Word2Dict – Lemma Selection and Dictionary Editing Assisted by Word Embeddings

Nicolai Hartvig Sørensen, Sanni Nimb
Society for Danish Language and Literature
E-mail: nhs@dsl.dk, sn@dsl.dk

Abstract

We describe the use of a tool that assists lexicographers with extending the lexical coverage of an online Danish dictionary. The tool is based on a word embedding model (word2vec) trained on a large Danish corpus, and it presents semantically related lemmas already included in the dictionary and, importantly, their definitions. Furthermore, lemma candidates, i.e. words from the corpus which are *not* included in the dictionary, are presented in the tool, supplemented by information on corpus frequency. The tool thereby facilitates the lemma selection as well as the process of writing consistent definitions across synonyms and near synonyms. We discuss the shortcomings of the tool and the semantic model when it comes to identifying words similar in meaning from different genres and registers. We also look closer into whether it does in fact benefit the dictionary-making process or not by studying a number of previously edited words, including their synonyms, and comparing them with the output data from the tool.

Keywords: lemma selection, word embedding, word2vec, semantic similarity, dictionary-making process

1 Background

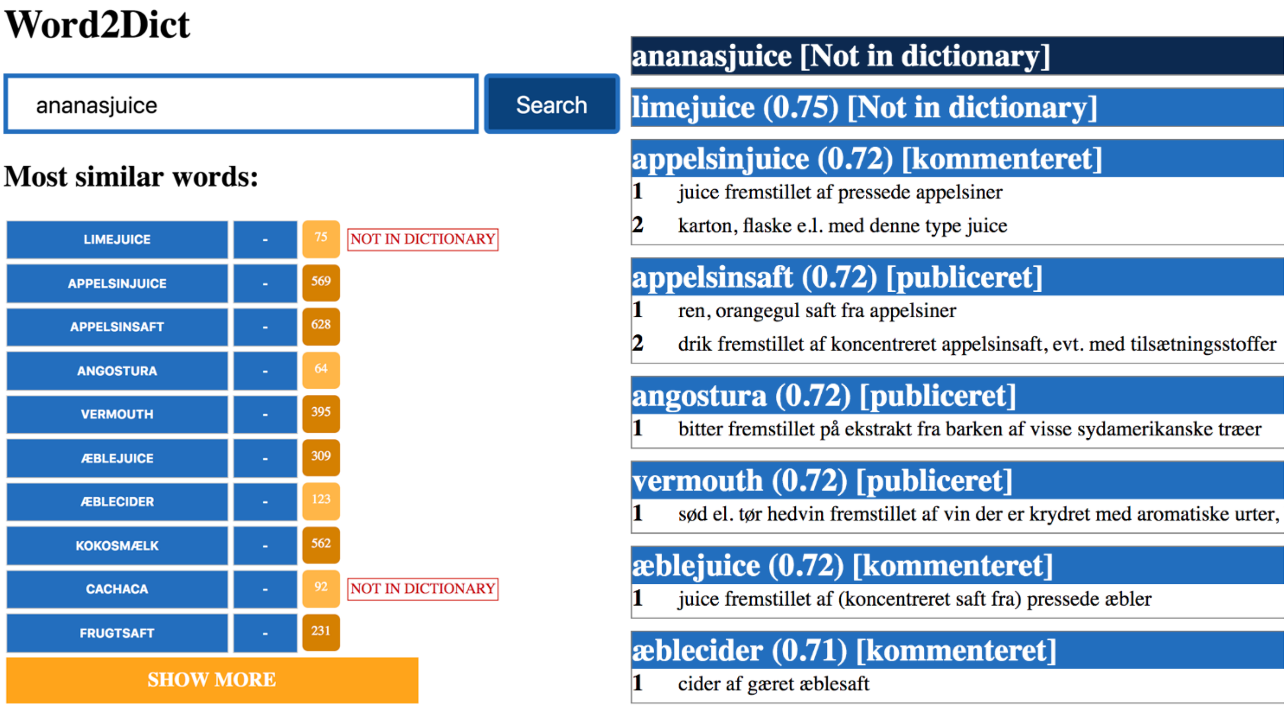
When compiling a new dictionary, it is a well-known and well-proven strategy to edit the lemmas in a semantic order rather than strictly alphabetically in order to ensure consistency of definitions among semantically similar words (cf. e.g. Lorentzen 2004). This strategy was adopted when the printed version of the dictionary we are working on, *Den Danske Ordbog* (*The Danish Dictionary*, henceforth DDO), was compiled 1992-2005. However, when the task consists of augmenting the lemmata of an existing dictionary by adding either completely new or formerly neglected lemmas, as is the case with the current DDO project, it is less obvious how to carry out the process. How do you in a fast and consistent way compare new lemma candidates to already described lemmas within the same semantic field in order to ensure the consistency of the definitions? And additionally, how do you obtain the extra advantages of such a semantically driven workflow in order to identify other lemma candidates in the same field? In this paper, we describe the use of a lexicographic tool that we have developed based on a word embedding model in order to present a number of words that are most semantically related to the lemma that the lexicographer is describing, see Figure 1.

