Towards automatic linking of lexicographic data: the case of a historical and a modern Danish dictionary

Towards Automatic Linking of Lexicographic Data: the case of a Historical and a Modern Danish Dictionary

Ahmadi S.[1], Nimb S.[2], McCrae J.[1], Sørensen N.[2]

[1] *Insight Centre for Data Analytics, National University of Ireland, Galway*
[2] *Society for Danish Language and Literature (DSL), Copenhagen, Denmark*

Abstract

Given the diversity of lexical-semantic resources, particularly dictionaries, integrating such resources by aligning various types of information is an important task, both in e-lexicography and natural language processing. The current study aims at analyzing the automatic alignment of word senses of the same lemmas across two comprehensive monolingual Danish dictionaries, the historic *Ordbog over det danske Sprog* and the modern *Den Danske Ordbog*. We report our efforts in creating a gold-standard dataset and show that semantic similarity measures can be efficiently used to create statistical models to automatically align senses across dictionaries.

Keywords: semantic similarity detection; dictionary linking; natural language processing; e-lexicography

1 Introduction

During the past decades, there have been many efforts in applying natural language processing (NLP) techniques for word sense alignment (WSA) where senses of identical words are aligned across various lexical resources. This task is proved to be beneficial to many applications such as semantic role labelling (Palmer 2009) and information extraction (Moro et al. 2013). But also within e-lexicography and the work of publishing a series of online monolingual dictionaries is the alignment of senses of identical lemmas very relevant, opening up for new ways of presenting word information to the users of the resources.

The Society for Danish Language and Literature (DSL) has been publishing scholarly edited Danish dictionaries for more than 100 years, since 2005 also in the form of online dictionaries. Two of these, the modern dictionary *Den Danske Ordbog*[1] ("The Danish Dictionary", henceforth DDO) covering the senses of more than 100,000 Danish lemmas from 1955 till today, and the historic, retro-digitized dictionary *Ordbog over det danske Sprog*[2] ("Dictionary of the Danish Language" henceforth ODS), covering 220,000 Danish lemmas from 1700 till 1955, are in the Danish society considered to be key lexical resources of the Danish language. Both are available at the same public dictionary site of DSL, *ordnet.dk* which has more than 100,000 daily users. At the site, the lemmas in the two dictionaries are connected at a string-based level (by implying exact string similarity), allowing for hits across the two resources when a word is looked up by a user. In the XML structure, ODS has by the use of semi-automatic methods been supplied with links at lemma level to a number of DSL's retro-digitized historic dictionaries covering Danish before 1700 (Svendsen et al. 2020). Opposite to this, DSL's lexical resources for modern Danish are all linked not only at lemma level, but also at sense level. This means that DDO shares sense ID numbers with not only a Danish thesaurus, *Den Danske Begrebsordbog* (Nimb et al. 2014), but also the Danish WordNet DanNet (Pedersen et al. 2009) as well as the Danish FrameNet lexicon (Nimb 2018), see Pedersen et al. (2018).The semantic linking between DDO and the thesaurus constitutes the basis of the compilation of the FrameNet lexicon (Nimb et al. 2017) as well as the presentation of groups of near-synonyms and thematically related words from the thesaurus in the online DDO (Nimb et al. 2018). However, the important but challenging task of linking the modern resources to the historic dictionaries still remains to be carried out. The future online publishing of digital dictionaries at *ordnet.dk* will to a very high degree benefit from such links, opening up for new ways to present the elder vocabulary to dictionary users. For example functions from the modern dictionary like "Ord i nærheden" could be easily transferred to the online ODS and thereby give new insights into the Danish vocabulary and conceptualization in older times.

Linking identical lemmas at sense level in the two key dictionaries DDO and ODS is the obvious first step to take. Once a method is developed for lemmas where we know for sure that there are sense matches between the dictionaries, the method can be applied on the rest of the vocabulary, and senses can be matched across lemmas, not only between identical lemmas. DDO and ODS are similar in many ways since they are compiled by the same institution, DSL, and since DDO from the beginning was planned to be a modern follow-up to ODS. The DDO project was initiated approx. 40 years after the last volume of ODS had been published, and there is an overlap in the different language periods of Danish that they describe: the middle period of the 20th century. Furthermore, DDO is to a high degree inspired by the lexicographic style that ODS had already established for dictionaries being compiled at DSL, both w.r.t. the method, structure and content. ODS has been edited on the basis of approx. 2.5 million manually collected sentences with precise source citation, DDO on the basis of a 40 million text corpus, and both dictionaries use authentic language examples as sense documentation.

