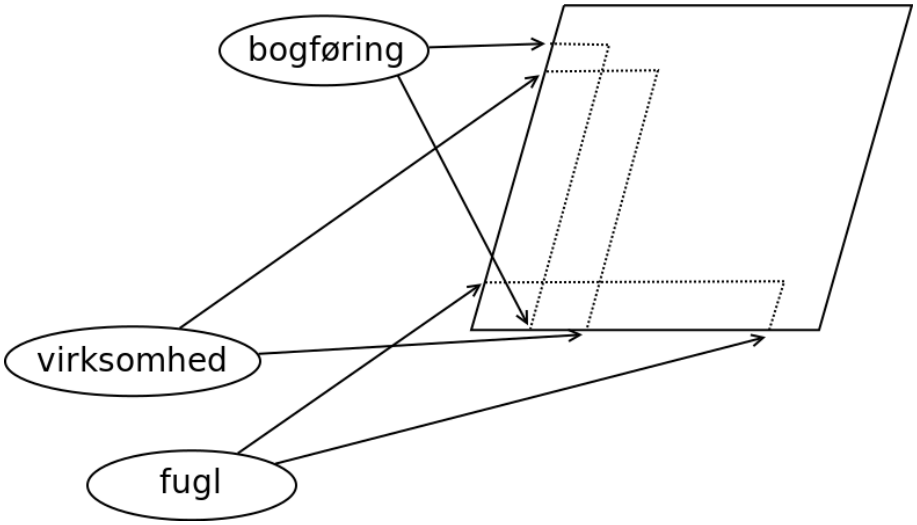
Semantics from corpora

Danish corpora size wrt. words



[1] According to https://visl.sdu.dk/corpus_linguistics.html

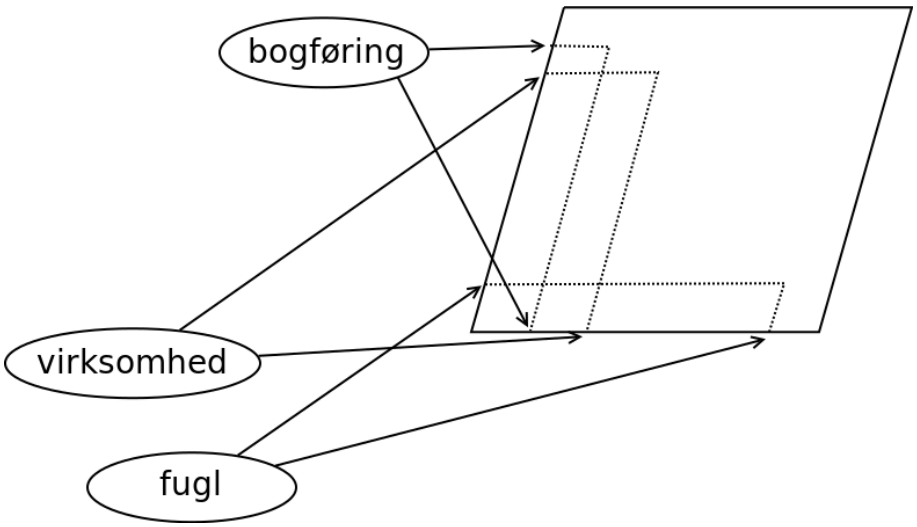
Other Danish embeddings



Dasem Word2vec model (243416 × 100) 90 MB (compressed). Not distributed (yet), but corpus handling and training available in Dasem at https://github.com/fnielsen/dasem. The word analogy gives: kvinde + konge - mand = monark, adel, tronfølger, . . .

Thomas Egense, *Word2Vec dictionary for 30 million Danish newspaper pages*, distributed from LOAR Repository. (2404836 × 300) 6.4 GB: The word analogy gives: kvinde + konge - mand = konqe, konges, dronning, . . .

FastText



fastText at https://fasttext.cc from Facebook is a standalone program and associated Python wrappers for word embedding.

Also uses character n-grams, — probably good for compound- and morphological-rich language.