Using distributional methods for suggesting semantically similar words is not new in the field of lexicography. Sketch Engine (Kilgarriff & Tugwell 2001), for instance, includes an “automatic thesaurus” for several languages (e.g. for Slovenian (Krek & Kilgarriff 2006)). But to the best of our knowledge, this is the first time it is being used specifically to assist in the editing process to compare the entry with previously written entries.

2 The Word2Dict Tool

“Word embeddings” is a group of techniques used to investigate the semantic similarity of words by mapping the context of the words in a large corpus into multidimensional vectors. In these models, each type in the corpus is represented by a vector in vector space, and the similarity of the words is based solely on the similarity of the context in which they occur. No morphological or syntactical information is used. Word2vec (Mikolov et al. 2013a; Mikolov et al. 2013b) is a very efficient word algorithm that produces word embeddings.

We used the version of the word2vec algorithm implemented in the Gensim Python package (cf. Řehůřek 2017; Řehůřek & Sojka 2010) to train a model based on the Danish corpus used by the lexicographers of DDO (cf. Asmussen 2017). The corpus included at the time of the training roughly 920 million running words, mainly newswire, but also, material from magazines, transcripts from the Danish Parliament, and some fiction, among others, spanning the years 1982 to 2017. We trained the model with 500 features, a window size of five, a minimum occurrence of five for all types, and used the default choice of CBOW instead of skip-gram. The corpus included 6.3 million types, five million of which occurred less than five times. The training took roughly 18 hours on a 2017 MacBook Pro.



LIMEJUICE	-	75	NOT IN DICTIONARY
APPELSINJUICE	-	569	
APPELSINSAFT	-	628	
ANGOSTURA	-	64	
VERMOUTH	-	395	
ÆBLEJUICE	-	309	
ÆBLECIDER	-	123	
KOKOSMÆLK	-	562	
CACHACA	-	92	NOT IN DICTIONARY
FRUGTSAFT	-	231	

ananasjuice [Not in dictionary]

limejuice (0.75) [Not in dictionary]

appelsinjuice (0.72) [kommenteret]

1 juice fremstillet af pressede appelsiner

2 karton, flaske e.l. med denne type juice

appelsinsaft (0.72) [publiceret]

1 ren, orangegul saft fra appelsiner

2 drik fremstillet af koncentreret appelsinsaft, evt. med tilsætningsstoffer

angostura (0.72) [publiceret]

1 bitter fremstillet på ekstrakt fra barken af visse sydamerikanske træer

vermouth (0.72) [publiceret]

1 sød el. tør hedvin fremstillet af vin der er krydret med aromatiske urter,

æblejuice (0.72) [kommenteret]

1 juice fremstillet af (koncentreret saft fra) pressede æbler

æblecider (0.71) [kommenteret]

