Trawling the corpus for the overlooked lemmas

Nathalie Hau Sørensen[1], Nicolai Hartvig Sørensen[1], Kirsten Lundholm Appel[1], Sanni Nimb[1]

[1]The Society for Danish Language and Literature, Chr. Brygge 1, 1219 Copenhagen K
E-mail: nats@dsl.dk, ka@dsl.dk, nhs@dsl.dk, sn@dsl.dk

Abstract

Lemma selection is a significant part of lexicographic work, also in the case of the online Danish Dictionary (DDO), a corpus-based monolingual dictionary updated twice a year based on the prior identification of good lemma candidates by means of statistical corpus methods as well as introspection. All low frequent word forms have until now been discarded in the statistical process, but in this paper, we present a method to also identify lemma candidates among these. Our hypothesis is that some words are too inconspicuously mundane to be noticed by introspection and at the same time so infrequent that they are overlooked by statistical measures. The method is based on different automatic measures of "lemmaness" by means of language models, character n-grams, statistical calculations and the development of a compound splitter based on information in the DDO. We evaluate the method by comparing the generated list with the lemmas included in the online DDO since 2005. Two trained DDO lexicographers furthermore evaluate words from the top as well as the bottom of the list. Though there is room for improvement, we find that our method identifies a large number of lemma candidates which otherwise would have been overlooked.

Keywords: Neology detection; lemma selection; low frequent words

1. Introduction

Lemma selection is a significant part of lexicographic work, also in the case of the online Danish Dictionary (DDO, Det Danske Sprog- og Litteraturselskab (2023a)), a corpus-based monolingual dictionary updated twice a year based on the prior identification of good lemma candidates primarily by means of statistical corpus methods. The corpus is extended monthly and consists of texts from the past 40 years, in total around 1.1 billion tokens. All low frequency word forms have until now been discarded in the automatic process of finding lemma candidates, but in this paper, we present a method to also identify low frequent ones among the discarded and noisy data. The method is based on different automatic measures of "lemmaness" by means of language models (word2vec, NER), character n-grams as well as statistical calculations and the development of a compound splitter based on lemma compound information registered in the DDO. We evaluate the method by comparing the resulting list of lemma candidates with the lemmas having been included in the online DDO since 2005. Two trained DDO lexicographers furthermore evaluate parts of the top as well as the bottom of the list.

The Danish Dictionary DDO was initially published as a printed dictionary 2003-2005, at that time describing the senses of 60,000 lemmas. Since 2009 the dictionary has been published online and today it describes more than 100,000 lemmas. The online publication gives us the opportunity to continuously update the content and thereby reflect the changes in the Danish vocabulary. Currently, the online DDO is updated twice a year. In Nimb

et al. (2020), we describe a corpus-based method of detecting new senses, new collocations as well as new fixed expressions of already included DDO lemmas; a method that has already resulted in many revised entries in the updates. The dictionary also publishes a substantial number of new lemmas in the updates, including both neologisms and words which were not included in the printed version due to space restrictions. Related to this part of the editorial work, we are interested in developing automatic methods to supplement our existing procedures of detecting lemma candidates. In spite of basing it on statistical corpus methods, we are aware that among the low frequency words in the corpus, some good candidates are still being overlooked, i.e. lemmas which become more and more relevant in the current phase of extending the DDO from 100,000 lemmas to up to 200,000 lemmas [1]. They hide among the large amount of noisy data which is discarded in our statistical methods, and which at a quick glance contains far more undesirable noise than good lemma candidates – meaning that the lexicographers would have to check the data manually word by word.

Lemma selection principles vary across different dictionary projects. In the case of the DDO, the selection is highly based on well-developed statistical corpus methods resulting in monthly lists of good candidates to choose from, supplemented by lemma suggestions from users, new words in the Danish orthographic dictionary *Retskrivningsordbogen* published by the Danish Language Council every year, and of course the editors' own notifications of relevant words not yet covered by the DDO, e.g. as a benefit of the editorial work with the Danish Thesaurus where lexical gaps in DDO were sometimes discovered (Lorentzen & Nimb, 2011). In any case, the overall criteria of lemma selection in the DDO project is always a certain representation in the DDO corpus. For words with a frequency lower than 50/1,000,000,000, not only corpus frequency but also other aspects should always be considered by the editor. The occurrences should appear in different texts published over a number of years, normally at least three years. In some cases, when a very low frequency lemma (i.e. a lemma occurring less than 10 times in the corpus) is used outside the domain of newswire, e.g. within specific domains, or in daily life, maybe also by specific age groups or in oral rather than written language, occurrences found by searching the internet are included.

The automatic method we suggest is inspired by the introspective judgments of low frequency words which have until now been randomly discovered by the DDO lexicographers. It aims at identifying lemma candidates in the large amount of noisy low frequency corpus word forms by measuring "lemmaness" based on a combination of NLP techniques. We evaluate the method by comparing the identified lemma candidates with the lemmas added to the DDO dictionary over the last 15 years, but also by introspective judgments of the results carried out by two experienced DDO lexicographers with the knowledge of the specific editorial principles of the DDO project.

1.1 Background

We know for a fact that low frequency lemmas are relevant to include in the DDO. The user log reveals that users do query low frequency words that are not yet in the dictionary (Trap-Jensen et al., 2014). When applying the DDO in a sense annotation task on 2000 sentences from Wikipedia in the ELEXIS project (Martelli et al., 2023; Pedersen et al.,

[1] the predecessor, the ODS dictionary published 1918-1955 contains around 220,000 Danish lemmas

2023), we also found a surprisingly high number of lemma candidates among the word forms that were not represented in the dictionary, even though they only occurred once in the text. Among these were also candidates that our normal detection procedures had not discovered. Our hypothesis is that some words are simply too inconspicuously mundane to be noticed by introspection and at the same time so infrequent that they are overlooked by statistical measures. We are also aware that our statistical corpus methods depend on the quality of the corpus. The ideal corpus contains a broad collection of different text genres. However, the DDO-corpus mainly contains newswire because Danish texts from other genres have turned out to be difficult to obtain due to copyright issues. In a corpus with a lion's share of newswire texts, some mundane words used in daily life may be underrepresented.