For Danish, there are two different pre-trained models:

wiki-da (312'956 $\times$ 300) based on Danish Wikipedia (Bojanowski et al., 2016)

cc.da.300 (2'000'000 $\times$ 300) based on *Common Crawl* and Danish Wikipedia (Grave et al., 2018)

Evaluations

Word intrusion evaluation

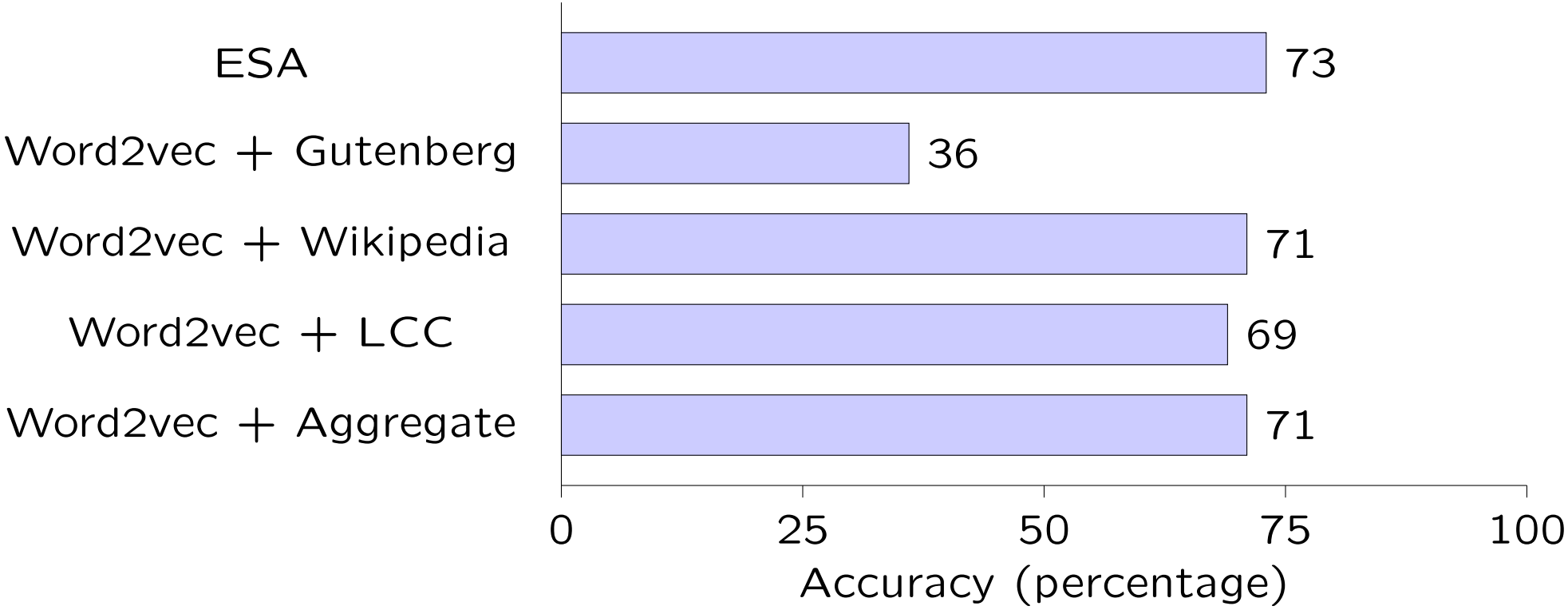
Detection of the odd-one-out with different semantic models.

word1	word2	word3	(outlier) word4	ESA	Word2vec Gutenberg	Word2vec LCC	Word2vec Wikipedia	Word2vec Aggregate
æble (apple)	pære (pear)	kirsebær (cherry)	stol (chair)	stol	stol	stol	stol	stol
stol (chair)	bord (table)	reol (shelves)	græs (grass)	græs	stol	bord	reol	bord
græs (grass)	træ (tree)	blomst (flower)	bil (car)	bil	træ	bil	bil	bil
bil (car)	cykel (bike)	tog (train)	vind (wind)	vind	tog	vind	tog	tog
vind (wind)	regn (rain)	solskin (sunshine)	mandag Monday	mandag	mandag	mandag	mandag	mandag

Five first rows in dataset: Here the Explicit Semantic Analysis (ESA) model (Gabrilovich and Markovitch, 2007) detects all five correct, while the word2vec models selects the wrong term multiple times.

Dataset available at https://github.com/fnielsen/dasem/blob/master/dasem/data/four_words.csv

Word intrusion evaluation



Accuracy in percentage for guessing the odd-one-out among four terms.

Bigger is better. ESA better than Word embedding.