1 cider af gæret æblesaft

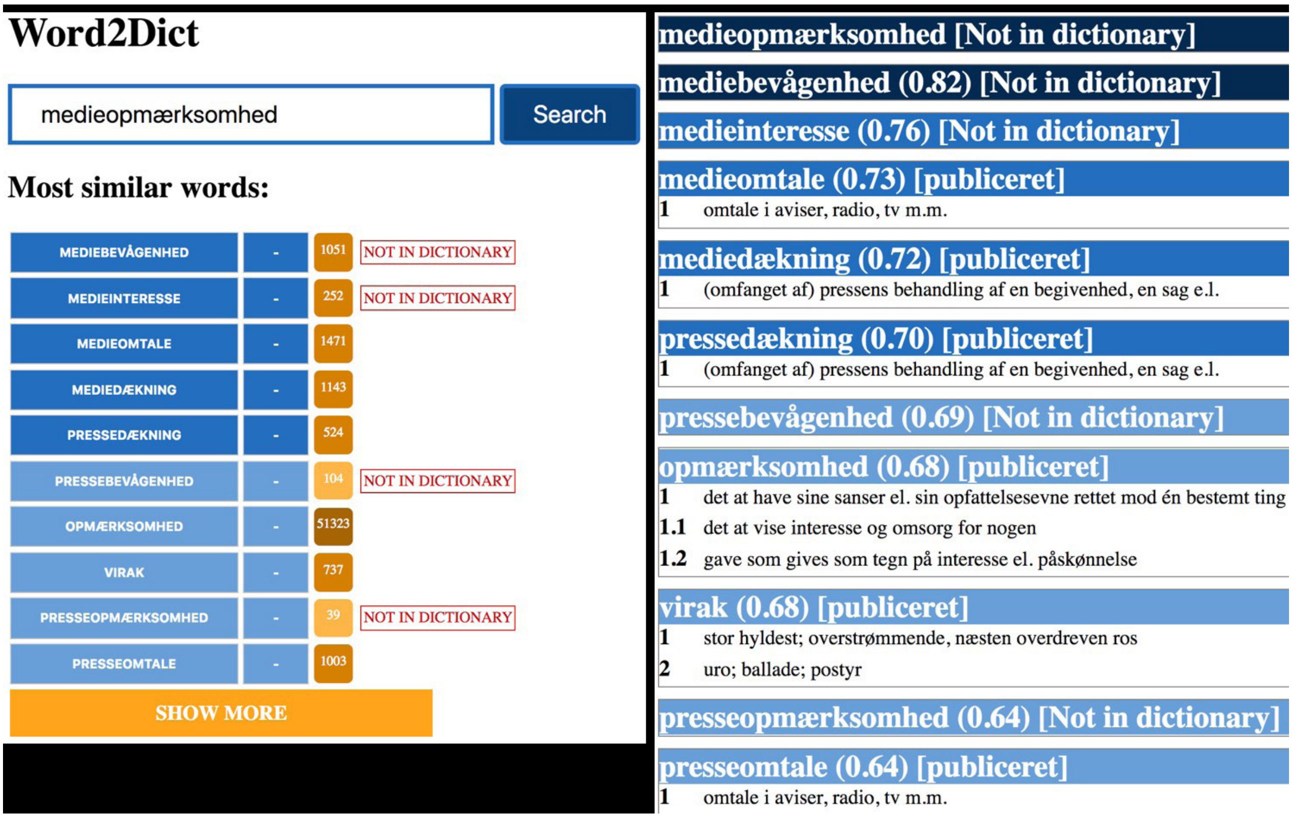
Figure 1: A search for *ananasjuice* (‘pineapple juice’). To the left (the top-half of the interface) we see the most similar words according to the context in which they appear in a corpus. Frequency counts for each word and whether or not the word is included in DDO is also displayed. The frequency counts are color-coded for quicker visual decoding: the darker the color, the higher the frequency. To the right (the bottom-half of the interface), definitions of the words already in the dictionary are shown, as well as their editorial status (e.g. “publiceret” (‘published’)) and the similarity score from the model (e.g. “0.75”, “0.71” – 1.0 equals identical).

