UPDATING THE DICTIONARY: SEMANTIC CHANGE IDENTIFICATION BASED ON CHANGE IN BIGRAMS OVER TIME

Sanni NIMB, Nicolai HARTVIG SØRENSEN, Henrik LORENTZEN

Society for Danish Language and Literature

Nimb, S., Hartvig Sørensen, N., Lorentzen, H. (2020): Updating the dictionary: semantic change identification based on change in bigrams over time. Slovenščina 2.0, 8(2): 112–138

DOI: https://doi.org/10.4312/slo2.0.2020.2.112-138

We investigate a method of updating a Danish monolingual dictionary with new semantic information on already included lemmas in a systematic way, based on the hypothesis that the variation in bigrams over time in a corpus might indicate changes in the meaning of one of the words. The method combines corpus statistics with manual annotations. The first step consists in measuring the collocational change in a homogeneous newswire corpus with texts from a 14 year time span, 2005 through 2018, by calculating all the statistically significant bigrams. These are then applied to a new version of the corpus that is split into one sub-corpus per year. We then collect all the bigrams that do not appear at all in the first three years, but appear at least 20 times in the following 11 years. The output, a dataset of 745 bigrams considered to be potentially new in Danish, are double annotated, and depending on the annotations and the inter-annotator agreement, either discarded or divided into groups of relevant data for further investigation. We then carry out a more thorough lexicographical study of the bigrams in order to determine the degree to which they support the identification of new senses and lead to revised sense inventories for at least one of the words Furthermore we study the relation between the revisions carried out, the annotation values and the degree of inter-annotator agreement. Finally, we compare the resulting updates of the dictionary with Cook et al. (2013), and discuss whether the method might lead to a more consistent way of revising and updating the dictionary in the future.

Keywords: corpus statistics, bigrams, dictionary update, semantic change, Danish

1 INTRODUCTION AND MOTIVATION

The Danish Dictionary (DDO) was originally edited from 1994 to 2003 based on studies of Danish word senses in corpus texts from 1983-1992, in total 40 million tokens (cf. Norling-Christensen and Asmussen, 1998). It was initially published in print 2003-2005 and at the time it described the senses of 66,000 lemmas (cf. Lorentzen, 2004). Since 2009 it has been available online at ordnet.dk/ddo, and in recent years the main focus has been to update it with new lemmas. Today, 25 years after the first editorial work was carried out, the dictionary covers 100,000 lemmas, and time has come to update the earliest edited ones by supplying them with new senses, new fixed expressions, new collocations, and also new citations. After the first published version of the dictionary, this has only been done sporadically, as a result of user suggestions and whenever the lexicographers observed new ways of using a word in the language. When it comes to citations, the dating of these in the dictionary can be used as an indicator since entries with only older ones probably need an update. The editorial staff is currently going through all senses which are only illustrated with a citation from the 1980s. However, presenting more updated citation information would also be relevant in many other cases, but these are hard to find systematically, as are those cases where there is a need for new collocations or even more importantly, for a slightly different sense description or even a new sense, maybe in the form of a fixed expression. Our aim is to be able to supply the current practice building on suggestions from users and editorial observations with a more systematic approach across the whole vocabulary, based on corpus statistics.

2 METHOD

It is a well-established fact that collocational change might indicate sense change (Tahmasebi et al., 2018; Pollak et al. 2019; Traugott, 2017). For instance, Pollak et al. (2019) compare automatically extracted collocations from computer-mediated communication (such as blogs and social networks) with those from a general language reference corpus and discover not only topic/genre-related new words, but also new meanings of previously lexicographically described vocabulary. In contrast to this, the present paper is based on the comparison of sets of automatically extracted collocations from corpora which are similar in composition and genre, but which instead cover different timespans. We describe a method where the collocational change in these corpora is used as input for lexicographers in their search for new meanings of already included vocabulary in a dictionary. We initially calculate the statistically significant variation in bigrams in a corpus and create a dataset of those that are estimated to be new in Danish texts. Independently of each other, two lexicographers judge whether, at a first glance, the bigrams indicate the need for a semantic revision of the lemmas involved, and if so, should it be 1) in the form of a defined sense or fixed expression, or 2) in the form of a collocation added to an existing sense with no need of explanation? Afterwards, the lemmas represented by the bigrams which were marked as 1) or 2) either by one or both lexicographers are more thoroughly inspected, leading to a revision in the dictionary when required, otherwise not. The judgments of the data are based on a set of internal guidelines to be followed by editors of the dictionary when new lemmas, senses and fixed expressions are to be added.

In this paper, we study and discuss the relation between annotation value (1 or 2), inter-annotator agreement and the final type of update to be carried out. We conclude that especially when the annotators agree that the bigram is semantically relevant, but disagree upon which exact type of semantic change it indicates, we find many new senses. Finally, we compare our findings with Cook et al. (2013).

In the next section we describe the statistical method that we estimate to be suitable for our purpose, as well as the computational creation of the dataset.

3 CREATING THE DATASET

Since 2005, the Society for Danish Language and Literature has collected newswire data of roughly the same size daily. The newswire corpus consists of 20 to 40 million tokens for each year, 512 million running words in all. It consists of articles that are randomly selected from major Danish newspapers each day (due to license restrictions the corpus is not publicly available, but see korpus.dsl.dk/resources.html for other Danish corpora from DSL that are).

The homogeneous data type, the relatively even distribution, and the sufficiently long time-scale make this corpus ideal for investigating our hypothesis. If lexical data in the form of a token or e.g. a bigram has not occurred at all in the initial period of the text collection, but occurs regularly in the more recent corpus texts, it might indicate that it is a neologism or, in the case of bigrams, either a new expression in the language, or a new way of using one (or more) of the words involved. We have previously used this method to identify potential new single lemmas for DDO, but have never evaluated the method formally. We divided the corpus by year, and selected all tokens which do not appear at all in the first 3 years, 2005-7, but appear frequently during the remaining 11 years. The set of tokens was checked by a lexicographer who removed proper nouns and errors, and now it is used as input to lexicographers in the task of supplying DDO with new lemmas. However, it has not been studied to which degree these lemma candidates do end up being included as new lemmas in the dictionary. This paper describes the same method carried out on bigrams, but takes it a step further. In this case not just one, but two lexicographers check and annotate the output data independently of each other. Furthermore we also check how useful the remaining manually selected part of the data turns out to be when it comes to the concrete task of updating the dictionary, and study the relation between the initial annotations and the usefulness. The updates that we decide upon are either carried out immediately or listed as future tasks in the editorial process of keeping the dictionary up to date.

Once again, we use the corpus text collection divided by year, and now collect all the bigrams which do not appear at all in the first three, but appear with a certain frequency during the next 11 years. Our method is easily reproducible.