We have previously studied the relation between corpus frequencies and lexicographic relevance in DDO (Trap-Jensen et al., 2014). Where all words represented among the top 100,000 most frequent forms were indeed well established in the language and for sure relevant DDO-lemmas, it turned out that frequency was less useful as a criterion when it came to the identification of relevant lemmas among the rest of the corpus words. When examining Spanish neologisms with corpus frequency and perception surveys, Freixa & Torner (2020) also found that even though frequency is an important factor to determine the degree to which a word is institutionalised, there are other factors (e.g. loan words, transparency of derivations and compounds) in play when considering whether a word should be included in the dictionary.

1.2 Related work

In Halskov & Jarvad (2010) a previous attempt to automatically identify neologisms in Danish is described. Due to the lack of available NLP tools for Danish at the time, the method did not yield very good results seen from a DDO perspective. Some of the problematic areas were named entities and compounds. There is a very high amount of the latter in Danish, also among low frequency words and new lemmas. An analysis of neologisms in the DDO updates showed that 52% of the neologisms are in fact compounds (Trap-Jensen, 2020). It is a challenge to distinguish ad hoc compounds from the more established compounds of which the senses are relevant to describe in the DDO dictionary, not only automatically but also by introspection. However, in the experiment in Halskov & Jarvad (2010), the aim was primarily to identify simplex neologisms, which are more important to include in the Danish orthographic dictionary than easily spelled compounds.

A popular approach to neologism detection is the use of exclusion lists (i.e. list of words that are already in the dictionary or otherwise beneficial to remove from the investigated data). In this approach, a corpus is preprocessed before the exclusion list is used to remove the already included lemmas. A series of filters and postprocessing steps can then be applied to get a list of potential neologisms or lemma candidates. The *NeoCrawler* (Kerremans et al., 2012) removes noisy tokens by using character trigrams to calculate whether a token is a probable English word, and we apply a similar technique on the Danish data (see section 2.3.5).

The exclusion list approach relies on the quality of the pre- and post-processing steps. In a corpus, many of the unknown words (words not registered already in a dictionary) are not neologisms but instead named entities, spelling errors and derivatives. In Langemets

et al. (2020), which describes an experiment with detecting Estonian neologisms, they conclude that only 10% of the words in an automatically derived list of lemma candidates are in fact good candidates to include in the dictionary. The results can be explained by a high number of derivatives and semantically transparent compounds in the final candidate list as well as the poor quality of NLP tools for Estonian.

Alternatively, the exclusion list approach can be combined with machine learning. For instance, Falk et al. (2014) uses supervised machine learning to identify neologisms in a French newswire corpus. For each unknown word, they extract a range of features related to form, spelling, and theme. The result is a ranked list representing a word's probability of being a neologism according to a trained model. However, supervised machine learning approaches require that we have manually annotated data to train on which can be time-consuming to obtain. Since we use a simple weighted average, we avoid the need of manually annotated training data.

In this work, we present an automatic method for lemma selection based on both exclusion lists and a scoring mechanism that imitates the editorial principles of the DDO lemma selection of low frequency lemmas. The main focus is not only to detect neologisms in the traditional sense, but also to detect the overlooked lemmas in the entire DDO corpus with texts from 1982-2022, i.e. lemmas that could have been added to the corpus-based dictionary since the project was initiated in 1992. Like the approaches mentioned above, we use a series of preprocessing and filtering steps to remove the worst noise. We take a similar approach to Falk et al. (2014) by also extracting a range of features. Each feature corresponds to a post-processing step, however we do not remove candidates on the basis of only one of these. Instead, we calculate a combined score as a weighted average of each feature. The idea is that in the cases where a feature is not realised, or represents an error, another feature might balance the score.

In the next section we describe the method and the various steps involved in it, initiated by a description of the corpus that we use in our experiments. In section 3, we evaluate the results. Section 4 discusses some pitfalls of our approach, and finally we conclude in section 5.

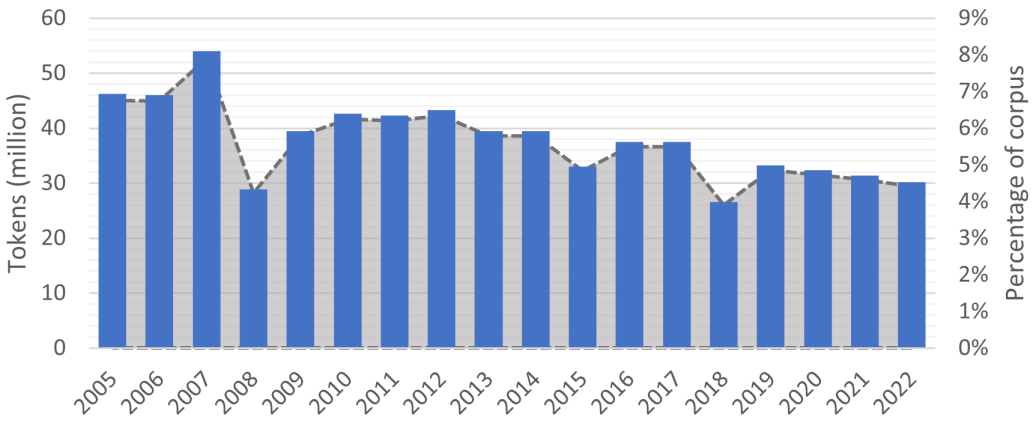

Figure 1: Distribution of tokens across the 18 years present in our corpus. Tokens are measured in millions.