We then implemented the Word2Dict tool as a CherryPy web app, by modifying the open-source Gensim HTTP service (Řehůřek 2014) and included calls to a simple API for accessing DDO data. This means that the lexicographic tool combines the word2vec model with on the fly lookups in DDO.

This combination makes it possible to immediately determine whether or not the words returned by the word2vec model are already included in the dictionary – words which are not included are marked by “not in dictionary”, as shown in Figure 1 (left). The higher up the word is, the more similar to the query word (in this case *ananasjuice*). For the words that are included, their definitions are looked up and displayed immediately in the bottom half of the interface, see Figure 1 (right). This allows the lexicographer to be inspired by – and steal from – existing definitions of semantically similar words, and thus ensure greater consistency in the descriptions within semantic domains. The second most similar word to *ananasjuice* in the model is *appelsinjuice* (‘orange juice’), and DDO describes this word with two main senses (the juice itself, and a container with that juice). In this case, the editor may conclude that the two definitions of *appelsinjuice* might serve as an excellent starting point when editing the new word *ananasjuice*.

The Word2Dict tool also makes it easy to consider the words not yet included in the dictionary as new lemma candidates. Figure 2 demonstrates a search for *mediaopmærksomhed* (‘media attention’) with several similar words not yet included in the dictionary.

The DDO project uses corpus frequency as an important indicator of relevance. Therefore, information on corpus frequency is also presented in the tool, supplying the editor with important knowledge when deciding whether to accept or discard the lemma candidate. In this case the tool reveals that *mediabevågenhed* (‘media attention’) and *medieinteresse* (‘media attention’) are rather frequent words, while *pressebevåghenhed* (‘attention from the press’) and *presseopmærksomhed* (‘attention from the press’) are less so. In addition to this, the similarity score suggests that both *mediabevågenhed* (0.82) and *medieinteresse* (0.76) might be considered as good synonyms of the query word, which they indeed are – as indicated by the English glosses above.



Word2Dict

Most similar words:

MEDIEBEVÅGENHED	-	1051	NOT IN DICTIONARY
MEDIEINTERESSE	-	252	NOT IN DICTIONARY
MEDIEOMTALE	-	1471	
MEDIEDÆKNING	-	1143	
PRESSEDÆKNING	-	524	
PRESSEBEVÅGENHED	-	104	NOT IN DICTIONARY
OPMÆRKSOMHED	-	51323	
VIRAK	-	737	
PRESSEOPMÆRKSOMHED	-	39	NOT IN DICTIONARY
PRESSEOMTALE	-	1003	

medieopmærksomhed [Not in dictionary]

mediebevågenhed (0.82) [Not in dictionary]

medieinteresse (0.76) [Not in dictionary]

medieomtale (0.73) [publiceret]
1 omtale i aviser, radio, tv m.m.

mediedækning (0.72) [publiceret]
1 (omfanget af) pressens behandling af en begivenhed, en sag e.l.

pressedækning (0.70) [publiceret]
1 (omfanget af) pressens behandling af en begivenhed, en sag e.l.

pressebevågenhed (0.69) [Not in dictionary]

opmærksomhed (0.68) [publiceret]
1 det at have sine sanser el. sin opfattelsesevne rettet mod én bestemt ting
1.1 det at vise interesse og omsorg for nogen
1.2 gave som gives som tegn på interesse el. påskønnelse

virak (0.68) [publiceret]
1 stor hyldest; overstrømmende, næsten overdreven ros
2 uro; ballade; postyr

presseopmærksomhed (0.64) [Not in dictionary]

presseomtale (0.64) [publiceret]
1 omtale i aviser, radio, tv m.m.

Figure 2: The most similar words to *mediaopmærksomhed* (‘media attention’), top-half of the interface to the left, bottom-half to the right.

We believe that the Word2Dict have the potential of speeding up the editing process without sacrificing the quality of the work. It has, however, proven difficult to reliably measure the speed-up, as the Word2Dict tool also reveals hidden inconsistencies in the existing entries that need to be fixed in order to decide on the semantic structure of the new word. It seems unjust to blame the tool for earlier shortcomings. However, we make an informal estimate of the benefits of editing the dictionary when also including relevant lemma candidates revealed by Word2Dict in section 3.3 below.