In this paper, we will discuss how natural language processing techniques can be applied in the task of aligning the senses of identical lemmas in the two dictionaries, a task which otherwise would be a time-consuming and difficult challenge due to the high number of senses in both of them. The main objective of the study is to evaluate the performance of various automatic methods to carry out the linking, such as string similarity measures and word embeddings. As a preliminary study of its kind for ODS and DDO, we define our alignment task as detecting sense candidate pairs within a combination of all senses for two identical lemmas in two resources.

The rest of the paper is organized as follows. We first describe our lexicographic data, their similarities as well as their dissimilarities in Section 2. Then, in Section 3, we present our methodology where the preparation of the data, the manual

[1] https://ordnet.dk/ddo
[2] https://ordnet.dk/ods

annotation, and the models are introduced. Section 4 provides the results of our experiments indicating how sense length and various experimental setups change the performance of the alignment task. Finally, the paper concludes in Section 5 where we mention a few future directions in the same field.

2 Lexicographic Data

A monolingual dictionary can be considered as a knowledge repository which provides description of the vocabulary of a language with various information, particularly senses. Although two monolingual dictionaries such as DDO and ODS describe the same distinctive and unique ideas in the same language within a certain time period, the way they do it may be very different. The differences are mainly observed as follows:

Sense structure: senses in comprehensive dictionaries are typically organized in a hierarchy where semantically related concepts are provided as subsenses to a main sense. However, the sense granularity and the exact distinctions drawn between both main senses and subsenses of a lemma might differ quite a lot across monolingual dictionaries. Closely related concepts, e.g. the many cases of regular polysemy in the language (see among others Buitelaar 2000; Pustejovsky 1998), might be expressed as separate subsenses, but might as well be (indirectly) included in the main senses. This varies not only across dictionaries, but also within the same dictionary. Furthermore, the sense granularity of a dictionary is influenced by the specific editorial guidelines, according to for example the space available in printed versions, however also by the more subjective and individual judgments made by each lexicographer as stated by Kilgarriff (2003: 372): "any working lexicographer is well aware that, every day, they are making decisions on whether to 'lump' or 'split' senses that are inevitably subjective".

Definition content: The description style decided upon by the two dictionaries, as well as the lexicographer's individual description style, focus on meaning aspect and lexical word choice, may vary quite a lot. When the two dictionaries are compiled in different time periods, such differences become even more significant. In this case, spelling variations over time in the language might also be a factor that must be taken into consideration.

If we compare our two dictionaries, ODS is first of all a historical dictionary covering Danish from 1700- ~1950, where DDO is a modern dictionary covering Danish from around 1950. Its main focus is on the years after 1982 where the first corpus texts that it builds upon date from, see Lorentzen (2004).

ODS was published in 27 volumes describing 188,000 lemmas in the years 1918 to 1954, with a later addition of 5 supplementary volumes with 35,000 lemmas, published 1992-2005. It contains far more dialectal language than DDO, both at lemma and sense level.

DDO was edited in a much shorter period (1994-2003), and published in far less volumes, namely 6, in the years 2003-2005, at that time describing the senses of 66,000 lemmas. Today, the online version describes the senses of 100,000 lemmas. The dictionary focuses on general language, both w.r.t. lemma selection and sense descriptions. DDO still only covers half as many lemmas as ODS. Also at sense level, ODS is more extensive. Since ODS describes Danish in a 250-year period, and DDO only in a 50-year period, ODS covers far more historic senses per lemma.