1. We calculate the statistically significant bigrams for the complete newswire corpus 2005 - 2018 (~ 512 million tokens), see [3.1] below for details;
2. We divide the corpus into 14 sub-corpora, one for each year;
3. We count the occurrences of the bigrams for each sub-corpus, i.e. each year, separately;
4. We make a dataset of all bigrams that meet the following two requirements:

a. The bigram does not occur in the first three years, 2005, 2006, and 2007, 3 being the lowest number of years that we felt would prevent accidental gaps in the distribution of the bigram.

   b. The bigram occurs at least 20 times in the following time period of 11 years,

   (--> frequency ~20/400 million = 0.00000005).

The output of the process is a dataset of 745 bigrams considered to be new in Danish. These bigrams are listed and used as input for the manual annotation task.

3.1 Calculating the statistically significant bigrams

In order to calculate the statistically significant bigrams we developed a small Python script using the Phrases module of the Gensim package (Řehůřek and Sojka, 2010; Řehůřek, 2020). We used the so-called *original scorer* algorithm based on the bigram scoring function developed by Mikolov et al. (2013) for calculating the bigrams.

The bigrams are calculated using the formula:

$$\text{score} = (\text{count}(w_i, w_j) - m) * \text{count}(\text{vocab}) / \text{count}(w_i)*\text{count}(w_j)$$

where count(w_i, w_j) is the frequency of the bigram, count(vocab) is the size of the vocabulary, count(w_i) is the frequency of the first word, count(w_j) is the frequency of the second word, and m is the minimum frequency of the bigrams.

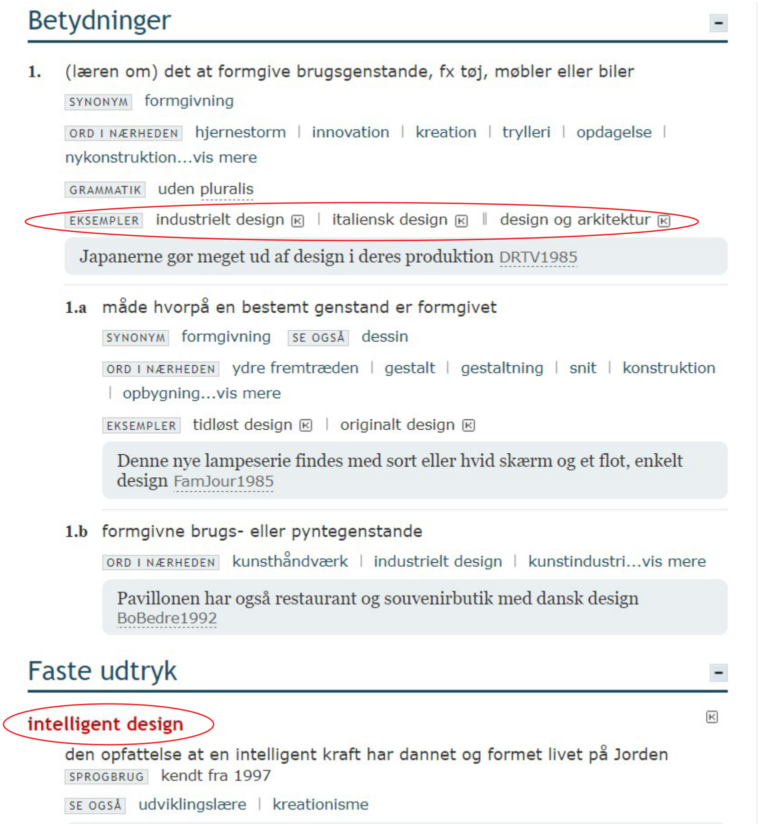
We chose the minimum frequency of bigrams to consider (m) to be 5 and we chose the threshold of 7 for significant bigrams. This threshold was chosen based on manual inspection in order to select only the most significant bigrams without letting too much noise into the dataset. This threshold removes arbitrary, ad-hoc bigrams like *nævne nogle* ('mention some', score 3.9) and *skal betale* ('must pay', score 1.2), but keeps wanted bigrams like *offentlig institution* ('public institution', score 8.8) and *monopolagtige tilstande* ('monopoly-like conditions', score 385.0). However, any fixed threshold must of course be expected to give some unfortunate results. In our case we find that some bigrams that are clearly non-collocational are included in the dataset (e.g. *stormer flyet*, 'raid the plane', score 7.3), and some excellent ones are excluded (e.g. *stor betydning*, 'great importance, score 6.8). We have not investigated the perfect threshold for this experiment, but it is clearly a task we wish to perform.

4 MANUAL ANNOTATION OF THE DATASET

We established the following five questions for the manual annotation task. The categories we chose are closely related to the type of information described in the dictionary which is to be updated with new semantic information.

1. Is the bigram likely to represent a new sense of one of the words, possibly in the form of a fixed expression, to be included in the dictionary?
2. Is it instead more likely to represent a new collocation, both words being transparent in sense?
3. Is the bigram (part of) a proper noun? For example the title of a Danish movie *Den skaldede frisør* (English title: Love is all you need), or a Danish tv-program *Den store bagedyst* (corresponding to the English program: The Great British Bake Off).
4. Is it a grammatical construction, for example *anno 2013* ('in the year 2013'), *arvelovens paragraf (X)* ('section (X) of the Inheritance Act').
5. Is it not at all relevant to include in the dictionary? *Eurozonens tredjestørste* ('the third largest of the Eurozone', *din smartphone* ('your smartphone').

The first 2 categories are particularly important in the semantic update task. In Figure 1, the DDO entry *design* is shown, and here we see how the two categories are used. Category 1 refers to defined senses in the dictionary which can be expressed as either a main sense or subsense (1., 1.a and 1.b in Figure 1), or in the form of a multiword unit where the lemma is included, initiated by the headline 'Faste udtryk' ('Fixed expressions') in the figure illustrated by *intelligent design* ('intelligent design'). Category 2 refers to the use of bigrams (or trigrams) as examples of how the word combines with other words in this sense, e.g. *industrielt design* ('industrial design') and *italiensk design* ('italian design'). We have chosen to call only these example bigrams 'collocations' in this paper. Others use the term 'collocations' differently. In a similar work, Pollak et al. (2019) use it in a broader sense, corresponding to the entire set of bigrams that they operate with, due to the fact that this only contain noun lemmas and their collocates. They operate with only bigrams containing noun lemmas in the dataset. Only their term 'collocationally new collocations', which is used to define one of the 7 core categories among their initially extracted collocations, correspond to what we call 'collocations'.