A more basic question therefore becomes whether or not the model returns relevant words at all. Bearing in mind that the word2vec model knows nothing about the morphology of the words in the corpus or syntax of the sentences, only the word forms in the context, does it return words that are relevant to lexicographic work? Does it reflect the mind of a lexicographer sufficiently well? This will be tested in section 3.

3 Testing Word2Dict

Since 2015 the DDO dictionary has been extended with more than 8,400 lemmas. Many of these are compounds (especially noun compounds) which were already occurring with a mid-to-low frequency in Danish corpora texts from the 1980s and 1990s. However, the sense descriptions were left out of the dictionary due to space limits when the first, printed edition of DDO was compiled. Other lemmas that have been included in recent years are new words in the Danish language since 1990, e.g. English loanwords. All the newly included lemmas, most of which are nouns, are represented with a mid-to-low corpus frequency in the large corpus of approximately one billion words mainly consisting of newswire, which is today used for lemma selection and corpus investigations in the dictionary-making process.

Almost 4,000 of the 8,400 lemmas contain information on “related words”, i.e. on synonyms, antonyms and “see also” words (presented to the user with the label *Se også* (‘see also’) – a large part of these are near-synonyms) in separate fields. These have been manually selected by the editor without the use of any tool, e.g. based on subjective judgment and, maybe more importantly, to a high degree by consulting the lexical description of the approximately 65,000 lemmas which were already part of the printed edition of DDO (some of them being synonyms of these). In order to find out whether the Word2Dict tool supplies the editor with the same semantically related words that he or she would include manually, based on a subjective judgment, we compared the output data of the tool with a randomly selected number of the newly included DDO lemmas containing related words.

We studied 100 randomly chosen lemmas of 1,332 lemmas with related words which were edited and included in DDO in 2016. Of the 1,332 lemmas, 1,160 (87%) are nouns, 132 (10%) are adjectives, 36 (3%) verbs and only three are adverbs, and the study we present therefore mainly focuses on nouns. We wanted to see whether the related word (or words) included in the dictionary entry by the editor was in fact also part of the output of the Word2Dict tool. We also wanted to see whether it happened to be no. 1 or 2 on the list, considering this to confirm the high quality of the Word2Dict output. Furthermore, we studied whether the output contained highly relevant words that are *not* already mentioned in the dictionary entries, but in our opinion ought to be included and maybe even replace the related words selected by the editor without the use of the tool. First, we exemplify the investigation method by presenting two quite different results:

1. When we studied the noun lemma *fejlinformation* (‘erroneous information’) in DDO, we found the noun synonym *misinformation* (‘misinformation’). Word2Dict gives as output of *fejlinformation* the following list of nouns: *vildledning* (‘deception’), *misinformation* (‘misinformation’), *forhaling* (‘delay’), *fordrejning* (‘misrepresentation’), *fejlfortolkning* (‘misrepresentation’), *obstruction*

(‘obstruction’), *ignoring* (‘ignoring’), *mistænkeliggørelse* (‘casting suspicion on’), *manipulation* (‘manipulation’), *fortielse* (‘concealement’). In this case the Word2Dict output matches the manual selection very well, since the dictionary synonym *misinformation* is the second word on the list. Furthermore, the list supplies us with several other very relevant nouns: *vildledning*, *fordrejning*, *manipulation*, *fortielse*, of which *fordrejning* is not yet included in DDO. In this case the tool would have been very useful in the editing process of the lemma, having facilitated a faster and more consistent dictionary editing flow, since the editing of several near-synonymous noun lemmas could have been carried out at the same time, assuring relevant references between the words.

2. In contrast to the example above, in the case of the noun *mediebranche* (‘media business’) Word2Dict does not supply us with any words of interest. The DDO editor has chosen to present the noun synonym *medieverden* (‘the media world’)), but we got as output from the Word2Dict tool only a list of words denoting other types of working branches (i.e. co-hopynoms of the noun *branche* (‘line of business’)), none of which we found relevant to include in the DDO entry.