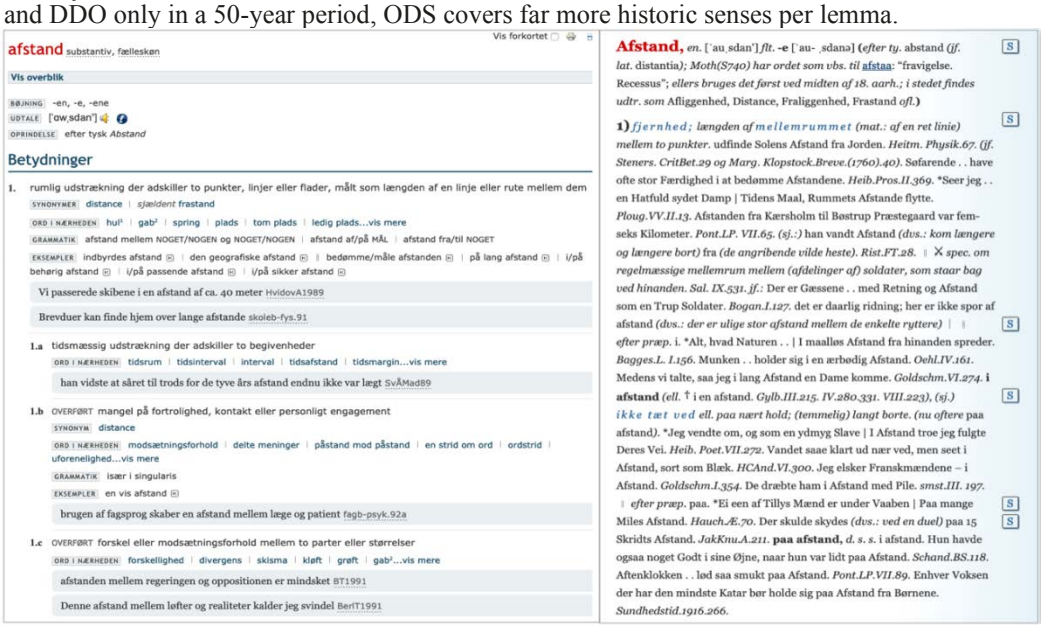

Figure 1: The noun *afstand* ("distance") in the two Danish monolingual dictionaries, DDO (left) and ODS (right).

We know for a fact that there is an overlap in the lexicographic content of DDO and ODS, both at lemma and sense level. The editors of DDO were generally advised to consult ODS when establishing the DDO descriptions of identical lemmas, since many of the modern senses were already registered and described in ODS, see an example, the noun *afstand* ("distance"), in figure 1, where the first senses are similar. From our studies and the extraction of datasets (see below), we also know that around 86% of the central lemmas in modern Danish are included in ODS (Pedersen et al. 2019).

If we look further into the entries of the two dictionaries that we want to link at sense level, we find many resemblances between them, but also many differences.

Sense structure: Both make use of a hierarchical structure with main senses and subsenses, however in different ways.

The order of main senses as well as subsenses is in ODS based on etymology but in DDO on corpus frequency. DDO establishes only main senses proved by concrete textual examples, before the closely related senses to it are listed in the form of subsenses. These might represent either a broader, a narrower, or a figurative use to a higher degree of the main sense, and also have to be manifested in concrete examples in the language. Opposite to this, ODS operates with "main" senses in the structure which are in fact rather a kind of heading or very broad "summing up" sense description for a series of subsenses to be listed, which are then the only ones to be manifested in concrete language. This has of course the consequence that very often two senses in ODS, namely both the heading "main" sense and one of its subsenses are semantically related to the same one sense in DDO. So, in this case, ODS splits in more senses than DDO does. When it comes to sense granularity, they also differ in other ways, since ODS often "lumbs" content that DDO would instead express as several senses in the structure, by using formulations like: "*også om*" ("also about"), "*dels .., dels*" ("both .. and"), "*også uegentl*" ("also figurative") in one and the same definition. So, in these cases DDO splits in more senses than ODS does. Furthermore, the difference in size might have influenced the sense granularities of the two dictionaries. The DDO editors were often encouraged to rather "lumb" the senses of the less frequent lemmas due to the limited space in the printed edition, e.g. in the cases of regular polysemy. The editors of ODS had less restrictions on space, which might have had as consequence that they splitted senses more often.

Definition content: The time span between the edition of the first volumes of ODS and the most recent edition of lemmas in DDO is 100 years. This leads to many differences in lexicographic description style. The definition style of ODS is very compact, aiming at presenting as many details as possible in one and the same phrase. The editors of DDO focused instead on the communicative qualities of a definition. Where ODS uses many parentheses, additional words and phrases and a deep syntactic structure with many attributives and subordinate phrases in order to try to cover all aspects of a sense, DDO focuses on the prototypical aspects and prefers a more flat syntactic structure (see figure 2, the verb *lukke* ("to close"), and figure 3, the noun *standpunkt* ("view") for examples). When DDO makes use of supplementary explanations, these are easily identified automatically, always being initiated by a semicolon in the definition text, or being placed in two seperate XML-fields, one for connotative, one for encyclopedic information. These fields are not a part of the extracted data to be linked.