Wordsim353-da evaluation

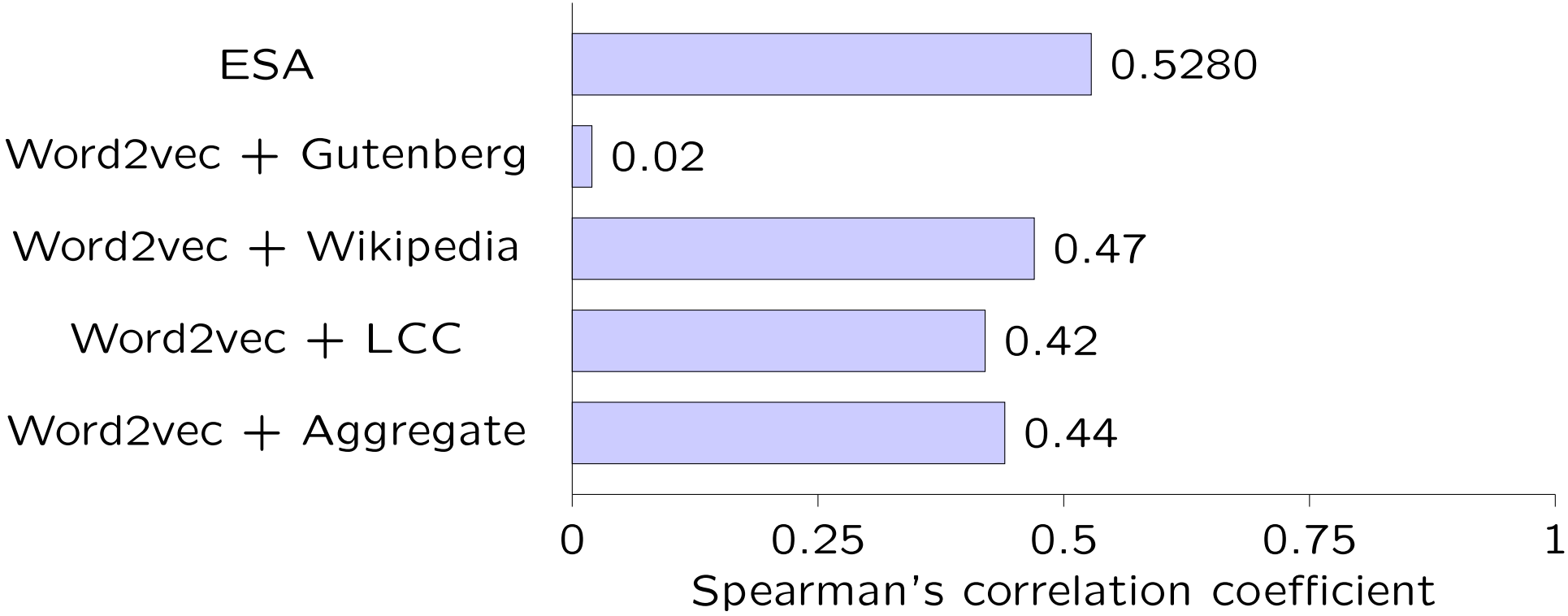
Danish translation of the classic English word list

Word 1	da1	Word 2	da2	Human (mean)	Problem
love	kærlighed	sex	sex	6.77	
tiger	tiger	cat	kat	7.35	
tiger	tiger	tiger	tiger	10	
book	bog	paper	papir	7.46	
computer	computer	keyboard	tastatur	7.62	
⋮					
football	fodbold	soccer	fodbold	9.03	1
⋮					

Only 319 word pairs used in the further analysis due to "problems".

Compute similarity with the semantic models and compare with the human annotation.

Wordsim353-da evaluation



Spearman's correlation coefficient between semantic model and human annotation on the wordsim353-da word pair data.

Bigger is better. ESA better than Word embedding.

AFINN sentiment word list evaluation

AFINN word list with 3552 Danish words labeled with sentiment between −5 and +5 available at https://github.com/fnielsen/afinn/:

```
absorberet       1
acceptere        1
accepterede      1
...
flagskib         2
flerstrengede    2
flerstrenget     2
flop            -2
flot             3
flov            -2
fluekneppende   -3
flueknepperi    -3
```

Prediction of the sign of the sentiment label from AFINN word list.

Predicting AFINN word sentiment

Accuracy for a number of classifiers trained to predict sign of AFINN sentiment score from the representation in the word embedding:

Classifier	Gutenberg	Wikipedia	LCC	Aggregate
MostFrequent	0.596 (0.019)	0.632 (0.027)	0.653 (0.006)	0.646 (0.013)
AdaBoost	0.644 (0.015)	0.754 (0.016)	0.806 (0.009)	0.829 (0.010)
DecisionTree	0.564 (0.018)	0.645 (0.019)	0.716 (0.011)	0.721 (0.020)
GaussianProcess	0.660 (0.020)	0.741 (0.022)	0.784 (0.014)	0.812 (0.011)
KNeighbors	0.615 (0.017)	0.711 (0.022)	0.765 (0.011)	0.796 (0.014)
Logistic	0.694 (0.015)	0.779 (0.016)	0.832 (0.011)	0.853 (0.009)
PassiveAggressive	0.624 (0.051)	0.723 (0.036)	0.792 (0.024)	0.830 (0.030)
RandomForest	0.622 (0.017)	0.722 (0.024)	0.774 (0.009)	0.791 (0.008)
RandomForest1000	0.672 (0.012)	0.777 (0.020)	0.825 (0.010)	0.860 (0.011)
SGD	0.653 (0.021)	0.758 (0.018)	0.808 (0.024)	0.836 (0.020)