Betydninger -

1. (læren om) det at formgive brugsgenstande, fx tøj, møbler eller biler

SYNONYM formgivning

ORD I NÆRHEDEN hjernestorm | innovation | kreation | trylleri | opdagelse | nykonstruktion...vis mere

GRAMMATIK uden pluralis

EKSEMPLER industrielt design | italiensk design || design og arkitektur

Japanerne gør meget ud af design i deres produktion DRTV1985

1.a måde hvorpå en bestemt genstand er formgivet

SYNONYM formgivning SE OGSÅ dessin

ORD I NÆRHEDEN ydre fremtræden | gestalt | gestaltning | snit | konstruktion | opbygning...vis mere

EKSEMPLER tidløst design | originalt design

Denne nye lampeserie findes med sort eller hvid skærm og et flot, enkelt design FamJour1985

1.b formgivne brugs- eller pyntegenstande

ORD I NÆRHEDEN kunsthåndværk | industrielt design | kunstindustri...vis mere

Pavillonen har også restaurant og souvenirbutik med dansk design BoBedre1992

Faste udtryk -

intelligent design

den opfattelse at en intelligent kraft har dannet og formet livet på Jorden

SPROGBRUG kendt fra 1997

SE OGSÅ udviklingslære | kreationisme

Figure 1: The noun lemma *design* in DDO.

Two of us, both experienced lexicographers, annotated the output of 745 bigrams independently of one another with one of the 5 categories listed above. We both have a good knowledge of the lexical content of the DDO, and are very familiar with the task of updating the dictionary with new lemmas, senses etc. Table 1 shows an extract of one of the two independently annotated lists of bigrams.

Table 1: *The list of bigrams with frequency information and annotation, one annotator*

Bigram	Frequency	Annotation
amerikanske=internetgigant	23	2
amerikanske=jobmarked	32	5
amerikanske=medicinalselskab	57	5
amerikanske=whistleblower	74	5
analyserer=kulturelle	123	5
anbefalinger=fordeler	94	5
andengenerations=bioethanol	32	2
anno=2012	124	4
anno=2013	111	4
anno=2015	113	4
anno=2017	103	4
annoncerede=ordrer	26	5
antisemitiske=hændelser	25	2
anvendte=billedmateriale	422	5
arabiske=forårs	45	1
arabiske=opstande	21	2
arabiske=revolutioner	30	2
arktiske=kyststater	26	2
arktiske=stater	46	2

To compare our annotation task with similar work carried out by Pollak et al. (2019), they instead initially annotated a dataset manually (not double-annotated) in only three categories (p. 190): ‘non-relevant data’ (corresponding to 4 and 5 in our task), ‘proper words and abbreviations’ (corresponding to 3 in our task), and finally ‘core results’, which correspond to our categories 1 and 2. Afterwards the ‘core results’ in their study were annotated by two linguists (again not double-annotated) into 7 more specific categories, some of which are related to their specific interest in non-standard vocabulary and therefore not relevant to our case. But their 4 categories: ‘lexically’, ‘collocationally’, as well as ‘semantically new vocabulary’, and finally ‘terminology’, are all covered by the content of our first 2 categories: ‘new sense or fixed expression’ or 'new collocation’.

Pollak et al. (2019) apparently do not double-annotate the data, and as we shall see, the double annotation is in our case an important part of our method, and likewise plays an important role in the analysis and conclusions. Nor do Pollak et al. (2019) investigate to which degree the annotated data in each case entails an update in a practical lexicographic project, and what exact type of update that ends up being carried out on the basis of each bigram in the dictionary. Our study allows us to compare on the one hand the annotations and the inter-annotator agreement, on the other hand the different types of resulted updates, and to draw some conclusions based on the combinations.

The output of the annotation task that we carried out – two lists with 745 annotated bigrams – was subsequently compared in order to calculate the inter-annotator agreement. The results are discussed in the next subsection.

4.1 Inter-annotator agreement and relevant data

The overall inter-annotator agreement was 85% in the annotation task described above. However, there was almost 100% agreement between the two lexicographers on whether the data was unlikely to influence the semantic description in the DDO (the categories 3, 4 and 5, covering proper nouns, grammatical constructions or simply not relevant information to include in a dictionary). This data, 1/3 of the statistically significant bigrams, was therefore discarded as non-relevant for further lexicographic inspection, a share which corresponds roughly to the 37,4% of the extracted data which was found irrelevant in the Slovene study (Pollak et al., 2019, p. 191). The high inter-annotator agreement indicates that the task of discarding non-relevant bigrams from the automatically extracted list could probably have been carried out by just one experienced lexicographer.

The bigrams said to belong to either category 1 or 2 by both lexicographers, and thus likely to influence the semantic description of one of the lemmas (or both), constituted 482 bigrams, corresponding to 2/3 of all statistically significant bigrams. These were selected as highly relevant for a more thorough lexicographic inspection.

4.2 Frequency

Our choice of a frequency criteria of 0.00000005 seems suitable for our purpose of finding enough data to initiate a more systematic update process of the dictionary. A large part, namely more than 1/3 of the new bigrams, had a frequency between 20 and 30 (of 400 million tokens), and most of them, 3/4, had a frequency lower than or equal to 50. If the initial criteria on frequency had been raised from 20 to 50, we would only have obtained 1/4 of the relevant data that was found. It might even pay off to also check bigrams with a frequency between only 10 and 20 in the corpus, since more than a third of the relevant bigrams had 30 or less occurrences.

5 LEXICOGRAPHIC INSPECTION OF THE BIGRAMS AGREED UPON TO BE RELEVANT DATA

Figure 2 illustrates how the 745 statistically significant bigrams are overall distributed in non-relevant and relevant ones as described above and, maybe more importantly, how the relevant 2/3 (482 bigrams) are further divided into three groups: two groups with those where the lexicographers agreed upon the type of semantic update (both chose category 1, or both chose category 2) and one where they disagreed (the one chose category 1, the other chose category 2), or put differently, agreed upon it to be either category 1 or 2 (and not any of the categories 3, 4 or 5).

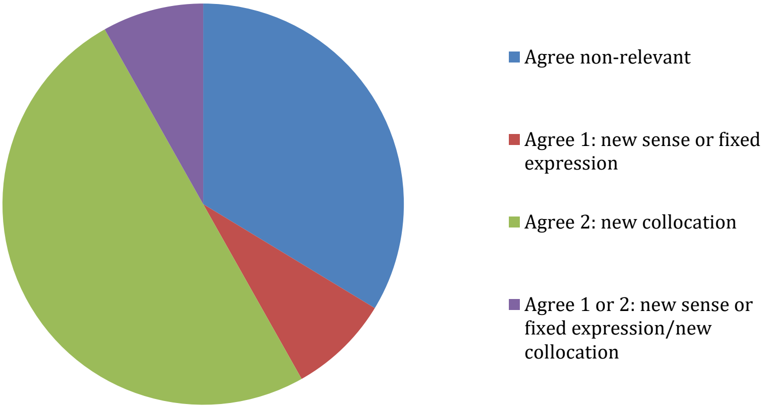


Figure 2: Double annotation of 745 statistically significant bigrams results in 4 groups: one with bigrams agreed upon as being non-relevant, one with bigrams agreed upon to represent 1) a new sense or fixed expression, one with bigrams agreed upon to represent 2) a new collocation, and finally one where the one annotator chose 1) new sense or fixed expression, and the other chose 2) new collocation.