	Danish definition	English translation
verb *lukke* ('to close')	ODS: *trække, lægge, skyde hen for (over) en aabning, saaledes at denne spærres, udfyldes, tilstoppes; især m. h. t. et dertil beregnet og anbragt (i aabningen passende) spærremiddel, fx. klap, lem, dør; m. h. t. dør olgn. ogs. undertiden: (trække til og) laase ell. stænge* DDO: *bevæge noget dertil indrettet hen foran eller hen over en åbning så den spærres*	ODS: "pull, place, shoot over (over) an opening so that it is blocked, filled out, clogged; in particular w.r.t. a specially designed and arranged (in the aperture) blocking means, e.g. a clap, limb, door; w.r.t. doors or the like also sometimes: (pull and) lock or close" DDO: "move something to the front of or across an opening to lock it"

Figure 2: The verb *lukke* ("to close") descriptions of the same sense in DDO and ODS differ to a high degree w.r.t. syntax and description style. Where DDO focuses on the prototypical type of closing something, ODS tries to cover all possible ways of doing it, with all types of objects.

Also when it comes to the content of the definition text, we find many differences between the two dictionaries, either due to the time span between the edition of the two dictionaries, or simply to the lexicographer's individual choices in each case. See figure 3 for an example (the lemma *standpunkt* ("view'") where there is no word at all in common between the two definitions, even though they convey the same meaning.

	Danish definitions in ODS and DDO	English translations
noun *standpunkt* ('view')	ODS: *om en persons åndelige stade som forudsætning ell. baggrund for hans anskuelser, synsmåde ell. handlemåde; synspunkt; ogs. om den anskuelse, hvortil man er kommet, det grundsyn, man anlægger på noget, ell. (i videre anv.) om stadium ell. trin i en persons åndelige ell. sociale udvikling ell. i en sags, et forholds udvikling olgn.* DDO: *opfattelse af og holdning til et bestemt spørgsmål el. anliggende*	ODS: "about a person's spiritual state as a prerequisite or background to his views, mode of view or mode of action; point of view. about the view to which one has come, the basic view that one is applying to something, or (further use) about the stage or step of a person's spiritual or social development or the development of a case, a relationship, etc." DDO: "perception of and attitude to a particular issue or matter"

Figure 3: Different word choice: The two definitions of the noun *standpunkt* ("view") in ODS and DDO describe the same sense but have no lexical content words in common. The ODS definition is furthermore an example of the complicated definition style of the dictionary.

Figure 4 illustrates that definitions might also focus on different aspects of word meaning, i.e. different qualia roles (Pustejovsky 1995: 76). In ODS, "honey" is described by focusing on how it is produced, the AGENTIVE role: "factors involved in its origin or bringing it about", in DDO mainly by focusing on how it is used, the TELIC role: "its purpose and function" having as consequence that the resulting definitions become very different.

	Danish definitions in ODS and DDO	English translations
noun *honning* ('honey')	ODS: *plantesaft, der er opsuget af bier, omdannet i deres tarmkanal og atter gylpet op* DDO: *sød klæbrig masse som bier danner af blomsters nektar, og som fx spises på brød eller bruges som ingrediens i mad*	ODS: "sap/plant juice which is soaked up by bees, transformed in their intestinal tracts and regurgitated" DDO: "sweet sticky mass that bees form from the nectar of flowers and which for example is eaten on bread or used as an ingredient in food"

Figure 4: Different meaning aspects: In ODS, "honey" is described with the focus on the biological process behind where DDO instead focuses on the resulting food and how it is consumed.

But we also find many quite parallel definitions in the two dictionaries, both w.r.t. syntactic style and lexical choice (however, the lemmas may be in different morphological forms), see Figure 5 for examples.