Table 1: Classifier accuracy for sentiment prediction over *scikit-learn* classifiers with Project Gutenberg, Wikipedia, LCC and *aggregate* corpora Word2vec features. The *MostFrequent* classifier is a baseline predicting the most frequent class whatever the input might be. *SGD* is the stochastic gradient descent classifier. The values in the parentheses are the standard deviations of the accuracies of 10 training/test set splits.

Explicit semantic representation

Why explicit semantic representations?

> "When ConceptNet is combined with word embeddings acquired from distributional semantics (such as word2vec), it provides applications with understanding that they would not acquire from distributional semantics alone, nor from narrower resources such as WordNet or DBPedia. We demonstrate this with state-of-the-art results on intrinsic evaluations of word relatedness that translate into improvements on applications of word vectors, including solving SAT-style analogies." — (Speer et al., 2016)

DanNet

DanNet (Pedersen et al., 2009) inspired by the the English language (Princeton) WordNet.

BabelNet (Navigli and Ponzetto, 2010) is a multilingual wordnet, see, e.g., kaffemaskine.

(Collaborative InterLingual Index (Bond et al., 2016))

Wikidata

Wikidata at https://www.wikidata.org is a collaborative wiki for structured data.

Project from the Wikimedia Foundation with development in Wikimedia Deutschland in Berlin.

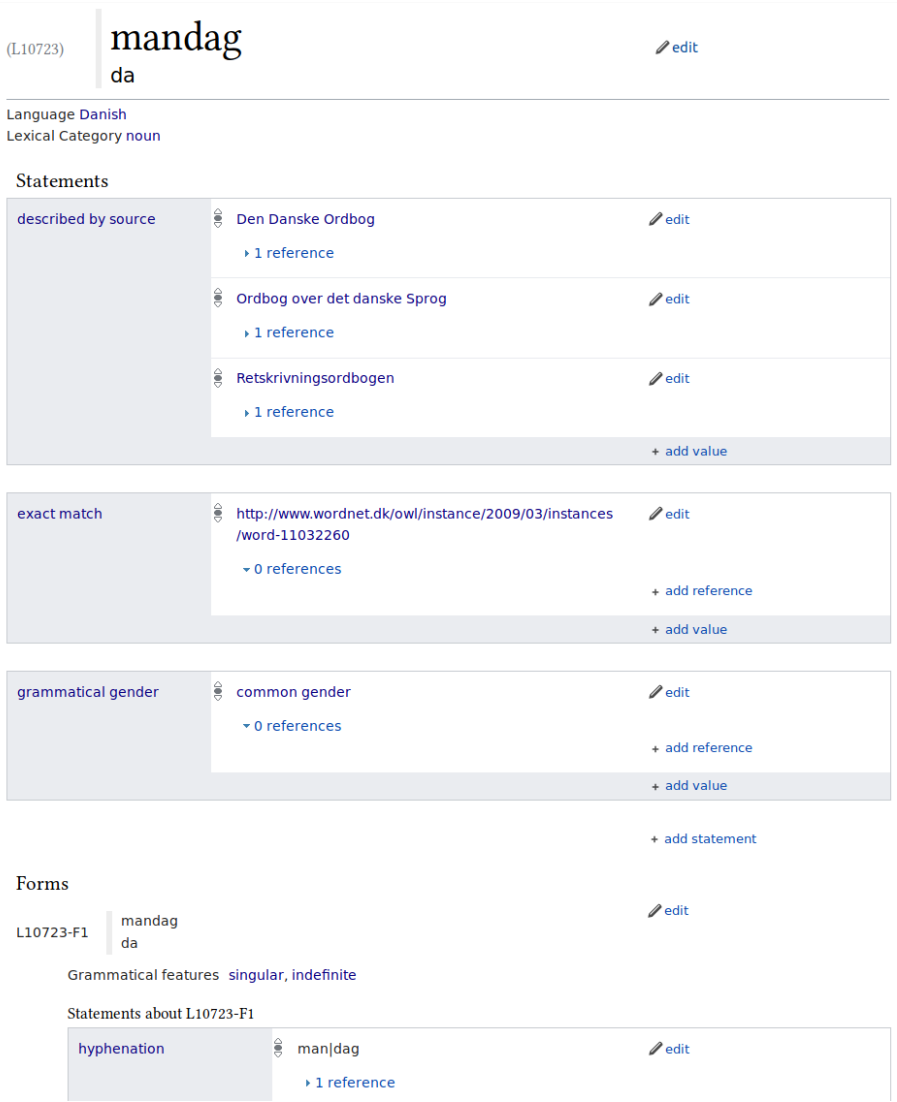
Close to 50 million "items" (concepts)

Items connected via properties: somewhere around 5000 to choose from.