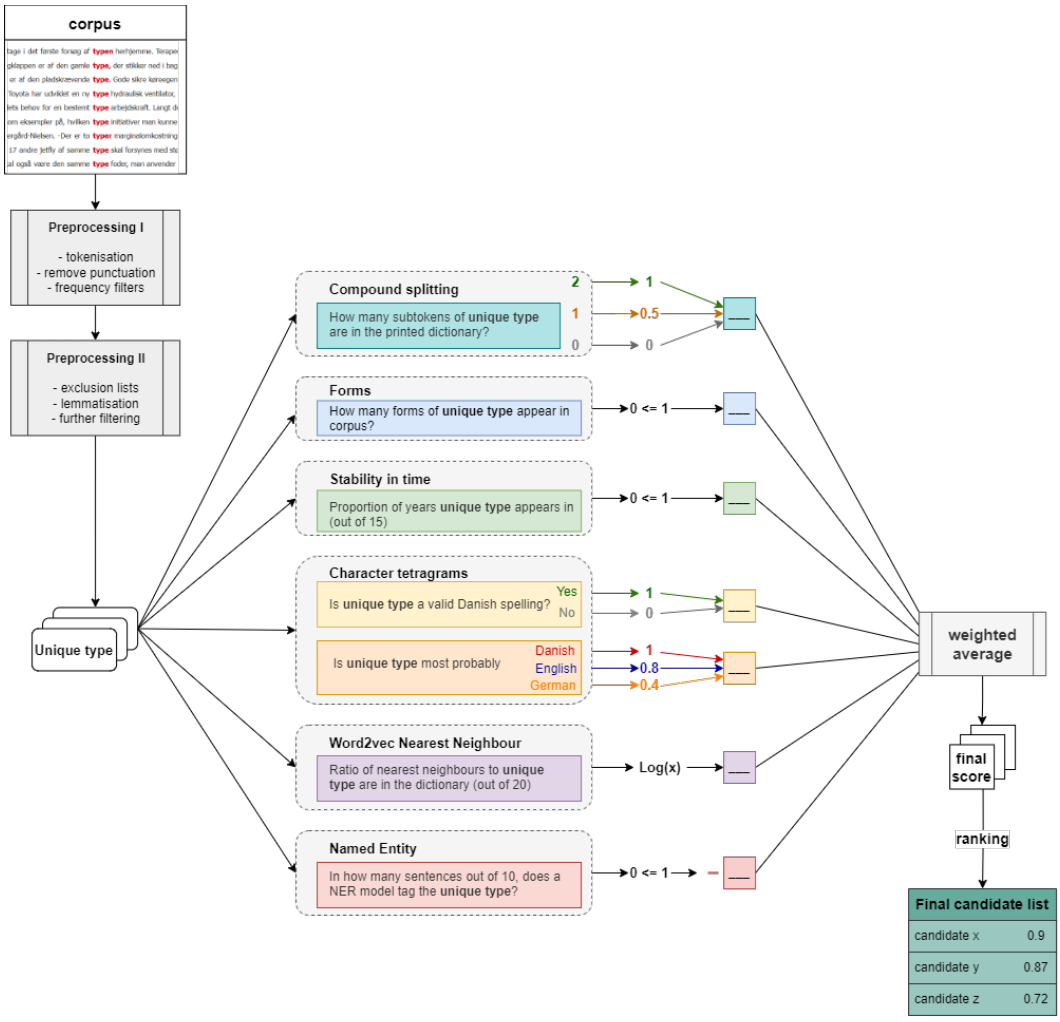

Figure 2: Overview of the automatic method from start to finish. The process begins with the corpus which is then preprocessed by two modules. Next, we extract features for each remaining word. Finally, the features are combined into a lemma score which the data is ranked after.

2. Methodology

Since the very beginning, the DDO dictionary project has been highly based on the access to high quality corpus data. The initial task in the project in 1991-1993 was to create the first part of today's DDO corpus specifically for the subsequent compilation of the dictionary in the years 1994-2005. At the time, the corpus contained 40 million running words of Danish produced between 1983 and 1992 and contained both written and spoken Danish from a wide range of media and genres (Norling-Christensen & Asmussen, 2012). It has since been extended with two similar batches from around 2000 and 2010 respectively and a batch of texts from Wikipedia in 2017. Since 2005 it has been extended with newswire texts every month. It currently holds roughly 1.1 billion running words.

In our main corpus the genre composition varies over time making word frequency comparisons across different years unreliable. We have therefore chosen a subset containing only one genre to eliminate this issue. For our current experiment we use a part of the corpus composed exclusively of newswire texts from 2005-2022. It contains a total of

683,277,792 running words, and the texts are roughly equally distributed across the time period ranging from 26,640,430 words in 2018 to 53,958,056 words in 2007. The distribution is shown in figure 1. A subdivision into years allows us to track changes over time, and the homogeneous nature of the corpus makes the different years directly comparable, as the text domain remains constant across the entire corpus.

The starting point of the automatic method is a division of the lemmas in the online DDO into two parts, one containing only the lemma entries from the first printed edition of the dictionary from 2003-2005, another one containing the lemma entries in the dictionary added from 2005 to now. In the experiment, we assume that we have no knowledge of the new lemmas. Instead, we use the information as a 'lemma selection gold standard' when we evaluate the results. Hereafter, we refer to the original 2003-2005 printed DDO edition with 60,000 lemma entries as *the printed dictionary* and refer to the current online edition containing 100,000 lemma entries as *the online dictionary*.

The method includes three main steps which are visualised in figure 2. In the first step, we apply a cleaning and filtering process to reduce the number of investigated tokens. In step two we assign scores to each string using different scoring mechanisms. Each score assignment aims at representing the information used by the lexicographer in the process of selecting lemma candidates. In the last step, the final score is calculated using a weighted average of the scores, and the list is sorted accordingly. The following subsections explain the individual steps in detail.

2.1 Preprocessing I

The first step is to extract all the word forms from the corpus which will constitute the initial list of candidates. Thus, we convert the corpus into a raw text format, and the text is tokenised simply using space as a separating character. We then preprocess the data by lowercasing all tokens and removing all punctuations except hyphens. Next, we implement a script which counts the frequency of all types for all years and groups the data by type.

To further reduce the list of candidates, we apply two filters. First, we only keep the types with a frequency of 10 or above, as this is the lower limit used by the editors for low-frequency words. Secondly, we choose only the types that do not appear in the years 2005 to 2007. We primarily add this filter to reduce the number of candidates, as some of the tools we have developed require a lot of computing power. The raw list contains more than 600,000 word types, but the reduced list contains only 79,944 word types. This ensures that we focus on word types that tend to be new in the corpus - although the relatively low frequency of many of the word types naturally makes it less certain that they are actually neologisms. In principle, we are equally interested in candidates that were previously overlooked by the editors, and in the final dataset we will include all word types.

At this step it still contains a high number of word forms that are unlikely to be good lemma candidates, as shown in Table 1 which contains several proper nouns, German words and unidentifiable strings, and only three possible lemma candidates out of 12 tokens. If we sort the list by frequency as seen in Table 2, we easily conclude that frequency alone is not a useful criterion when identifying lemma candidates. Even though some are very good candidates (e.g., *coronakrise*, and *covid-19*), we also find named entities among the

Word		Total	2005-2007	2008	2009	..	2018	2019	2020	2021	2022
stickstoffdioxid	(possibly German)	11	0	1	0	..	7	2	0	0	0
hoé	(unknown)	29	0	1	0	..	1	0	2	7	2
personalekrævende	('staff-intensive')	10	0	1	0	..	0	0	0	2	0
cleantech-virksomheder	('cleantech firms')	31	0	1	5	..	0	0	0	1	0
-kursus	('course (suffix)')	11	0	1	0	..	1	0	2	1	0
kosin	(unknown)	18	0	2	0	..	4	0	0	0	11
grega	(unknown)	12	0	1	0	..	0	0	0	0	0
zurückfallen	(German word)	13	0	2	0	..	0	2	2	2	0
husfliddk	(web url)	11	0	1	1	..	1	1	4	0	2
amap	(proper noun)	11	0	2	2	..	0	0	0	4	0
renneberg	(proper noun)	22	0	1	3	..	0	0	0	1	1
hotspot-indsatsen	('the hotspot effort')	10	0	1	1	..	0	0	0	0	0

Table 1: Random types in the corpus that do not occur in 2005-2007.