	Danish definitions in ODS and DDO	English translations
noun *klemme,* ('trouble') - (ODS sense 2)	ODS: ***knibe**, forlegenhed;* ***vanskelig situation***; DDO: ***vanskelig situation**; knibe*	ODS: "trouble, embarrassment; difficult situation;" DDO: "difficult situation; trouble"
noun *klemme* ('sandwich') (ODS sense 6)	ODS: ***tykt** (og mindre lækkert)* ***stykke smørrebrød** (især om* ***sammenlagte** (egl.: sammenklemte? ell. mindende om en (tøj)klemmes to led ell. flader?) stykker* ***smørrebrød**, der* ***medbringes** til arbejdsstedet)* DDO: ***tykt stykke smørrebrød**; to skiver brød som er* ***lagt sammen** omkring et* ***stykke** pålæg og* ***medbragt** i en madpakke*	ODS: "thick (and less delicious) piece of sandwich (especially about combined (maybe in fact squeezed or reminiscent of a clothespin's two joints or surfaces?) pieces of sandwiches brought to the workplace)" DDO: "thick piece of sandwich; two slices of bread that are put together around a piece of topping and brought in a packed lunch"
verb *sikre* ('to secure') (ODS sense 1.1)	ODS: ***beskytte** en ell. noget* ***mod angreb**, skade,* ***overlast**, forstyrrelse olgn.* ***v. hj. af forebyggende foranstaltninger*** DDO: ***beskytte mod angreb**,* ***overlast**, forringelser e.l.* ***vha. forebyggende foranstaltninger***	ODS: "protect somebody or something from attack, injury, nuisance, disruption, etc. using preventative measures"→ DDO: "protect against attack, nuisance, deterioration etc. using preventive measures"
noun *middag* ('noon'), ODS sense 1.	ODS: *det* ***tidspunkt midt på dagen** (kl. 12), da* ***solen står højest på himlen*** DDO: ***tidspunktet midt på dagen hvor solen står højest på himlen** (ca. mellem kl. 11 og 13)*	ODS: "that time in the middle of the day (12 noon) when the sun is highest in the sky" DDO: "the time in the middle of the day when the sun is highest in the sky (approximately between 11am and 1pm"
noun *søvn* ('sleep'), ODS sense 3.	ODS: ***materie** (pus),* ***afsondret i** øjet **(øjenkrogen)** under søvnen* DDO: ***materie** som* ***afsondres i øjenkrogene** mens man sover*	ODS: "matter (pus), secreted in the eye (eye hook) during sleep" DDO: "matter that is secreted in the corners of the eye while sleeping"

Figure 5: Resemblances in lexical choice: Examples of definitions in ODS and DDO where the two dictionaries make use of identical words or lemmas (in bold), and even identical phrases, to describe the same sense.

An important difference between the two dictionaries is that ODS in some cases presents metainformation in the form of precise sense references (numbers) for words in the definition text itself, typically when it consists of only a synonym or when the lemma is a derivation (e.g. a number reference from a verbal noun to the relevant verb sense). This we never find in the isolated definition data from DDO. We also very often find other types of metainformation inside the ODS definition

text, for example the lexicographers' guess regarding the etymology of the sense of the noun *klemme* ("sandwich") as seen in figure 5. Another very big difference is that ODS sometimes have no definition text at all to a sense in the structure, only examples.
Finally there are some divergences in the orthography of two dictionaries due to a Danish language spelling reform in 1948 where for example the letters "aa" were replaced by a new letter "å". See examples in figure 2: *aabning* → *åbning*, *laase* → *låse*). Many abbreviations are also spelled differently in the two dictionaries: ODS *p. gr. af* (*på grund af* "because/due to") → DDO: *pga.*, ODS *ell.* (*eller* "or") → DDO: *el.*, ODS: *ogs.*(*også* ("also") → DDO *også*, etc. The structure and content of ODS are described (in Danish) in a number of texts at https://ordnet.dk/ods/, see for example Jacobsen & Juhl-Jensen (1918).
Differences in XML structure: Our task considered, it is important to mention that the dictionaries to a very high degree differ when it comes to the number of markups in the XML structure. DDO was from the very beginning edited in a fine-grained XML structure with isolated content-named elements, e.g. one for the definition, another for the citation etc., constituting the perfect basis for the later online edition. Opposite to this, ODS has been retrodigizied based on the printed version in order to be published online in 2005, and is still in the process of being transformed into a well-defined XML structure. With regard to the semantic part, only the full sense content including citations, etc. has so far been identified automatically in the established digital manuscript, not the exact part of the sense description which constitutes the definition phrase that would be ideal to be compared to the definition phrase of DDO in our task. The definition text from ODS is often initiated by different types of metainformation on for example frequency, chronology, domain, as well as use, which is not part of the DDO definition text that it is compared with. Furthermore, metainformation can even be part of the definition text itself, as described above.
To sum up, in many cases of identical lemmas the two dictionaries differ quite substantially when it comes to structure as well as content, furthermore the extracted ODS definition text used as input for our task is very noisy compared to the extract from DDO.