Wikidata

Wikidata is multilingual, here “kaffemaskine”.

Wikidata lexemes

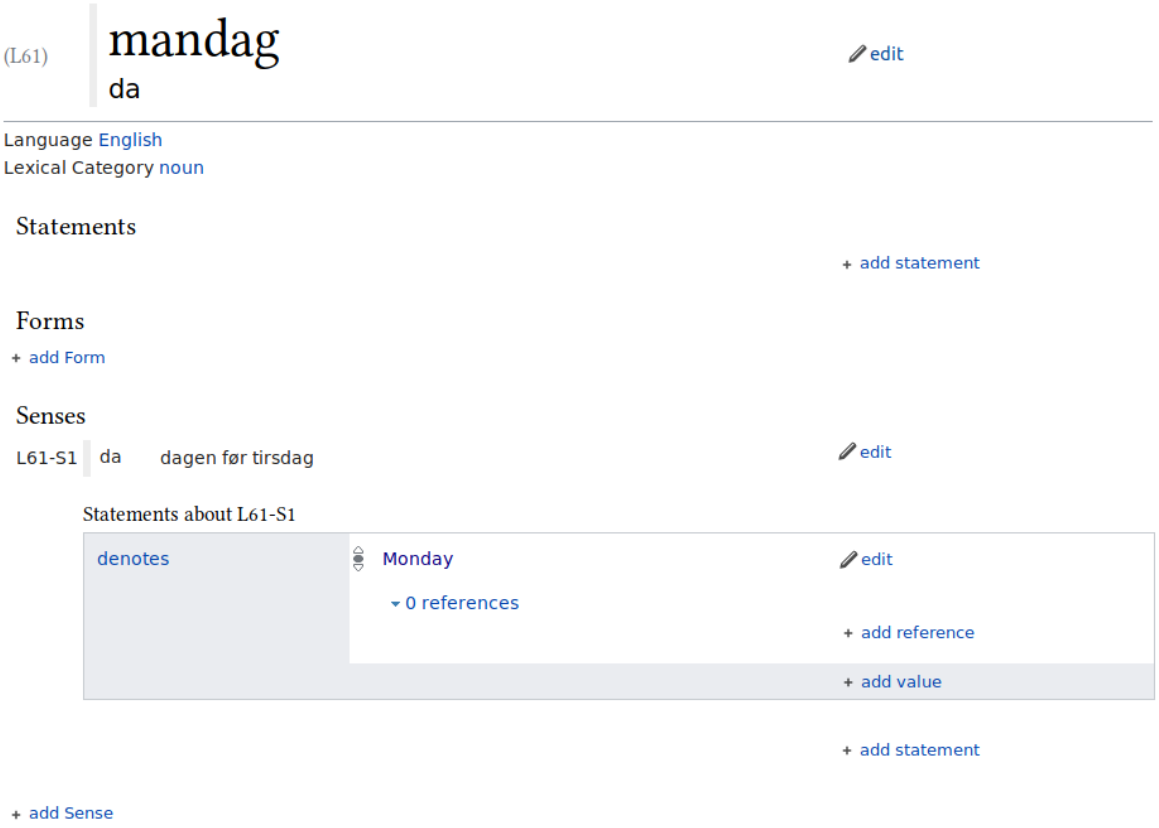


In 2018, Wikidata implemented support for lexemes.

Wikidata lexeme items describe words (lexemes), their language, word class, some grammatical features, e.g., grammatical gender.

. . . and records the multiple forms, e.g., for L10723 (“mandag”), mandag, mandagen, mandage, mandagene, mandags, etc. with their feature, grammatical number, definiteness, etc.