By dividing the relevant bigrams in this way we obtain a distinction between the relatively clear cases (the first two groups where the annotators agreed upon the type of update) in opposition to the more unclear, albeit relevant cases (the third group where the annotators disagreed on the type of update). Interesting data concerning sense change tends to hide in the unclear data, as we shall see in section 6.3.

Our next step was to thoroughly inspect the bigrams from all three groups with the purpose of updating one or maybe even both lemmas in the dictionary with new semantic information. As an example, the multiword expression *fri fagskole* ('free vocational school', a new type of educational institution in Denmark) was added to the noun entry of *fagskole* ('vocational school') based on the bigram *frie fagskoler* ('free vocational schools'). The collocation *streame musik* ('to stream music') was inserted in the verb entry *streame* ('to stream') based on the identical bigram *streame musik*, and the collocation *nordisk køkken* ('Nordic cuisine') was added to the noun entry of *køkken* ('cuisine') based on the bigram *nordiske køkkens* (genitiv: 'of the Nordic cuisine').

It turned out that the updates would not only consist in a new sense, fixed expression or collocation, but also a slightly changed definition, or an added citation illustrating the bigram. In some cases the lemma was even updated in more ways than one, e.g. the bigram *intelligente løsninger* ('intelligent solutions') entailed both a new collocation as well as a slightly changed definition in the adjective entry *intelligent*, which now includes the new digital and computerized aspect of the sense.

Other bigrams turned out to be of less relevance than originally expected during the initial annotation task when they were more thoroughly inspected. E.g. the bigrams *forbyde burkaer* ('to ban burkas', reflecting a political debate) and *levende myrer* ('live ants', a much debated dish at the famous Danish restaurant, Noma) did not entail any revision of entries in the dictionary, estimated to be connected to very specific former events, and therefore, from a linguistic and lexicographic point of view, less relevant to include in the DDO today.

After having closely studied 189 bigrams and the corresponding two lemmas in the dictionary, we ended up deciding upon 103 semantic updates to be carried out in the dictionary. However, 300 bigrams from the collocation group have not yet been thoroughly analysed, but based on our studies of 1/5 of the group, we estimate the total amount of bigrams leading to an update to be approx. 41% of all the bigrams annotated to be relevant (category 1 or 2), and thereby 27% of the initial dataset of automatically extracted and calculated bigrams. This will be discussed further in the next section, where we will study the relation between the annotations carried out and the resulting types of updates, and draw conclusions on how to profit in more than one way from the double annotation of the bigrams.

6 THE RELATION BETWEEN TYPE OF ANNOTATION AND TYPE OF RESULTING UPDATE IN THE DICTIONARY

In Table 2, the number of updates (some of which are not yet carried out but listed as future editorial tasks), are presented in relation to the annotated data.

Table 2: *Bigrams divided into three groups depending on inter-annotator agreement*

482 relevant bigrams (of 745 statistically significant bigrams)	**Agree 1**: 55 bigrams. Both annotators agree: new sense or fixed expression	**Agree 2**: 367 bigrams. Both annotators agree: collocation	**Agree 1 or 2**: 60 bigrams One annotator: collocation Another annotator: new sense or fixed expression
Number leading to update	All inspected 49 lead to update	1/5 inspected (a sample of 74 bigrams) 24 lead to update (estimate full set: ~120)	All inspected 30 lead to update

Note. For each group, the number of bigrams leading to an update is given.

The same data is illustrated in Figure 3. When at least one of the annotators estimate the bigram to represent a new sense or new fixed expression, the data very often turns out to be useful in the process of updating previously described lexicographical vocabulary with new semantic information, as illustrated by the first and last columns.

Furthermore, and perhaps quite surprisingly, Figure 3 also clearly shows that when both annotators agree that a bigram constitutes a new collocation, the bigram quite often does not result in any update at all.

Apart from studying the amount of updates made up by the bigrams of each annotation group, it is also interesting to find out what kind of updates the

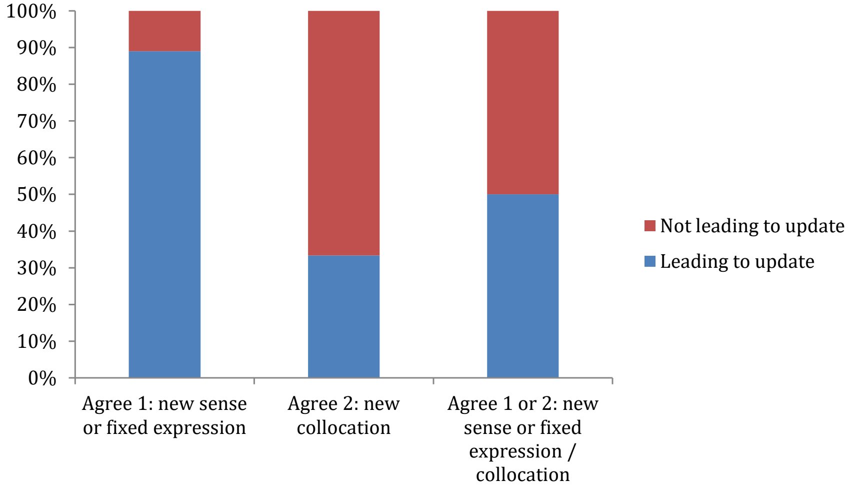


Figure 3: The figure illustrates how often the each of the three groups of relevant bigrams contained data which was useful in the task of updating the dictionary.

three different groups typically entail. Table 3 presents the number of specific updates in relation to the type of annotation.

Table 3: *Bigrams leading to updates and the types of updates that they entailed related to annotations*

Type of annotation leading to update → **Type of update**	**Agree 1**: Both annotators: new sense or fixed expression = 49	**Agree 2**: Both annotators: collocation = 24 of sample (estimation full set ~ 120)	**Agree 1 or 2**: One annotator: collocation. The other annotator: new sense or fixed expression = 30	**Estimated total number of updates** = 200
new lemma	22	2 (full set ~10)	2	34
fixed expression	19	0	8	27
new sense	1	3 (full group ~15)	7	23
changed definition	3	0	4	7
collocation	4	11 (full group ~ 55)	10	69
new citation	0	8 (full group ~40)	0	40

Note. The table also presents the estimated total number of updates entailed by the extracted dataset of bigrams.