Word		Total	2005-2007	2008	2009	..	2018	2019	2020	2021	2022
covid-19		11141	0	0	0	..	0	0	5761	4049	1331
coronakrisen	('the corona crisis')	14385	0	0	0	..	0	0	9742	3579	1064
brexit		10098	0	0	0	..	1254	3151	933	596	354
coronavirus	('corona virus')	8898	0	0	2	..	0	0	6260	2013	594
instagram	(proper noun)	7831	0	0	0	..	644	997	1033	1136	1130
macron	(proper noun)	7029	0	0	0	..	1026	1001	674	620	1305
-fhv	(unknown)	6142	0	1	286	..	516	765	492	82	125
ipad		4565	0	0	0	..	166	238	145	133	155
coronapandemien	('the corona pandemic')	3980	0	0	0	..	0	0	1396	1583	1001
e-mailfinansritzaudk	(an url)	3780	0	0	0	..	0	0	0	0	0
bnb	(proper noun, 'b'n'b')	3737	0	0	0	..	1	3	0	0	0
radio24syv	(proper noun)	3301	0	0	0	..	317	991	190	158	81

Table 2: Most frequent types in the corpus that do not occur in 2005-2007.

high frequency tokens. Additionally, these candidates are already identified by the normal corpus methods in the DDO project. Interestingly, there are unidentifiable strings even among the high frequency types (e.g. *-fhv*).

2.2 Preprocessing II

To further improve the list of candidates, we apply a second, more extensive preprocessing step which also includes the comparison of the data with lists of already registered word forms in other resources. The goal is to remove as many of the tokens that are highly unlikely to be lemma candidates as possible. First, we remove all numbers and URLs. Next, we use the previously kept hyphens to find a very common type of tokenisation error in Danish texts which are caused by the spelling rule of not writing the full form of a compound when listing it next to another one containing the same word component (e.g. we remove *morgen-* which occurs in the phrase *morgen- og aftenritual* ('morning and evening ritual')), corresponding to *morgenritual og aftenritual*. We also remove tokens starting with a hyphen (e.g. *-kursus* and *-fhv*).

To remove proper names, e.g. *Renneberg* and *Quintus*, we use lists of registered personal and place names published by *Danmarks Statistik*[2]. We are aware that some personal names are also common appelatives in Danish[3] but assume that these are already present in the dictionary. Names of organisations are identified and processed in a later step (see 2.3.3).

In a similar way, the list of inflected forms of lemmas appearing in the printed DDO is used to identify and remove all inflected forms of lemmas in the printed DDO. The list of inflected forms of lemmas added to the DDO since 2005 (the online dictionary) is, however, not used to lemmatise the remaining data, it is instead kept for evaluation purposes (see section 3). We use the CSTLEMMA lemmatiser (Jongejan & Dalianis, 2009) as an alternative[4], and sum up the corresponding frequencies at lemmma level (keeping also a list of the original tokens). This step greatly reduces the size of the dataset, and at the same time allows us to treat inflected forms of the same lemma in a similar manner.

Since the focus of the experiment is low frequency lemma candidates, we finally remove tokens and lemmas with a total frequency above 100 across all years; these are likely to have already been identified by our standard corpus methods. We also remove all words that only occur in texts from one or two years out of the 15 years of corpus data that we investigate, taking into consideration the editorial criteria of representation in texts from at least 3 years.

After these preprocessing steps, the list of lemma candidates is reduced from 79,976 to 36,172 candidates, which is still a very high amount of data which - even though it includes a high percentage of lemma candidates *middelklassedrenge* 'middle class boys', *lektiehjælper* 'private tutor', *fejltankning* ('misfuelling'), *envejsbil* ('one-way-car'), and *bodyage* (English loan) - still also contains a lot of named entities (e.g. *okmans*, *grunerwidding*, *jammerbugts*) as well as other noise, such as tokenisation errors (*erlyd*, and *dkandidat*).

2.3 Measuring "lemmaness"

In order to improve the list, we develop a measure of "lemmaness" which we call the *lemma score*. The lemma score is a weighted average of several subscores related to stability in time, adaptation to Danish orthography and morphology, and last but not least semantic similarity to known lemmas in the dictionary. By measuring these features, we reflect a large number of the criteria used by the lexicographer when selecting a lemma for the dictionary.

2.3.1 Stability in time

In the current list, we know that all words are represented in texts from at least three years, but we assume that an even more widespread representation correlates with the suitability to be included as a lemma in a dictionary. We do not take the frequency in each year into account, but simply assign a score if the word occurs just once. The 'stability in

[2] 'Statistics Denmark', a government authority: https://www.dst.dk

[3] *Sten* 'stone' is a Danish male first name

[4] We are interested in developing a method that can also be used for future detection of overlooked lemmas. Therefore, we do not want to base the lemmatisation only on the online dictionary

time' score is the number of years the candidate occurs (in any form) divided by the total 15 investigated years (2005-2022).

2.3.2 Form

A word's adaptation to Danish morphology highly indicates that it has been lexicalised in the language and that it is highly unlikely to be a named entity. Therefore, the previous lemmatisation is used to count the number of forms in the (lemmatised) groups of words, and assign a score according to the number. We do not take PoS into account, which in turn may discredit candidates with inherently fewer word forms in Danish (e.g. adverbs have fewer forms than nouns and adjectives). However, only a few candidates appear with more than three word forms. We assume that the disparity has minimal influence on the final score.