Wikidata: senses of lexemes

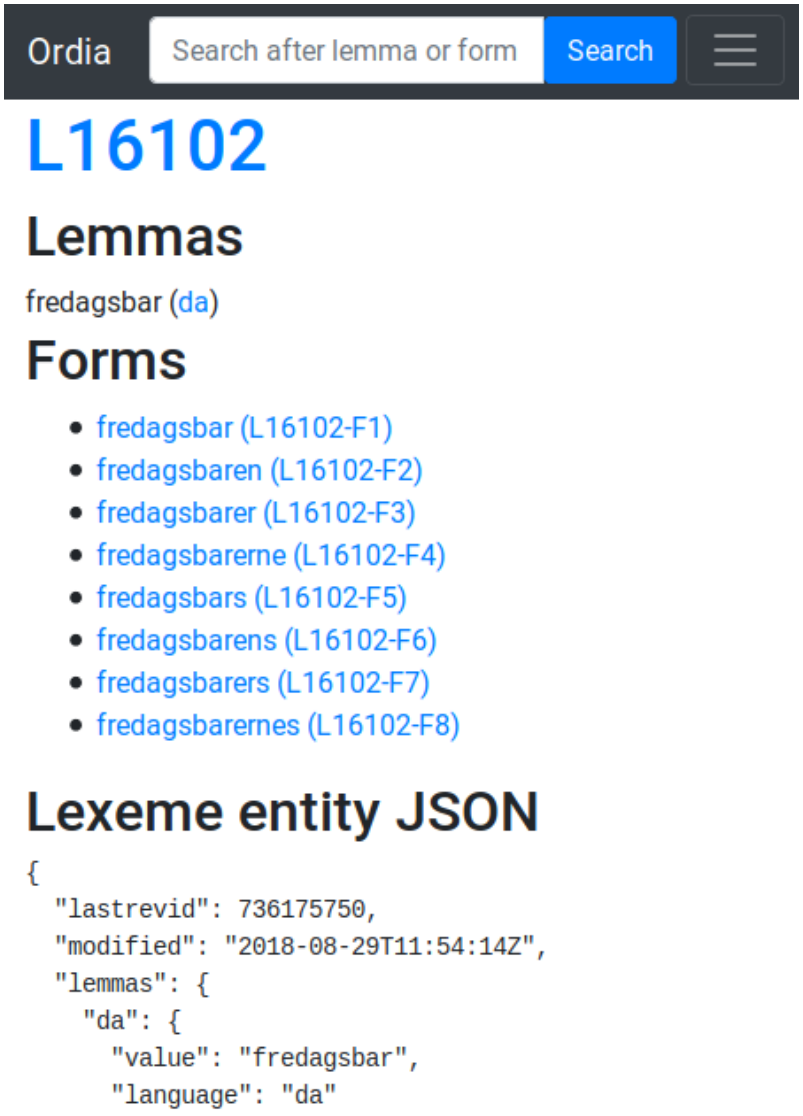


2018 August version of Wikidata does not enable users to link lexemes with senses → we have no semantics! :(

But senses for lexemes are under development and a test version is operational at https://wikidata.beta.wmflabs.org

For instance, we can say that the Danish noun spelled as "mandag" has a sense and that sense denotes the concept described by the Wikidata beta item Q487758, which is mandag/Monday/Montag.

Ordia

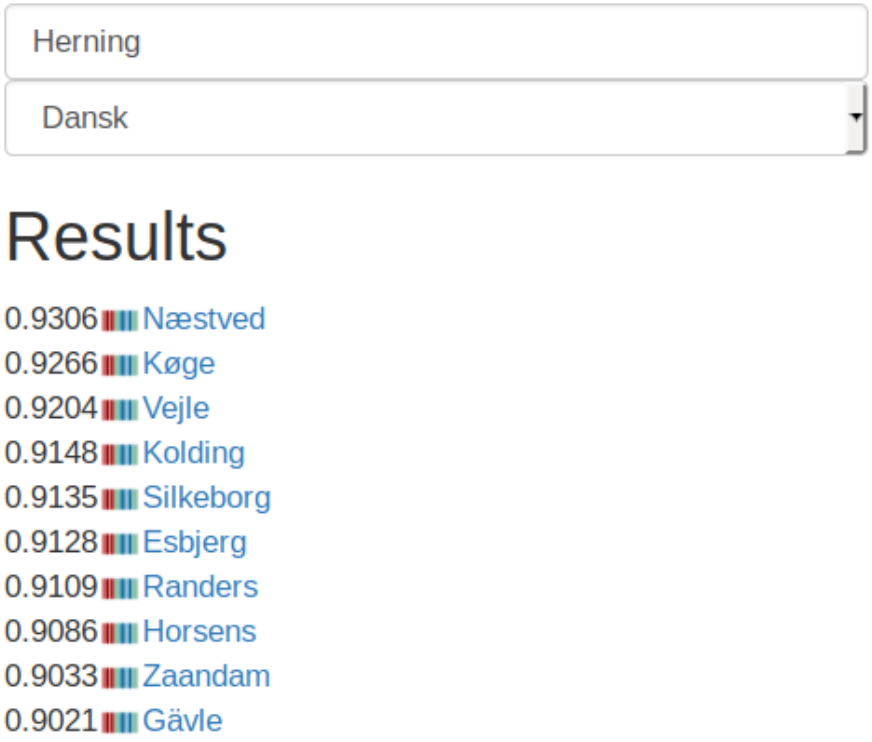


Tools for Wikidata lexemes are not plentiful: Querying and entering is a bit cumbersome, e.g., how many Danish lexemes are there? And the Wikidata Query Service SPARQL engine does not yet support Wikidata lexemes.