In Table 1 we present the results of the study.

In 60% of the cases, the output includes the editor’s manually inserted related word, and in 43% it is either no. 1 or 2 in the output list (see Table 1, case 3), and in only 18% of the studied cases did the tool not supply us with any lexicographically relevant related words (see Table 1, case 4). This, we believe, strongly indicates that the Word2Dict tool sufficiently well reflects the mind of a lexicographer.

3.1 Cases Where the Tool Did Not Find the Editor’s Related Word From the DDO Entries

In 23% of the 100 studied cases, we find that the output does contain relevant words, however the editor’s choice of synonym is *not* among the words found by the Word2Dict tool (case 2 in Table 1). In most cases this is due to a too low frequency of the synonym in our corpus, which is based on newswire and does not represent spoken language, more specific domains or older language, nor much literary language. In these cases, the DDO synonym is of course very relevant. For example, in the cases of the two old-fashioned, poetic nouns *hjerteven* (‘bosom friend’) described by the editor as a synonym in the entry *bedsteven* (‘best friend’) and *kvindekær* (‘fond of women’) described as a synonym in the entry *pigeglad* (‘fond of girls), these are not represented in the Word2Dict output. Nor is the specific language noun *veterinær* (‘veterinary’) on the output list of the common noun *dyrlæge* (‘vet’) in DDO, but in the dictionary the two nouns are presented as synonyms of one another. In these latter cases, the Word2Dict tool is not able to supply the editor with relevant words.

3.2 Cases Where the Tool Finds Relevant Words Not Discovered by the Editor

In far the most cases (82%, case 4 in Table 1) the list supplies us with more relevant synonyms, near-synonyms and antonyms than the ones that were manually selected by the editor for presentation in the DDO entry. In other words, the tool reveals relevant related words that seem to have been left out by the editors, either on purpose or because they were not aware of them. Most of the related word candidates on our output lists appear at a first glance to be more common words compared to the ones mentioned in DDO. The words have more or less the same meaning but are simply more frequent in the modern DDO-corpus based on newswire of today. We therefore guess that the difference between DDO and Word2Dict in this case is caused by the fact that the editors of the newly included lemmas in DDO first of all have based the sense descriptions and related words on the already described DDO-lemmas from the first, printed edition which was based on older Danish dictionaries and words and their meanings in corpus texts from 1982 to 1992. In this case, the introduction of Word2Dict as a supplementary editorial tool would make sure that newer related words from modern corpora are also presented in the entries.

Table 1: The study of the Word2Dict output when compared to already edited lemmas with related words (i.e. synonyms, antonyms and 'see also' words) in DDO