2.3.3 Named Entity

Even though many named entities were already removed from the list during the filtering process by use of data from *Danmarks Statistik*, many names of organisations, products, foreign personal nouns as well as creative artist nouns are still present in the candidate list. To automatically detect these, we use the ScandiNER model [5]. The model expects a sentence as input and it outputs NER-tags with their respective position in the sentence. Therefore, the model cannot identify whether a word is a named entity when seeing it in isolation. To circumvent the problem, we randomly select up to ten sentences from the corpus. Each sentence is then tagged with the model and we compare the tag position with the candidate's position in the sentence. If the two positions overlap, we increase the named entity score. The final named entity score is the total percentage of sentences where the named entity tag overlaps with the candidate position out of all selected sentences. The named entity score is subtracted from the total score.

2.3.4 Compound splitting

As mentioned above, the automatic analysis of the high number of compounds in Danish is a challenge. We chose to develop the compound splitter *DSLSplit*[6] in order to be able to split the many compounds in the most likely word components, so that we are able to analyse these based on existing lemma information in our resources. The compound splitter is characterised by having two modes, a "careful" and a "brute" one. We base the "careful" mode on CharSplit - a German n-gram based compound splitter (Tuggener, 2016) that we have adapted to Danish. The "brute" mode also uses probabilities to estimate the most likely split. It was trained using the manually added information on a part of the compounds in the DDO (30,211 compounds). This data turned out to be far from sufficient and it was therefore supplemented with automatically generated compound information based on the retro-digitised historic Danish dictionary *Ordbog over det danske Sprog* (ODS, Det Danske Sprog- og Litteraturselskab (2023c)). 168,321 compounds were identified automatically, modified to modern orthography and used as training data, even though

[5] Available at https://huggingface.co/saattrupdan/nbailab-base-ner-scandi

[6] DSLSplit and a more detailed description can be found https://github.com/dsldk/dslsplit

they contain some compounding errors. The brute method does not always correctly identify the first of the two compound elements when they are joined by an "e", or an "s" (these letters are also very often final or initial letters of simplex lemmas in Danish). In the present work, we run our compound splitter in *mixed* mode, meaning that the splitter first tries the "careful" approach, and if this method doesn't find a probable split, the "brute" method is applied.

Scores are assigned to the components of the compounds depending on whether they are lemmas in the printed DDO or not. If both are, we assign the highest score (1) to the candidate. If only one component is included in the dictionary, the candidate is assigned a low score (0.5) If none of the components are found in the dictionary or if the candidate cannot be split by the algorithm, it receives no score (0).

2.3.5 Language features

With the naked eye, it is evident that some candidates on the list do not follow the typical spelling of Danish words, either due to tokenisation or spelling errors, or because the word is of foreign origin. To identify this part of the data from the list of lemma candidates, we calculate the likelihood of a character sequence being in accordance with the standard Danish spelling. We apply the tetra-gram model in LexiScore[7]. The model is trained on the list of inflected word forms from the online DDO dictionary containing 641,971 words. Additionally, Laplace smoothing (k=100) is applied to offset the low-frequency tetra-grams. To determine the probability of a character sequence belonging to the target language, we multiply the probabilities for each tetra-gram, assigning a very low probability for any tetra-gram not found in the list of words (specifically, 1e-20). To avoid penalising longer words, the base probability is normalised after the length of the sequence. We set the probability threshold to 0.0001. A candidate above this threshold gets a Danish validity score of 1, while a candidate below the threshold is deemed invalid and gets a score of 0.

Some of the words of foreign origin are highly relevant lemma candidates. We checked 1363 new lemmas in the DDO (lemmas included in 2019-21, both neologisms and older words), and 8% contain non-Danish orthography (e.g. *gefühl*, *betting*, *aftersun*, *ajvar*), and we know for a fact that many neologisms in Danish are loanwords from especially English, but also German. LexiScore also contains tetra-gram models for English(trained on the Moby Crosswords word list[8]) and for German (trained on the German Aspell dictionary[9]). We also include an extra Danish model trained exclusively on head words in DDO. This allows us to compare the probability of a character sequence being the most typical for either of the three languages. The language origin feature score is shown in Table 3.

Highest probability	**Danish**	**DDO head**	**English**	**German**
Language origin	1	0.9	0.8	0.4

Table 3: Language origin feature score

[7] The source code of LexiScore is available at https://github.com/dsldk/LexiScore. New languages are easily added from simple word lists.
[8] Available at https://www.gutenberg.org/files/3201/files/
[9] Available at https://ftp.gnu.org/gnu/aspell/dict/0index.html

2.3.6 Semantic model

Synonyms as well as near synonyms of already included lemmas are very likely to be good lemma candidates. For instance, the new lemma *daikon* 'daikon, type of Japanese radish' was added to the online dictionary in 2021, having a sense very close to three lemmas already included in the printed DDO, namely *kinaradise* ('Chinese radish'), *radise* ('radish'), as well as *ræddike* ('black radish').

One way of checking whether a word is a good lemma candidate or not is to investigate whether it is semantically related to any of the already included lemmas. Word embedding models like word2vec (Mikolov et al., 2013) build on the assumption that a word's meaning can be estimated from its distribution in text, in line with the distributional hypothesis summed up in the famous line 'you shall know a word by the company it keeps' (Firth, 1957). Or, put differently: Semantically similar words typically appear in the same context. A word embedding model creates a vector representation (i.e. a word embedding) of each token in a corpus. By computing the distance between two embeddings, we are able to estimate their semantic similarity. In order to find similar lemmas to the already known ones from the dictionary, we simply need to search through the semantic space created by the word embedding model.

In the experiment we use a model that we have previously trained on texts in the DDO-corpus published before October 2019 when the corpus contained over 1 billion raw tokens, 7.17 million word types, and 2.79 million sentences. We use a model that is trained similarly to the one in Sørensen & Nimb (2018) with opensource Gensim package for Python (Řehůřek & Sojka, 2010), and use the model to find the 20 nearest neighbours of a lemma candidate in our list. The candidate is assigned a value if the neighbour is included as lemma in the DDO. We calculate an overall semantic feature score based on the number of neighbours being DDO lemmas, however with a logarithmic function to decrease the influence of having more than 5. We find that it shouldn't count drastically more having a higher number than this.

2.4 The final lemma score

For each candidate, the features are combined into the final lemma score through a weighted average. The weights are set manually depending on how reliable we find each feature. We adjusted the weights after inspecting the lemma score in a preliminary experiment on only the first ten years of the corpus data. For instance, we found it beneficial to lower the initial weight of the compound feature to lower the advantage of compound candidates. We further discuss this problem in section 4. In descending order, the final weights are: semantic feature (0.6), origin language (0.5), compound (0.4), form (0.2), stability in time (0.1), valid Danish spelling (0.1), and named entity score (-0.8). Finally, the candidate list is ranked after the final lemma score. Examples of candidates and their respective lemma scores and rank is visible in Table 4.