Lucas Werkmeister's tools: Wikidata Lexeme Graph Builder and Wikidata Lexeme Forms.

The Ordia webservice at `https://tools.wmflabs.org/ordia/` with rudimentary search and display.

Wembedder



Wembedder: Knowledge graph embedding of items in Wikidata (Nielsen, 2017) running at https://tools.wmflabs.org/wembedder/, — but also downloadable.

Related to RDF2Vec (Ristoski and Paulheim, 2016).

For instance, finding related items based on word2vec-based knowledge graph embedding. Here for the town *Herning*.

Pre-trained models distributed from Zenodo at https://zenodo.org/record/823195 and https://zenodo.org/record/827339.

Wembedder within Python

Wembedder Python session using the imported Gensim (Řehůřek and Sojka, 2010) functionality:

```
from wembedder.model import Model
model = Model.load()
model.most_similar("Q21178")  # Næstved
```

Gives Herning (Q27393) is the top most similar Wikidata item:

```
[('Q27393', 0.9306493997573853), ('Q21184', 0.929047 ...
```

Concept analogies: "Denmark is to Copenhagen what Germany is to":

```
model.most_similar(positive=['Q6581072', 'Q12097'],
                   negative=['Q6581097'])
```

Results in: Berlin, Frankfurt am Main, Köln, München, ...

The usual suspect with Wembedder

“Man is to king as woman is to . . . ?”

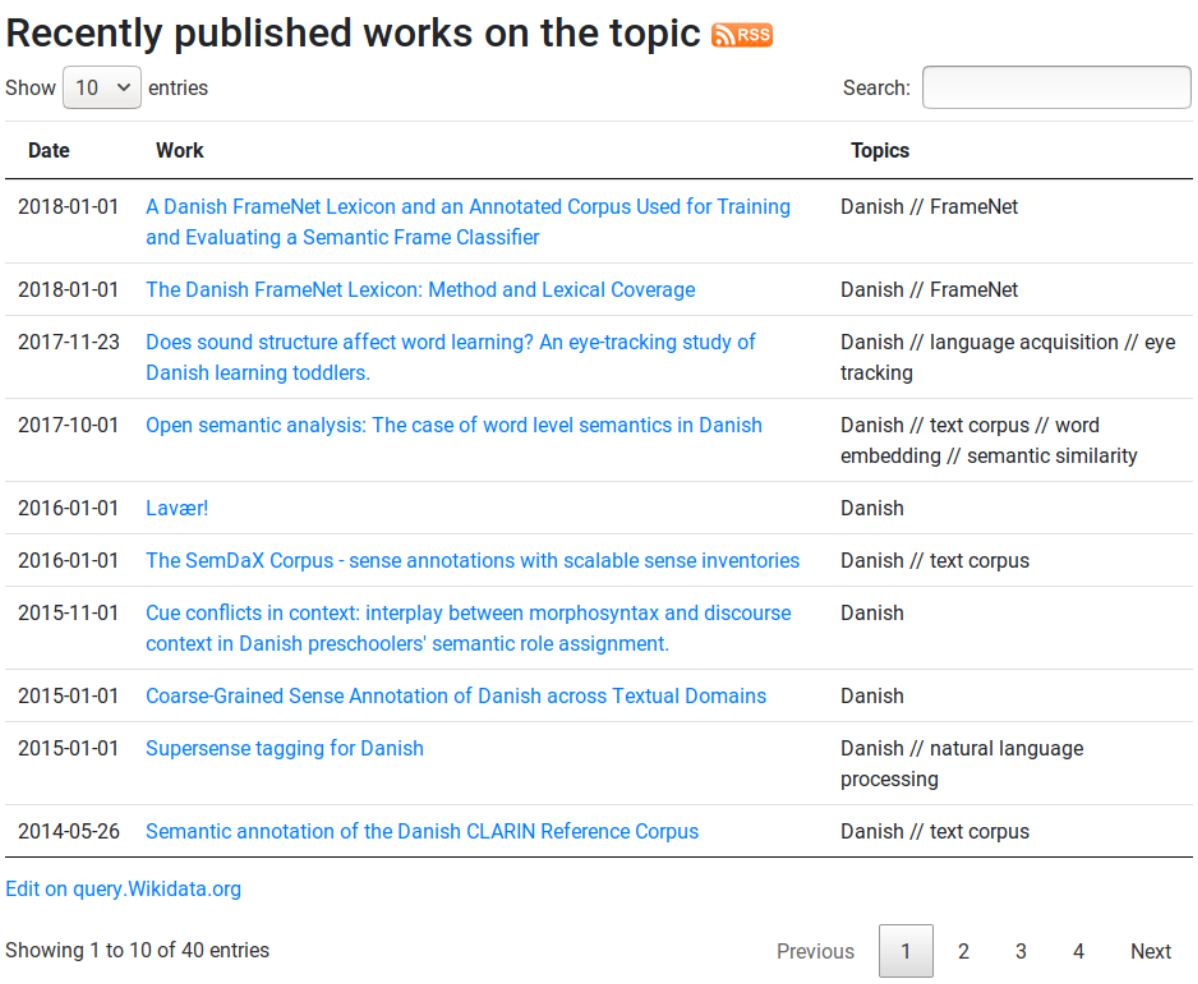
Q6581072 (kvinde/weiblich/female/kvinna), Q12097 (konge/König/king/?), Q6581097 (mand/männlich/male/madur)

```
model.most_similar(positive=['Q6581072', 'Q12097'],
                   negative=['Q6581097'])
```