	Comment on output from Word2Dict	**100% = 100 lemmas studied**	**Examples**
(1)	No useful output (often due to low corpus frequency of the input lemma)	5%	*aktieobligation* ('special kind of stock option'), *anvendelsesmulighed* ('application'), *normalpris* ('normal price'), *originaleksemplar* ('original copy')
(2)	Output does contain relevant words, however it does not contain the editor's manually selected related word	23%	*amatørcykelrytter* ('amateur racing cyclist') without the synonym *amatørrytter* *aktualitetsstof* ('news') without the synonym *nyhedsstof* *bedsteven* ('best friend') without *hjerteven* ('bosom friend') *diktaturstat* ('dictatorship') without the synonym *diktaturland* *kødfri* ('without meat) without *vegetarisk* ('vegetarian') *letsindighed* ('carelessness') without the synonym *skødesløshed*
(3)	Output includes the editor's manually selected related word within the first 10 words	60% (28% as no. 1 on list, 15% as no. 2)	*afterparty* ('afterparty') contains the synonym *efterfest* as no. 1 on the list *auktionsfirma* ('auction house') contains the synonym *auktionshus* as no. 1 on the list *blomstringsperiode* ('flowering period') contains *blomstringstid* ('time of flowering') as no. 2 on the list *actionkomedie* ('action comedy') contains *krimikomedie* ('crime comedy') as no. 2 on the list *borgmesterstol* ('mayoralty') contains the synonym *borgmesterpost* as no. 1 on the list *bundkamp* ('match between relegation candidates') contains the synonym *bundopgør* as no. 8 on the list
(4)	Output includes a number of relevant related words which were not (yet) presented by the editor as a related word in the DDO entry	82%	*actionkomedie* ('action comedy'): the synonyms *thrillerkomedie*, *dramakomedie*, *komediedrama* *bundkamp* ('match between relegation candidates'): the synonyms *bundgyser*, *bundbrag*, the antonym *topkamp* ('top-of-thr-table clash') *annektering* ('annexation'): the synonyms *indlemmelse*, *invasion*, *erobring*, *okkupation*
(5)	Output contains lemma candidates among the related words (= words not yet included as a lemma in the dictionary)	32%	*actionkomedie* ('action comedy'): the near-synonym *thrillerkomedie* *bundkamp* ('match between relegation candidates'): the synonyms *bundgyser*, *bundbrag* *amatørcykelrytter* ('amateur racing cyclist'): the synonym *motionscyklist* *annektering* ('annexation'): the synonym *indlemmelse* *bilværksted* ('car repair shop'): the synonym *mekanikerværksted* *diktaturstat* ('dictatorship'): the near-synonym *etpartistat* ('one party state') *dystopisk* ('dystopian'): the near-synonym *mareridsagtig*

In a few cases we even found the output from Word2Dict to be more semantically precise than the synonyms selected by the editors. This is for example the case for *dystopisk* ('dystopian') with the related word *utopisk* ('utopian') in DDO. Word2Dict lists *apokalyptisk* ('apocalyptic'), *mareridsagtig* ('nightmarish'), and *futuristisk* ('futuristic'), all of which we consider more appropriate to describe in the entry of *dystopisk*. Another case is *eliteidrætsudøver* ('elite athlete') with the related word *elitegymnast* ('elite gymnast') in DDO, which is just one of a large selection of more specific hyponyms. We would suggest it to be replaced by synonyms like *topatlet* ('elite athlete'), *topidrætsmand* ('elite athlete'), *elitesportsmand* ('elite athlete'), and *eliteatlet* ('elite ahtlete'). Also in these cases we find that the dictionary-making process would benefit to a high degree from the Word2Dict tool, not only to ensure consistency but also to avoid aggravating omissions.

3.3 Word2Dict Used to Identify Semantically Related Lemma Candidates

In 32% of the studied word examples (case 5 in Table 1) we find lemma candidates among the groups of words which we find semantically related to the headword. This means that in 1/3 of the cases where Word2Dict gives an output, we identify words which not only mean almost the same as the input word, but which are furthermore very good candidates (with a relevant frequency in the corpus) to be described in the dictionary.

By editing words which are semantically related all at once, we are able to reuse and adapt definitions. We estimate that the use of the tool in such cases will speed up the dictionary-making process by at least 10-20% on average, once the editor is confident with using it. It should be taken into consideration that the time benefits vary quite a lot depending on the character of the semantic domain. Those with many relevant lemma candidates (e.g. semantic areas which have not yet been fully described in DDO) allow the editor to greatly speed up the work in contrast to domains where the tool reveals fewer or no relevant lemma candidates.

We also expect the quality of lexical work to improve. By comparing a group of words with related, however not completely identical senses, the editor becomes aware of the subtle differences between the words and is thereby in each case able to formulate a more precise definition. Furthermore, the method facilitates the creation of references between the words.

4 Conclusions and Future Work

While others have investigated the use of word embedding techniques for unsupervised identification of synonyms with some success (cf. e.g. Nguyen et al. 2015, Leeuwenberg et al. 2016), our focus is instead on presenting a set of semantically similar words used as input to *manual* lexicographic inspection. Investigating near-synonyms, co-hyponyms and hyponyms, as well as synonyms, might offer the lexicographer a better overview of how the semantic domain previously has been described in the dictionary.