3. Evaluation

We evaluate the final candidate list in two ways. First, we compare the list with the updates from the online dictionary carried out since 2005 in order to see to which degree the automatic method is able to identify the lemmas which have in fact been selected by

Candidate	Semantic	Origin	Compound	Form	Time	Spelling	Named Ent	Lemma score	Rank
havesaks	0.9	1	1	0.1	0.5	1	0	**1.47**	3
campingstol	0.65	0.8	1	0.1	0.44	1	0	**1.34**	176
filologi	0.77	1	0.5	0.1	0.56	1	0	**1.24**	771
olieforum	0.1	1	1	0.1	0.63	1	1	**0.38**	28937
fwd	0	0	0	0.1	0.19	0	1	**-0.69**	36166

Table 4: Selection of candidates with their respective lemma score and rank.

the lexicographers in the last 17 years based on the existing methods in the DDO project. But we also want to evaluate the quality of the remaining candidates on the list. We expect to find a high number of lemma candidates not yet added to the dictionary and would like to measure the recall accordingly. Therefore, two experienced DDO lexicographers manually check a random selection of words from the top and bottom of the list.

3.1 Comparison with dictionary updates from 2005-2022

The online dictionary has been updated with approximately 38,000 new lemma entries since 2005, 30% with morphological information, 70% also with sense descriptions. We assume that all entries added to the dictionary are deemed to be good candidates by the editors and hence can be used as a gold standard in the evaluation. We will refer to the lemmas included in the dictionary updates as *true lemmas*.

In the final list of 36,172 candidates, we find 1550 true lemmas which account to less than 5% of the candidates. Although it seems like a low coverage, we have to consider that we have removed the candidates with the highest frequency and only include the candidates that did not appear in 2005-2007 (see section 2.2). In figure 3, the cumulative sum of true lemmas in the candidate list is presented. From the figure, we can see that the ratio of true lemmas is highest in the top section of the list. In fact, 10% of the true lemmas are covered by the top 1% of the highest scoring candidates, and 50% of the true lemmas by the top 17%. Additionally, we see almost no true lemmas in the bottom section of the list as the curve on figure 3 levels out after around rank 25,000.

3.2 Qualitative analysis

The majority of the words in our generated list of lemma candidates have not (yet) been included in the DDO, and therefore cannot be evaluated by a comparison with the 'gold standard' of entries established in the DDO since 2005. To evaluate the quality of these, we instead extract a subset to be annotated by two DDO lexicographers. We consider a lemma to be correctly identified as a lemma candidate, i.e. a true positive, when at least one of the two lexicographers find it relevant to present on a lemma candidate list. The subset is composed of a random selection of 394 candidates; 199 from the top 1500 (the best candidates), and 195 among the last 1500 (the worse candidates) (after having removed all the true lemmas) [10]. The subset is shuffled to obscure the rank of the candidates. On this subset, we calculate the recall to be 92.6%

[10] Originally, we extracted 200 from each selection. However, some of the candidates appeared in the dictionary in another form, e.g., *alterego* appeared as *alter ego*.

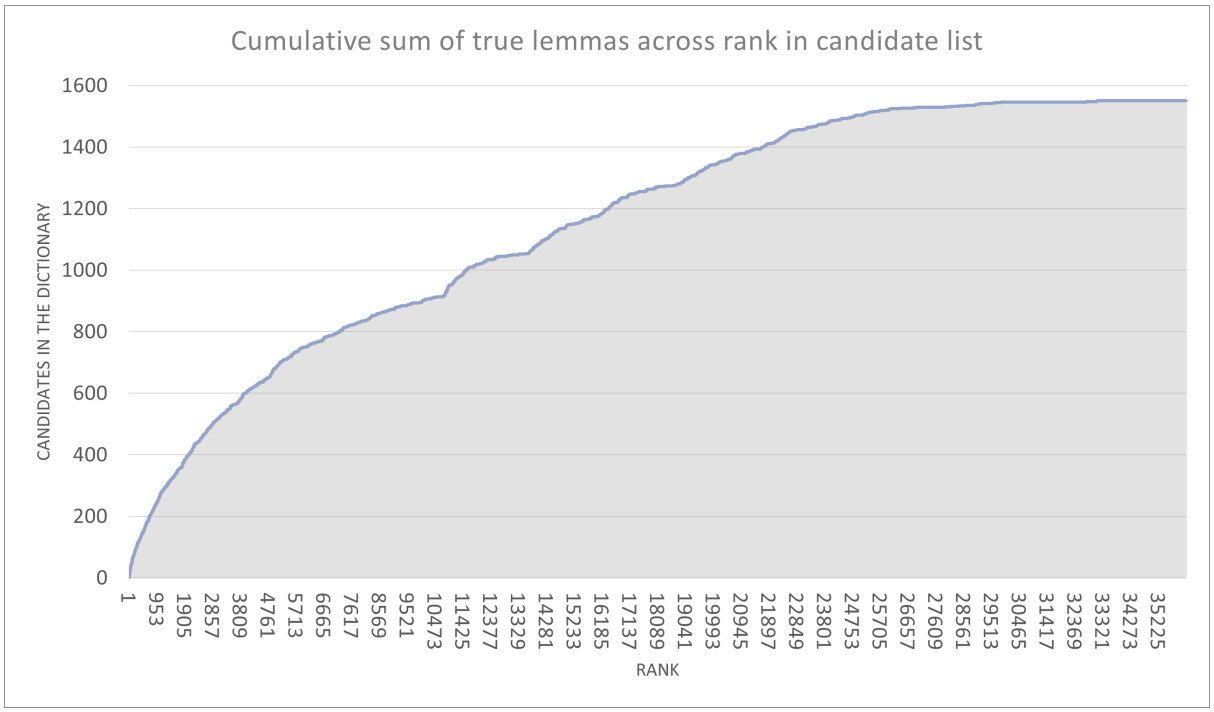

Figure 3: Cumulative sum of the true lemmas (i.e. lemmas added to the dictionary since 2005) across the rank in the final candidate list

Among the 199 top candidates, 94% (188) are estimated to be a relevant candidate by at least one of the two annotators. The 6% (11) non-relevant fall into four categories: ad hoc compounds (e.g. *accelerationssamfund* 'acceleration society' and *haveoplevelse* 'garden experience'), named entities (e.g. the bakery company *Lagkagehuset*), tokenisation errors (*haftafgørende* → *haft afgørende*), and domain specific terms (*hjertertaber* 'lit. heart loser', a bridge term). The ad hoc compounds constitute more than half.