We also estimate how many updates the dataset will lead to when the total set of annotated data is thoroughly studied. Around 27% of the automatically extracted bigrams lead to an update, which constitutes around 41% of the bigrams annotated as relevant for the semantic revision of the dictionary by both lexicographers. A little over 1/3 of the updates take the form of a new collocation in the dictionary, 1/4 take the form of a new senses or fixed expression, equally distributed. 1/5 is in the form of new citations, and almost 1/5 are new lemmas. See Figure 4.

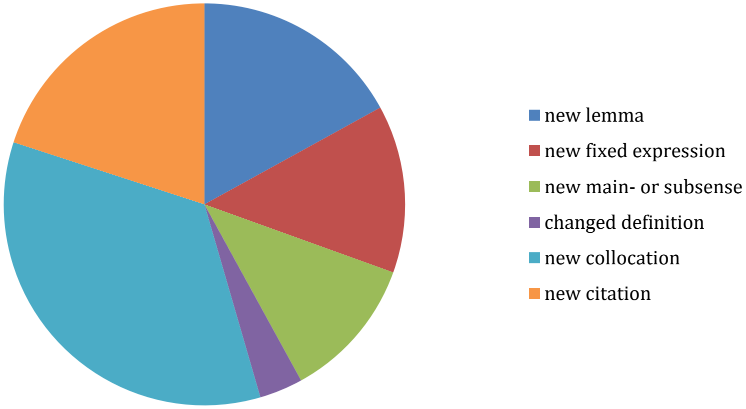


Figure 4: The share of the different types of updates entailed by the information on extracted bigrams.

In the next 3 subsections, we will go into detail with the data from each group.

6.1 Agree 1: Both annotators agree that it is a new sense, maybe in the form of a fixed expression

The two lexicographers agreed that a rather small, but valuable part of the semantically relevant bigrams represented a new sense or fixed expression. Here we find the most useful data when it comes to updating the already included lemmas in the dictionary, since almost all of it leads to revisions when the bigrams and the two corresponding dictionary entries are thoroughly inspected. See Figure 5.

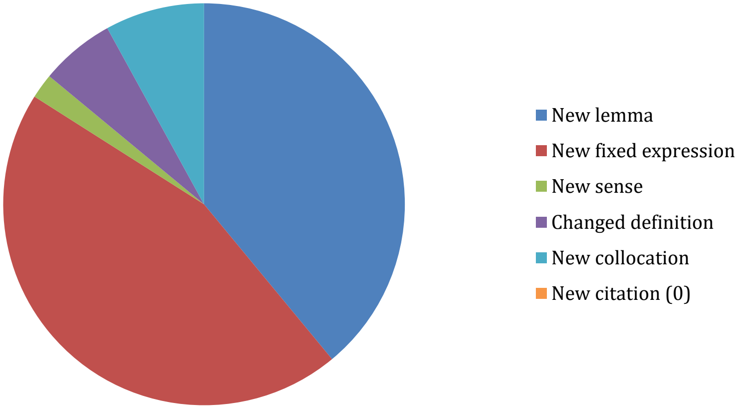


Figure 5: The distribution of different types of semantic updates entailed by the group of bigrams agreed to be a new sense or fixed expression by the two annotators.

Somewhat surprisingly, almost half turned out to constitute new lemmas based on an English multiword expression (e.g. *urban farming*, *augmented reality*). Danish neologisms are highly influenced by English, and loans from multiword expressions are often written in one word when they are included in Danish dictionaries, due to Danish spelling rules (*street food* → *streetfood*, *game changer* → *gamechanger*), if not, simply constituting a lemma entry spelled in two word. Pollak et al. (2019, p. 192) also deal with such loan words from English.

A substantial part of the bigrams in the group leads to a new fixed expression in the dictionary as foreseen by the annotators. In contrast to this, only very few led to the addition of a new main sense or subsense. More frequently they led to a change in existing definitions of the lemmas so that they now include the new phenomena described by the bigram. This was the case of the adjective *præhospital* 'prehospital' (based on the bigram *regionens præhospitale*), and *funktionel* ('functional'), based on the bigram *funktionelle lidelser* ('functional diseases'), see also other examples and a comparison with Cook et al. (2013) in section 7. Another rather small part led to new collocations in the entries. It is worth noticing that only among the bigrams in this group do we find the cases where the semantic information they represent had already been included in the dictionary, discovered during recent editorial work carried, for example due to user suggestions. In fact this goes for 12% of the updates, and most of them are fixed expressions which apparently attract the attention to a much higher extent than new senses and collocations.

6.2 Agree 2, inter-annotator agreement: collocations

Now we turn to the other part of the relevant bigrams in which the type of update was agreed upon by the two lexicographers, in this case judged to be new collocations by both. This part constitutes the largest group of the relevant data by far, namely ¾ (367 bigrams), and we have not inspected all of them yet. Here we find bigrams like *tørrede tranebær* ('dried cranberries'), *syriske borgerkrig* ('Syrian civil war'), *klimatiske udfordringer* ('climate challenges'), and *brystforstørrende operation* ('breast enlargement surgery'). In our investigation, we have previously only studied one fifth (74 bigrams) in detail, however we estimate this to be a sufficient number to enable us to draw some conclusions. We have compared them with the current lexical description of the two lemmas in the dictionary and also studied the occurrences in the corpora. As seen in Figure 5 above, only one third of the studied ones lead to an update of the dictionary. Many of them turn out to be very topical, time-limited and related to specific political or economic events in recent years. Therefore they are discarded in the final analysis and not integrated in the dictionary. One example of this is the bigram *amerikanske droneangreb* ('American drone strikes').

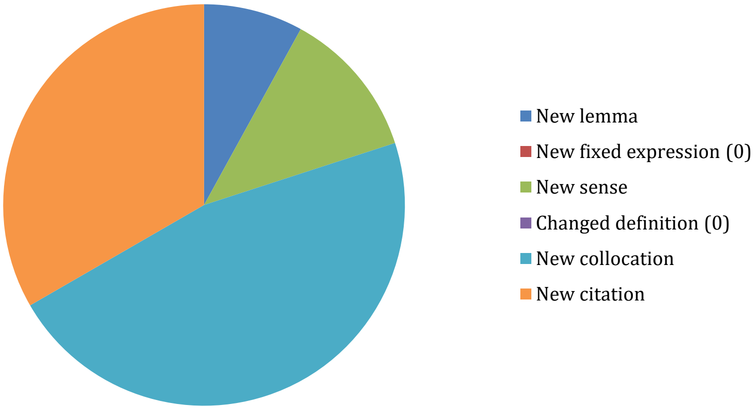


Figure 6: The distribution of updates entailed by the group of bigrams agreed to be collocations (category 2) by the two annotators.