By comparing the manually selected synonyms in the DDO with the output from the word2vec model, our small-scale study has proved that the model is able to supply us with a number of semantically related words from the same register and genre as the query word. Secondly, we have shown that the model to a great extent suggests words to be included in DDO as synonyms, antonyms and "see also" words, which the lexicographer would otherwise easily neglect. This strongly indicates that the already established definitions of DDO words being part of the output list constitute a very good starting point when it comes to defining the new lemmas to be included in the dictionary.

It is worth noting that the model is not able to identify semantically similar words belonging to very different registers, due to the fact that the textual context of such words differs greatly – this is an inherent feature of word embeddings. For this reason, the lexicographer should not rely solely on the Word2Dict tool when selecting synonyms.

In further work we plan to compare the word2vec with other models for Danish (e.g. the models recently made public from the Sketch Engine group, cf. "Models for download") and to fine-tune the training parameters. We want to study different aspects in order to improve the model, i.e. in order to be able to select the best window size when it comes to use the model for lexicographic work. The work will be carried out in collaboration with the Centre for Language Technology, University of Copenhagen. We also plan to add more features to the tool, for instance making it possible to see not only the definitions but also the synonyms of the already described DDO lemmas returned from the model. This feature would allow the dictionary editor to carry out a consistency check of the synonyms, as well as the definitions of similar words in a very efficient way. Finally, we plan to use the tool in the editing process of a Danish thesaurus published in print 2014 (Nimb et al. 2014), which is currently being extended with more words and expressions and which we plan to publish online in the coming years, depending on funding.

References

Asmussen, J. (2017). Hvor kommer ordene fra? In Nyhedsbrev fra Det Danske Sprog- og Litteraturselskab nr. 5, pp. 6-7.

DDO = *Den Danske Ordbog*. Det Danske Sprog- og Litteraturselskab. Accessed at: http://ordnet.dk/ddo [28/03/2018].

Kilgarriff, A. & Tugwell, D. (2001). WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography. In *Proc. ACL workshop on COLLOCATION: Computational Extraction, Analysis and Exploitation,* pp. 32-38.

Krek, S. & Kilgarriff, A. (2006). Slovene Word Sketches. In T. Erjavec, J. Žganec Gros (eds.) *Proceedings of the 5th Slovenian and 1st International Language Technologies Conference.* Ljubljana, Slovenia.

Leeuwenberg, A., Velab, M., Dehdaribc, J. & van Genabith, J (2016). A Minimally Supervised Approach for Synonym Extraction with Word Embeddings. In *The Prague Bulletin of Mathematical Linguistics, Number 105,* pp. 111–142.

Lorentzen, H. (2004): The Danish Dictionary at large: Presentation, Problems and Perspectives. In G. Williams & S. Vessier (eds). *Proceedings of the Eleventh EURALEX International Congress,* pp. 285-294.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. In *ICLR Workshop*. eprint arXiv:1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In Proc. Advances in Neural Information Processing Systems 26, pp. 3111–3119.

Models for download. Accessed at: https://embeddings.sketchengine.co.uk/static/index.html [01/04/2018].

Nguyen, N.T.H., Miwa, M., Tsuruoka, Y. & Tojo, S. (2015). Identifying synonymy between relational phrases using word embeddings. In *Journal of Biomedical Informatics*, Volume 56, pp. 94-102.

Nimb, S., Trap-Jensen, L. & Lorentzen, H. (2014). The Danish Thesaurus: Problems and Perspectives. In *Proceedings of the 16th EURALEX International Congress,* EURAC research, Bolzano, Italy, pp. 191-199.

Řehůřek, R. & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks*. Valletta, Malta: University of Malta, pp. 46–50.

Řehůřek, R. (2013). *Word2vec in Python, Part Two: Optimizing.* Accessed at https://rare-technologies.com/ word2vec-in-python-part-two-optimizing/ [24/11/2017].

Řehůřek, R. (2014). *Word2vec as an HTTP service.* Accessed at https://github.com/RaRe-Technologies/ w2v_server_googlenews [05/10/2017].