Among the candidates with the lowest score, 92% (180) are annotated to be irrelevant lemmas by at least one of the two annotators. The remaining 15 words were judged to be relevant. When we take a closer look at these, four received a low score in the automatic process because of lemmatisation errors (e.g. *facebookven* falsely lemmatised as *facebookv*), and the remaining part received a low score because they had either features resembling a named entity, or (in 7 cases) are (entirely or partly) words of foreign origin (*twitterfeed*, *techcrunch*), (tuaregoprører 'tuareg rebels'), or maybe even a combination of both.

4. Discussion

In spite of the promising results, we are aware of some pitfalls in the approach that we will address in the following subsections.

4.1 The impact of the corpus composition

A limitation of our study is the composition of the corpus. The corpus is a newswire corpus and does not reflect the language of the average speaker. First, the text is both written and edited by professional writers and large groups of language users are not represented, e.g. children, teenagers, as well as adults from other professions. Secondly, the texts in it

only cover a specific range of topics that are considered newsworthy, and they represent only a rather formal style. This means that a great deal of words and expressions used in everyday language are most likely missing from the corpus and thus will still not be uncovered by our methods.

On the other hand, if we had access to a more balanced corpus, our method may be trivial as the corpus frequencies would be more representative. Still, it could be interesting to investigate whether our method can be used in other domains to uncover even more lemma candidates. We are particularly interested in a corpus that contains more everyday Danish. In future work, we are therefore considering including Danish internet forums. Here, we need to consider whether the weights can be directly transferred to other text types and whether we need to examine a new type of feature connected to the domain.

4.2 Balancing of the weights

One of the challenges in the experiment was to balance the weights. This was done manually, and we aimed at finding the optimal balance between the features. Although the current weights successfully distribute a large number of relevant candidates among the top ranks, we see space for improvements in the middle range of the list. This becomes evident when studying the true lemma curve on figure 3 since the curve suddenly increases around rank 10,000. From the qualitative analysis in section 3.2, we also know that the named entity feature has too high a negative impact on some of the relevant lemmas. Likewise, the language features may penalise foreign words too much. The question is how many missed lemmas we can accept. Adjusting the feature scores may cause more noise to appear higher up in the list. We need to further investigate the impact of the individual features to refine our approach in the future.

One idea is to split our data into a training and test set. The weights can then be set on the training set before we evaluate on the test set. Even though we have a good standard in the form of the dictionary updates, these only give us the positive cases. We still need to collect a sample of non-lemma cases, e.g. cases of words that were actively discarded by the lexicographers. Simply using the words that are not in the dictionary would not be representative as they may be good, overlooked lemmas. Now that we have gathered some information about the characteristics of irrelevant words, we are able to conduct more experiments.

4.3 Are compounds too prominent?

The compound feature rewards compounds of which the components are already included in the printed dictionary. Thus, we might overemphasise compounds at the expense of simplex words and derivations. The question is how problematic this is for the lemma selection process. Compounds are prominent in Danish neologisms. A study by Trap-Jensen (2020) estimates that 52% of neologisms in recent DDO updates are compounds. However, the percentage of compounds is much higher if we look at both actual neologisms and words that were previously overlooked or fell outside the scope of the lemma selection when editing the printed dictionary. In the last three updates of DDO (November 2021, June 2022, and November 2022 (Det Danske Sprog- og Litteraturselskab (2023b))) a total of 938 lemmas were added to the dictionary. Of these, 738 are compounds, equivalent to

78.7%. Thus, it is not unreasonable to expect our method to also identify a high number of compounds. In addition, DDO describes a number of common derivational affixes and suffixes, and the current compound splitter is therefore able to split certain derivations. In the future, it is worth exploring whether we can update the compound splitter to differentiate between actual compounds and derivations and thereby give them different scores.

With the high number of compounds, we also face the problem of ad hoc compounds. The automatic compound feature cannot distinguish compounds that have been established in the language from ad hoc ones. Since ad hoc compounds constitute more than half of the high ranking irrelevant words in the qualitative analysis, they seem to cause a more general problem to our method. We simply lack more information about the characteristics of compounds which are produced on the fly. For instance, are certain words more productive than others as components in ad hoc compounds? This is something we plan to investigate further.

4.4 The "newness" criterion

A large number of the highest ranking candidates are in fact not new in the language, although we have disregarded word types that occur in 2005-2007. The main purpose of disregarding these word types was to reduce the size of the dataset rather than filtering it for actual neologisms. When the editors of DDO include more lemmas in the dictionary, they are not only searching for neologisms, but also previously overlooked words like *julekugle* (eng. 'Christmas ornament') and *havesaks* (eng. 'garden sissors'). Therefore, the scope of the study is wider than just neologisms in a strict sense. Nevertheless, we plan to expand the method on all word types in the entire newswire corpus (i.e. words that also occur in 2005-2007) to identify even more overlooked lemma candidates.

4.5 Are foreign words unfairly penalised?

We introduced LexiScore in order to automatically identify and filter out noise coming from web addresses, failed tokenisation, and also non-Danish texts from the corpus. However, many neologisms in Danish are direct borrowings, especially from English. Some domains are also naturally described by means of loan words, e.g. culinary terms like *nduja* 'italian pork sausage'. In the corpus, we find many examples of full sentences from other languages, for instance through a direct quote. For a foreign word to be considered a loan in Danish, it has to occur in a context with a majority of Danish lemmas. In our experiment we are limited to only study the character composition without being able to include the context. A better way to calculate the language origin feature might be to look at the combined probability of the candidate and its closest context (+/- two or more words) to see whether they belong to a specific language. Alternatively, LexiScore can be used to identify and subsequently remove longer sequences of English and German texts during the preprocessing steps to reduce the number of foreign words in the data.

5. Conclusion

The lemma score method that we have presented is a useful contribution to the task of identifying the new lemmas to be added to the DDO dictionary. The approach has enabled

us to effectively sort out a large amount of irrelevant words in the extracted corpus data so that only a minimum of noise is left. Where a manual inspection in the initial candidate list (before the scoring) showed that only roughly every 20th word was relevant, we now find relevant words in up to 94% of the cases in the top of the list. The fact that we find half of the lemmas added since 2005 in the top 17% of the generated list also proves the high quality of the method. In the daily lexicographic work it means that it is now a manageable task for the lexicographers to manually inspect the list. In this way our method speeds up the process of lemma selection in the DDO project significantly.