Figure 6 illustrates how those of the category 2 bigrams that did result in an update are distributed when they are to be implemented in the dictionary. Almost half of them are added in the form of a collocation as also foreseen by both lexicographers, i.e. *trådløs opladning* ('wireless charging') which has been added to the adjective *trådløs*, *politiets vagtchef* ('police officer on call') which has been added to the noun *vagtchef* ('officer on call'), *ulovlig overvågning* ('illegal surveillance') which has been added to the noun *overvågning (*'surveillance'), and *kriseramte banker* ('crisis-stricken banks') which has been added to the adjective *kriseramt* ('crisis-stricken'). But Figure 6 also reveals that quite a lot of the bigrams that were estimated to be collocations in the first place instead have led to the adding of a new citation representing the bigram. It is worth noticing that only this group of bigrams (agreed upon to be collocations by both lexicographers) leads to this type of update in the dictionary. This suggests the future use of the same method in the task of updating citations in the dictionary, as a supplement to the criteria we use at the moment where we only look at entries with old citations from specific magazines. Another interesting fact about the updates based on the collocation group is that none of the data had already been discovered and included in the dictionary by other editors in the period since the bigrams were extracted for our experiments, indicating that this type of information, which is in fact highly needed in order to keep the dictionary content up to date at a more general level, would probably have been overlooked without the statistical investigation of bigrams.

However, the group of collocations also contains the highest amount of inapplicable data. It contains a lot of time-limited bigrams which according to the editorial guidelines of the DDO are not relevant to include in the dictionary. This is due to the fact that we are dealing with bigrams extracted mainly from newspapers. From a structural point of view, they are of course typical collocations: adjective + noun, verb + object etc., which is also why the two lexicographers easily agreed upon their status as such at first hand, but from a more pragmatic point of view they are not, and we should probably have been aware of this problem from the beginning. We can also conclude that very few bigrams in this group led the lexicographers on the track of new senses or new lemmas. One rare example is the loanword *big data* based on the English multiword expression. The lemma *data* is already part of the DDO which is why both lexicographers annotated it as a new collocation. However, since it is a term and a direct new loan pronounced in English it has instead to be included at lemma level in the dictionary.

6.3 Agree 1 or 2: inter-annotator disagreement whether it is a collocation or rather a new sense, maybe in the form of a fixed expression

The third and last part of the data selected for further lexicographic inspection consists of 60 bigrams that the two lexicographers agreed to be highly relevant. They disagreed, however, upon how to include them in the dictionary structure. While one annotator estimated that the bigram was most likely to represent a new sense or fixed expression, the other believed that it was more likely to represent a new collocation. In fact, only half of the bigrams in this group entailed a dictionary update. See Figure 7 for the distribution of the different types of updates.

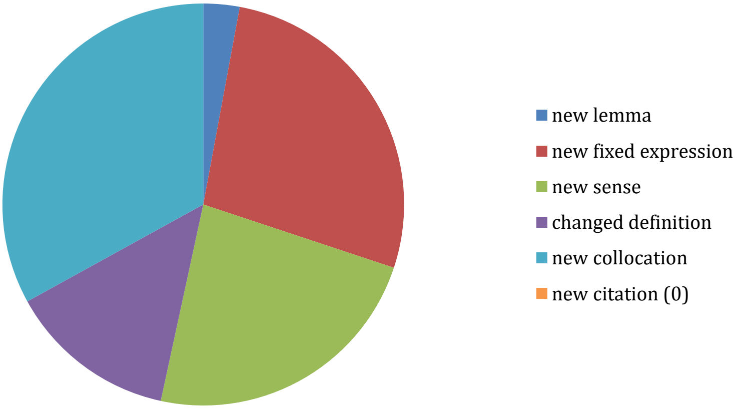


Figure 7: The distribution of updates entailed by the bigrams agreed to be relevant. However the annotators disagreed upon whether the bigram represented a new sense or fixed expression, or rather a collocation.

The vast majority of those which entailed an update did so in the form that was suggested by either one or the other annotator, more or less equally distributed. For the first time, we find quite a lot of new senses and not only fixed expressions. One third of the bigrams were included as collocations (e.g. *bæredygtig omstilling* ('sustainable conversion', *mentalt helbred* ('mental health')), almost another third as a fixed expression (*bibelske dimensioner* ('biblical proportions'), *pædagogiske assistenter* ('teaching assistents', new job title)), and, particularly interesting, one quarter in the form of a new main sense or subsense. E.g. the new subsense of the noun *boble* ('bubble') discovered from the bigram *glas bobler* (lit. 'glass of bubbles' – i.e. 'a glass of sparkling wine, e.g. champagne') was included in the dictionary, and the adjective *mobil* ('mobile') is planned to be provided with a new sense triggered by the bigrams *mobile bredbånd* and *mobilt internet* ('broadband/internet via a cellular phone').

Some of the bigrams will result in several changes. In the case of the new concept *selvkørende bil* ('self-driving car') which is also a part of the new data described in Pollak et al. (2019, p. 193), the definition of the adjective entry *selvkørende* needs to be changed in DDO, as does the entry of *bil* ('car'). The entry will be extended with a new fixed expression with its own definition.

It is worth noticing that this group of bigrams is the one reveals the largest amount of new senses by far. Several bigrams lead to the inclusion of a new main sense or subsense in the dictionary. Many also entail the need of a changed definition for one of the lemmas. For instance, a revision of the definition of *digital* ('digital') is needed due to the bigram *digital dannelse* ('digital code of conduct/digital education'), likewise a revision of the definition of *cannabis* ('cannabis; marijuana') was needed due to the bigram *medicinsk cannabis* ('medicinal marijuana'). We also found one new lemma in the group, the adjective *æresrelateret* ('honor-related'), due to the bigram *æresrelaterede konflikter* ('honor-related conflicts'). This lemma would also be discovered by single lemma extraction methods, but since it very often occurs together with *konflikter* in our data, this should be added as collocational information when the new lemma is included and edited.

Among the discarded data in the group were bigrams that had only been frequent for a short period of time (based on the study of the occurrences in our corpus), others were considered to be terminology which is not suitable for inclusion in the dictionary. As in the case of the agreed collocations, it's worth noticing that no lexical information discovered from our study of this group of bigrams had been registered in the dictionary by other editors since the data was extracted, and it would probably have been hard to discover without the use of statistical methods.

6.4 Conclusions on annotation and resulting updates

Our computational measure of the appearance of new bigrams in homogenous newswire corpora combined with double annotations of the output dataset and the entailed updates of the dictionary allow us to draw a number of conclusions.