3 Methodology

3.1 Data Preparation

Based on our knowledge of the many differences between the two dictionaries regarding both structure and content, the datasets for linking task are created following these steps:

- Extracting identical lemmas in DDO and ODS: After normalizing the spelling variations, we extract lemmas with identical spelling and with subsequent manual corrections.
- Extracting senses in ODS and DDO: This was a challenging process as different reference keys that are used for senses were dealt with differently. Due to the complexity in extracting senses, we did not take multi-word expressions into account in the extraction process.
- Normalizing orthographies: In ODS, an old Danish orthography is used, as formerly mentioned. We automatically converted that orthography to the modern one using a mapping between characters. The mapping consists simply of 24 mappings like "kjø → kø" and was constructed by philologists at DSL working on 19th century Danish literature (see for example Bjerring-Hansen et al. (2019)).
- Summarizing senses: As described above, the ODS sense descriptions are often very detailed and syntactically complex (see figures 2 and 3 for examples) and the borders between definition text, usage examples and idioms still remain to be fully identified in the XML structure. For the experiments in this study, in addition to the full original text, we create three other datasets in such a way that the number of space-separated tokens is limited to only 15, 20 and 25 tokens. The performance of the alignment task with respect to the number of tokens is shown in Section 4.
- Unifying sense hierarchy: The senses in both dictionaries are provided in a hierarchical form to represent semantically-related concepts. For our task, we bring all the senses along with subsenses at the same level. Having said that, the sense hierarchy structure can explicitly provide information about the semantic relationship between senses and therefore should preferably be considered in later experiments with the data.
- Dataset creation: Entries are linked using a common ID, called metaID, in ODS and DDO. Using this ID, senses of the same headwords in the two dictionaries are brought together for the annotation task.

3.2 Manual Annotation

In the manual linking process where the training data was established, we annotated the senses of a large number of lemmas which were initially linked between ODS and DDO (meaning that they are etymologically the same words). The lemmas were picked out randomly among a selection of "core concept lemmas", already having been identified in DDO, constituting of a total of 4,646 DDO lemmas of which at least one sense constitutes the Danish equivalent of one of the 5000 core/base concept synsets in Princeton Wordnet (Pedersen et al. 2019). Approximately 75% of these DDO core concept lemmas are polysemous, and even though they only constitute 5% of the total number of lemmas in the dictionary, they cover more than 20% of its senses (Pedersen et at. 2019). The lemma selection thereby represents a high degree of polysemy which makes it highly suitable for our task. The DDO core concept lemmas cover both nouns, verbs, adjectives

and adverbs, and 86% of them have a lemma match in ODS, confirming that even though the DDO core concept lemmas were selected via and English selection, they are in fact central lemmas also in the Danish language. We excluded senses from fixed expressions in our dataset. Table 1 summarizes the sense statistics of the annotated data set.

Resource	Nouns	Verbs	Adjectives	Adverbs	Other	All
ODS	2176 (282040)	983 (119163)	4 36 (60599)	0 (0)	0 (0)	3595 (461802)
DDO	1036 (12326)	383 (4045)	248 (2228)	0 (0)	0 (0)	1667 (18599)

Table 1: The statistics of the annotated data based on (Ahmadi et al. 2020). The numbers in parentheses refer to the overall number of the tokens in senses.