Gives Q719039 (dronning/Königin/queen consort), Q385468 (Elizabeth, female given name), . . .

Resources

Scholia: Showing all science with Wikidata



Recently published works on the topic RSS

Show 10 entries Search:

Date	Work	Topics
2018-01-01	A Danish FrameNet Lexicon and an Annotated Corpus Used for Training and Evaluating a Semantic Frame Classifier	Danish // FrameNet
2018-01-01	The Danish FrameNet Lexicon: Method and Lexical Coverage	Danish // FrameNet
2017-11-23	Does sound structure affect word learning? An eye-tracking study of Danish learning toddlers.	Danish // language acquisition // eye tracking
2017-10-01	Open semantic analysis: The case of word level semantics in Danish	Danish // text corpus // word embedding // semantic similarity
2016-01-01	Lavær!	Danish
2016-01-01	The SemDaX Corpus - sense annotations with scalable sense inventories	Danish // text corpus
2015-11-01	Cue conflicts in context: interplay between morphosyntax and discourse context in Danish preschoolers' semantic role assignment.	Danish
2015-01-01	Coarse-Grained Sense Annotation of Danish across Textual Domains	Danish
2015-01-01	Supersense tagging for Danish	Danish // natural language processing
2014-05-26	Semantic annotation of the Danish CLARIN Reference Corpus	Danish // text corpus

Edit on query.Wikidata.org

Showing 1 to 10 of 40 entries Previous 1 2 3 4 Next

Scholia is a webservice from https://tools.wmflabs.org/scholia/ generating an overview of science from the information in Wikidata.

For researcher profiles, scientometrics, bibliographic reference management, information discovery (find relevant papers, scientific meetings, researchers, funding opportunities, . . .).

Recently published works on the topic *Danish*

Danish resources

References

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching Word Vectors with Subword Information.

Bond, F., Vossen, P., McCrae, J. P., and Fellbaum, C. (2016). CILI: the Collaborative Interlingual Index. *Proceedings of the Eighth Global WordNet Conference*, pages 50–57.

Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. *Proceedings of the 20th international joint conference on Artifical intelligence*, pages 1606–1611.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning Word Vectors for 157 Languages. *Proceedings of the 11th edition of the Language Resources and Evaluation Conference*.

Navigli, R. and Ponzetto, S. P. (2010). BabelNet: building a very large multilingual semantic network. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225.

Nielsen, F. Å. (2017). Wembedder: Wikidata entity embedding web service. DOI: 10.5281/zenodo.1009127.

Pedersen, B. S., Nimb, S., Asmussen, J., Sørensen, N. H., Trap-Jensen, L., and Lorentzen, H. (2009). DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, 43:269–299. DOI: 10.1007/S10579-009-9092-1.

Ristoski, P. and Paulheim, H. (2016). RDF2Vec: RDF Graph Embeddings for Data Mining. *The Semantic Web – ISWC 2016*, pages 498–514. DOI: 10.1007/978-3-319-46523-4_30.

Speer, R., Chin, J., and Havasi, C. (2016). ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 4444–4451.

Řehůřek, R. and Sojka, P. (2010). Software framework for topic modelling with large corpora. *New Challenges For NLP Frameworks Programme*, pages 45–50.