6.4.1 How useful was the automatically calculated dataset?

First of all, we can conclude that quite a lot, i.e. approx. 1/4, of the automatically extracted dataset leads (or will lead) to a resulting update in the dictionary, while 3/4 do not. In comparison, Pollak et al. (2019) find a little less "lexically, collocationally, or semantically new data that can be considered in the process of updating existing lexical resources for Slovene" (p. 197), namely 21.6%. The initial annotation by two lexicographers made it possible to discard many bigrams in the extracted dataset in an efficient and not very time-consuming way. The data that the lexicographers selected as most likely to be relevant turned out to be useful when more thoroughly inspected and compared to the content of the dictionary entries in almost half of the cases. Had the initial annotation task been carried out on the basis of more detailed and elaborated guidelines, we could probably have avoided even more 'noise' (bigrams not leading to any updates after all), for example the many time-limited bigrams. The automatic extraction of the bigrams can maybe also be tuned in a way so that such time-limited data is better avoided in the first place, and not even included in the output dataset. Pollak et al. (2019) also propose that the automatic extraction procedure should include language recognition in the preprocessing step in order to identify and remove the English bigrams from the list. However, this would entail that several new loan words would not have been discovered and included in the DDO.

6.4.2 New lemmas

We found far more lemma candidates in the dataset than expected, namely 4%, due to the fact that many English multiword expressions are to be integrated in the dictionary at lemma level. This is in line with the results of Pollak et al. (2019).

6.4.3 Fixed expressions

A little over 4% of the initial dataset ended up being included in the dictionary in the form of fixed expressions. They constitute 14% of the updates carried out. From our investigations, we can see that when a bigram is recognized by two lexicographers as a fixed expression, it very often holds true, and it almost surely will influence the semantic description of one or both lemmas that are part of the bigram in one way or another. Very few bigrams that had been annotated as a fixed expression by both lexicographers led to no update at all, so if you want to make sure you find relevant data for the updating task of a dictionary, then this a way to go. Furthermore we can conclude that when two lexicographers agree that a bigram is not a fixed expression but rather a collocation, we can also be sure that it is not. Fixed expressions also seem to be the easiest to discover without applying any systematic method, since around 1/6 of them had already recently been included in the dictionary.

6.4.4 New main senses and subsenses

We found quite a lot of new senses via the dataset. Around 3% of the automatically extracted bigrams led us to this information, and among the annotated relevant data one in every 20 bigrams revealed a new sense. Pollak et al. (2019) find a bit more (4.9% of the extracted data), but they state that many are found in non-standard colloquial language (p. 193), which might explain the higher amount – this type of language is not included in our corpus texts. Due to the method of double annotation, we discovered that new senses tend to hide between the more ambiguous data where the lexicographer is not so sure whether the bigram represents a sense or a fixed expression that needs to be explained to the dictionary user, or whether it is rather a collocation with transparent meanings of both words. However, new senses can also be found among bigrams which when presented to the lexicographers in the first place, were estimated to be merely collocations of already included senses in the dictionary. In contrast, new fixed expressions were in fact found only when both annotators estimated the bigram to be either a new sense or a fixed expression.

6.4.5 Collocations

Bigrams resulting in updates in the form of a collocation constitute 9% of the extracted data, and almost half of those that were annotated as category 2 by both lexicographers, also turned out to lead to a new collocation in the dictionary. Thereby they constitute the cases in which inter-annotator agreement is very high and at the same time they most often corresponded to the type of resulting update Pollak et al. (2019) find a higher percentage of ‘collocationally new collocations' in their extracted data (13.3%, p. 193), but the many collocations that we chose not to include in the dictionary after a more thorough investigation probably explains the difference. In contrast to the DDO update guidelines, Pollak et al. (2019) propose that such data should not necessarily be left out of dictionaries: "trending vocabulary that is often bound to specific political and social events", should instead be included in digital dictionaries. They advocate for "a faster and more fluid lexicography that focuses not only on the stable and established, but also on the changeable and variable aspects of language – which is where language users often need assistance" (p. 200). We find that the inclusion of such data would probably entail an ongoing and maybe time-consuming control with the already lexicographically described vocabulary in the DDO in order to be sure to avoid lexical information that has become outdated.

Since two thirds of the collocation bigrams did not lead to any updates, we can conclude that when two lexicographers independently of one another agree that a bigram is a collocation, it is much less likely to represent useful data for the semantic update of a dictionary than if at least one of them consider it a new sense or fixed expression as described above.

6.4.5 Citations

Many collocations were included in the form of a citation when the data was thoroughly inspected, and we are in fact pleased to have discovered a more systematic way of updating this part of the dictionary information across lemmas.

7 RESULTS COMPARED WITH PREVIOUS RESEARCH

In this section we compare our study with a similar project presented by Cook et al. (2013). They used a reference corpus from 1995 and a focus corpus from 2008 to identify new elements to be included in an English learner's dictionary (Macmillan). In their paper, they use three categories:

1. the uninteresting findings, which are mostly due to the many news stories in the corpus; certain items exhibit a sudden spike and then they disappear and never turn up again; one example of this is the word *junta* referring to the regime in Myanmar that would not accept humanitarian help from the outside world after a disastrous cyclone that caused

many deaths; another example is the word *candy* that popped up because some Chinese candy had been contaminated with melamine;

2. much more interesting are the cases where a dictionary entry should be changed in some way, it needs 'tweaking'; for instance the existing entry for *cleric*, which only referred to clerics typical of the Church of England, but in the 2008 corpus, clerics are often Muslim and this should be reflected in the entry; the example *video* is obvious: in the 1990s a video would be a video tape of the VHS type, but nowadays it is typically a digital recording of images and sounds distributed via online media;
3. the third category is cases where new senses should be included in specific entries in the dictionary, for instance the verb *to search* (= 'do a web search'), and *text* as in *text messaging*, *send someone a text* or *text someone*, a technology that was not yet available in 1995.

Let us take a look at our findings using more or less the same categories as Cook et al. (2013) We have a high number of irrelevant findings, which we first categorized as collocations without deciding if they would lead to an actual change in the entries for the two words (cf. Section 6.2). The high amount of newspaper texts in our corpus accounts for findings related to specific events and political discussions; *tibetansk flag* ('Tibetan flag') for instance refers to a demonstration where Danish police unlawfully removed a Tibetan flag so that it would not be seen by the Chinese president who was visiting Copenhagen.

As is the case for Cook et al. (2013) we have changed (tweaked) several dictionary entries, for instance *cannabis*, where the collocation *medicinsk cannabis* ('medicinal marijuana') shows that cannabis may also be used for medical purposes nowadays; or *intelligente løsninger* ('intelligent solutions'), which indicates a new nuance in the meaning of *intelligent* involving digital functions and computers - so this has been added to the definition (cf. Section 6.3).

The entirely new senses include the word *digital*; the current entry describes the situation in the 1980s and 1990s when you would distinguish between a digital watch and an analogue one; of course, this is not up to date and the entry *digital* needs a new sense that will account for collocations like *digitale indfødte* ('digital natives') and *digital mail*.