The greatest influence on the scores was provided by the word embeddings, and by the automatic identification of Named Entities. We find that the use of word embeddings and Named Entity recognition allows for a more efficient and accurate selection process. We believe that especially the idea of combining the scores ensures the good quality of the results. Instead of using each feature as a filter to remove noise, we consider all features at once to get a complete picture of each word's potential for inclusion. Thus, it is not detrimental if a word gets a low score in one feature if the scores in the other features are high enough to counterbalance the score.

Another advantage is that the method does not require an annotated dataset to train a supervised machine learning model. Since the list is going to be manually processed by the lexicographers, we don't need a very high accuracy of the ranking. We have shown that a weighted average yields good results when an annotated dataset is not available.

In conclusion, we believe that the data we have obtained is highly useful for the DDO lexicographers, as it allows them to select lemmas in a more efficient and objective manner which in turn also leads to a higher quality dictionary.

6. Acknowledgements

We would like to express our sincere gratitude to several individuals and organisations who have contributed to this work. First and foremost, we would like to thank senior editor Jonas Jensen for his invaluable help with the annotation of the data as well as with the proof-reading and commenting on the paper, and senior editor Thomas Troelsgård for supplying us with a comprehensive list of all inflected forms of lemmas in the written version of *Den Danske Ordbog*.

Finally, we would like to thank *The Carlsberg Foundation* and the Danish Ministry of Culture for their generous grants, which have enabled us to continue our work on the dictionary. Without their support, this project would not have been possible.

7. References

Det Danske Sprog- og Litteraturselskab (2023a). Den Danske Ordbog. https://ordnet.dk/ddo. Accessed on April 21, 2023.

Det Danske Sprog- og Litteraturselskab (2023b). Nyeste ord i DDO. https://ordnet.dk/ddo/nyeste-ord-i-ddo. Accessed on April 21, 2023.

Det Danske Sprog- og Litteraturselskab (2023c). Ordbog over det danske Sprog. https://ordnet.dk/ods. Accessed on April 21, 2023.

Falk, I., Bernhard, D. & Gérard, C. (2014). From non word to new word: Automatically identifying neologisms in French newspapers. In *LREC-The 9th edition of the Language Resources and Evaluation Conference*.

Firth, J. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, pp. 10–32.

Freixa, J. & Torner, S. (2020). Beyond frequency: On the dictionarization of new words in Spanish. *Dictionaries: Journal of the Dictionary Society of North America*, 41(1), pp. 131–153.

Halskov, J. & Jarvad, P. (2010). Manuel og maskinel excerpering af neologismer. *NyS, Nydanske Sprogstudier*, (38), pp. 39–68.

Jongejan, B. & Dalianis, H. (2009). Automatic training of lemmatization rules that handle morphological changes in pre-, in-and suffixes alike. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. pp. 145–153.

Kerremans, D., Stegmayr, S. & Schmid, H.J. (2012). The NeoCrawler: Identifying and retrieving neologisms from the internet and monitoring ongoing change. *Current methods in historical semantics*, 73, p. 59.

Langemets, M., Kallas, J., Norak, K. & Hein, I. (2020). New Estonian Words and Senses: Detection and Description. *Dictionaries: Journal of the Dictionary Society of North America*, 41(1), pp. 69–82.

Lorentzen, H. & Nimb, S. (2011). Fra krydderkage til running sushi – hvordan nye ord kommer ind i Den Danske Ordbog. *Nye ord, Sprognævnets Konferenceserie*, 1, pp. 69–85.

Martelli, F., Navigli, R., Krek, S., Kallas, J., Gantar, P., Koeva, S., Nimb, S., Sandford Pedersen, B., Olsen, S., Langemets, M., Koppel, K., Üksik, T., Dobrovoljc, K., Ureña-Ruiz, R., Sancho-Sánchez, J.L., Lipp, V., Váradi, T., Győrffy, A., Simon, L., Quochi, V., Monachini, M., Frontini, F., Tiberius, C., Tempelaars, R., Costa, R., Salgado, A., Čibej, J., Munda, T., Kosem, I., Roblek, R., Kamenšek, U., Zaranšek, P., Zgaga, K., Ponikvar, P., Terčon, L., Jensen, J., Flörke, I., Lorentzen, H., Troelsgård, T., Blagoeva, D., Hristov, D. & Kolkovska, S. (2023). Parallel sense-annotated corpus ELEXIS-WSD 1.1. URL http://hdl.handle.net/11356/1842. Slovenian language resource repository CLARIN.SI.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Nimb, S., Sørensen, N.H. & Lorentzen, H. (2020). Updating the dictionary: Semantic change identification based on change in bigrams over time. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, 8(2), pp. 112–138.

Norling-Christensen, O. & Asmussen, J. (2012). The Corpus of the Danish Dictionary. *Lexikos*, 8(1). URL https://lexikos.journals.ac.za/pub/article/view/955.

Pedersen, B.S., Nimb, S., Olsen, S., Troelsgård, T., Flörke, I., Jensen, J. & Lorentzen, H. (2023). The DA-ELEXIS Corpus - a Sense-Annotated Corpus for Danish with Parallel Annotations for Nine European Languages. In *RESOURCEFUL 2023 - Proceedings of the second workshop on Resources and Representations for Under-Resourced Languages and Domains, NoDaLiDa 2023*. Forthcoming.

Řehůřek, R. & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, pp. 45–50. http://is.muni.cz/publication/884893/en.

Sørensen, N.H. & Nimb, S. (2018). Word2Dict - Lemma Selection and Dictionary Editing Assisted by Word Embeddings. In *Proceedings of the 18th EURALEX International Congres: Lexocography in Global Contexts*. pp. 819–827.

Trap-Jensen, L. (2020). Language-Internal Neologisms and Anglicisms: Dealing with New Words and Expressions in The Danish Dictionary. *Dictionaries: Journal of the Dictionary Society of North America*, 41(1), pp. 11–25.

Trap-Jensen, L., Lorentzen, H. & Sørensen, N.H. (2014). An odd couple – Corpus frequency and look-up frequency: what relationship? *Slovenscina 2.0*, 2(2), pp. 94–113.

Tuggener, D. (2016). *Incremental coreference resolution for German*. Ph.D. thesis, University of Zurich.