A fourth category not mentioned by Cook et al. (2013) is new fixed expressions. As mentioned in section 5.4 this category is very salient in the list of bigrams and we have decided to include several of these. The most significant one is probably *sociale medier* ('social media'), which had already been discovered by other methods and added to the dictionary; other interesting examples are *assisteret reproduktion* ('assisted reproduction'), *cirkulær økonomi* ('circular economy') and *brændende platform* ('burning platform', i.e. a difficult situation that urgently needs taking care of); the expression refers to a fire on an oil platform in 1988 which resulted in many deaths.

A fifth category contains new lemma candidates, mostly of English origin; many of the English bigrams in the list may be included in our dictionary, either as headwords consisting of two words (*pulled pork*) or as a solid compound like *komfortzone* ('comfort zone' in English); even a pragmatic phrase like *oh, my god* and its abbreviation *omg* are lemma candidates if you take into account how common the phrase has become in everyday Danish, and the same goes for other English phrases that have been included in the DDO in recent years, such as *you name it*, *whatever*, and *take it or leave it*.

8 FINAL CONCLUSIONS AND PERSPECTIVES

In this final section we make a brief evaluation of our study: what are the overall pros and cons of this method and of our approach? On the upside, it provides the editors of the DDO with very useful input for updating senses, definitions, collocations, etc. In fact, the editors are so happy with it that the plan is to repeat the bigram calculation regularly, for instance every three years. It is also very encouraging that the material supports updates that have already been made - quite reassuring for a corpus-based dictionary. The material is a necessary supplement to other methods used by the dictionary editors to keep track of lexical and semantic change, like user suggestions, other corpus-linguistic data and good old editorial observations since it guarantees a systematic check across the entire vocabulary.

A drawback, of course, is that manual filtering is indispensable, but the good news is that one experienced lexicographer can fulfill the first phase (discarding non-relevant bigrams), whereas it takes two (or more) lexicographers to annotate the rest reliably and eventually make the actual changes in the dictionary. An important lesson from the experience is that a very large proportion of the bigrams consists of topical (time-limited) examples, which is due to the composition of the corpus (mostly newspaper material). Other types of corpus texts are too scarce for the time being, and this is a task that the dictionary staff intends to work on in the future, keeping in mind, however, that a homogeneous data type as well as an even distribution of text types over time is absolutely necessary in order to obtain good results with the statistical method that we have described in this paper.

Acknowledgments

The authors would like to thank the anonymous reviewers for their suggestions and careful reading of the manuscript. We would also like to thank our colleague Jonas Jensen for useful feedback and for proofreading the article.

REFERENCES

Dictionaries

DDO = *Den Danske Ordbog* [The Danish Dictionary]. Retrieved from https://ordnet.dk/ddo (17. 2. 2020)

Macmillan = *Macmillan English Dictionary*. Retrieved from https://www.macmillandictionary.com/ (17. 2. 2020)

Corpora

Korpus.dsl.dk = *Language Technology Resources for Danish*. Retrieved from https://korpus.dsl.dk/resources.html

Other

Cook, P., Lau, J. H., Rundell, M., McCarthy, D., & Baldwin, T. (2013). A lexicographic appraisal of an automatic approach for detecting new word-senses. In *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference* (pp. 49–65). Tallinn, Estonia.

Lorentzen, H. (2004). The Danish Dictionary at large: Presentation, Problems and Perspectives. In G. Williams & S. Vessier (Eds.), *Proceedings of the 11th EURALEX International Congress* (pp. 285–294). Lorient, France.

Mikolov, T., Sutskever, I, Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In *Advances in neural information processing systems 26*. Retrieved from https://arxiv.org/abs/1310.4546

Norling-Christensen, O., & Asmussen, J. (1998). The Corpus of The Danish Dictionary. *Lexikos (Afrilex Series) 8*, 223–242.

Pollak, S., Gantar, P., & Arhar Holdt, Š. (2019). What's New on the Internetz? Extraction and Lexical Categorization of Collocations in Computer-Mediated Slovene. In *International Journal of Lexicography, 32*(2), 184–206.

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

Řehůřek, R. (2020). *models.phrases – Phrase (collocation) detection*. Retrieved from https://radimrehurek.com/gensim/models/phrases.html (17. 2. 2020)

Tahmasebi, N., Borin, L., & Jatowt, A. (2018). Survey of Computational Approaches to Lexical Semantic Change [Preprint at ArXiv 2018]. Retrieved from https://arxiv.org/abs/1811.06278

Traugott, E. C. (2017). *Semantic Change*. Oxford Research Encyclopedias [Online publication]. doi: 10.1093/acrefore/9780199384655.013.323

POSODABLJANJE SLOVARJA: PREPOZNAVANJE SEMANTIČNIH SPREMEMB NA PODLAGI DIAHRONIH SPREMEMB BIGRAMOV

V prispevku preizkusimo metodo sistematičnega posodabljanja Danskega enojezičnega slovarja z novimi semantičnimi podatki o obstoječih lemah. Metoda temelji na hipotezi, da so diahrone spremembe bigramov v korpusnih podatkih lahko pokazatelj sprememb pomena ene od besed v bigramu. Pri metodi kombiniramo korpusno statistiko z ročnim označevanjem. V prvem koraku izmerimo kolokacijske spremembe v homogenem korpusu novic za 14-letno obdobje (2005 do 2018), tako da izračunamo vse statistično pomembne bigrame. Te bigrame potem preverimo v novi različici korpusa, razdeljenega na podkorpuse, pri čemer vsak podkorpus zajema obdobje enega leta. Nato izluščimo vse bigrame, ki se nikoli ne pojavijo v prvih treh letih, se pa pojavijo vsaj 20-krat v naslednjih 11 letih. Na podlagi tega postopka dobljenih 745 bigramov, ki jih obravnavamo kot potencialno nove v danskem jeziku, označita dva označevalca. Bigrami so glede na rezultate označevanja in ujemanje označevalcev bodisi izločeni bodisi razvrščeni v skupine glede na relevantnost za nadaljnjo obravnavo. Sledi temeljitejša leksikografska analiza, s katero določimo, do kakšne mere gre za nove pomene besed in posledično potrebo po spremembi pomenske členitve pri vsaj eni od besed v bigramu. Poleg tega analiziramo tudi povezavo med potrebnimi popravki, oznakami in odstotkom ujemanja označevalcev. V zadnjem delu prispevka primerjamo slovarske posodobitve s pristopom, ki so ga izvedli Cook idr. (2013), in podamo razmisleke o tem, ali tovrstna metoda lahko predstavlja doslednejše popravljanje in dopolnjevanje slovarskih gesel.

Ključne besede: korpusna statistika, bigrami, posodabljanje slovarja, semantične spremembe, danski jezik