In the manual annotation task, the hierarchical sense structure in ODS, including main sense as well as subsense numbers, is visible to the annotator while the DDO senses are presented in a random linear order with no information on the original sense numbers and hierarchical relations between senses. This facilitated the manual linking process since cases of potentially very different hierarchies in the two dictionaries did not disturb the picture. See figure 6.

pyramide (sb.)-19036640			
1) (massivt) bygningsværk af sten med firkantet grundflade og trekante	exact	4-	1-todimensional figur der har form som en trekant med s
2) (mat.) legeme, hvis grundflade er en polygon, og hvis trekantede sid	exact	3-	2-bygning el. konstruktion med form som et sådant grav
3) hvad der har form af en pyramide ell. kegle; ogs. i videre anv., om hv	exact	2-	3-rumlig geometrisk figur der fremkommer ved at der fra
3. 1) om (del af et) bygningsværk (tårn, spir olgn.); nu især (jf. Pyramid	narrow	2-	4-egyptisk gravmonument, ofte af meget store dimensio
3. 2) (især gart.) om træer (sjældnere andre planter). Pyramideaster, d	narrow	2-	
3. 3) om (lille) pyramideformet ting ell. figur; fks. til havepynt: små Pyra	narrow	2-	
3. 4) opstabling, opstilling af ting, der har form som en pyramide, tilspic	narrow	2-	
3. 5) om (del af) møbel (hylde, opsats olgn.), der tilspidses opefter; spe			
3. 6) (fagl.) krystalform bestående af to mod hinanden vendte pyramide			
3. 7) (anat.) om forsk. fremspring olgn. d. s. s. Nyrepyramide . Anat.(18			
3. 8) om (konkylier med stærkt opsvulmet nederste vinding af) forsk. f			

Figure 6: The senses of the noun *pyramide* (“pyramid”) in ODS (column 1 to the left) and DDO (column 4 to the right) in the sheet used for the linking task. The linking values (relation, e.g. “exact” and sense number, e.g. “4”) are annotated in the columns 2 and 3. In ODS the original sense numbers and sense order is kept, in DDO the sense numbers are ad hoc, and the order does not correspond to the one in the dictionary.

We operate with the following types of relations between senses in the two dictionaries:

- **none**: There is no match for this ODS sense in DDO
- **exact**: The sense in ODS corresponds to the sense in DDO, for example, the definitions are simply paraphrases, as seen in the examples in Figure 5, or they describe the same concept in rather different ways, as seen in the examples in figure 2, 3 and 4. Senses are also considered to be exact matches in cases where the only difference is due to the modernization of society. E.g. the ODS sense of the noun *passager* (“passenger”) “person traveling with mail coach etc.”, was considered an exact match to the DDO sense “person traveling with private or public means of transportation”.
- **broader**: The sense in ODS completely covers the meaning of the sense in DDO, but is also applicable to further meanings. E.g. the ODS sense of the noun *værge* (“guardian”): “a guardian of anything or anybody” is a broader sense of the DDO sense restricted to “a guardian in legal context” (i.e. a guardian for a child not yet legally competent or for an incapacitated adult).
- **narrower**: The sense in ODS is entirely covered by the sense of DDO, which is also applicable to further meanings. In ODS the adjective *spids* (“sharp”) has, for example, two specific senses, one about a sound and another one about a smell, where DDO covers both senses in one definition: “pungent in an unpleasant way (about smell, taste or sound)”. Therefore, both ODS senses are considered to be narrower than the “lumbed” DDO sense.
- **related**: There are cases when the senses may be related even though the definitions in ODS and DDO differ in key aspects. For example, the property of “being able to sleep”, a sense of the noun søvn (“sleep”) in ODS is considered “related” to “the state of sleeping” sense in DDO, however not identical. The noun *“bamse”* (teddy bear) is in ODS, described as a “fat, clumsy person, especially a child”, is in DDO described as a “fat, good-natured person”, and these two senses are also considered to be related. Also, cases of regular polysemy are considered to be “related” matches. E.g. ODS has only one sense for the noun “*ambassade*” (“embassy”), namely the organization sense, while DDO has two: the organization sense as well, but also the building sense. While the organization sense is an exact match to the sense in ODS, the building sense is considered to be only “related” to it.

3.3 Models

Using the annotated data, we predict the similarity scores between senses using a similarity function. The similarity function is a trained model based on the following similarity features given that A is a sense in the first resource, ODS, and $B$is a sense in the other resource, DDO:

1. **String metrics**
   - **Longest common substring**: the length of the longest substring that exists in both senses
   - **Length ratio**: the ratio of the number of space-separated tokens in each sense
   - **Average word length ratio**: the average length of words in each sense
   - **Jaccard, Dice, and Containment**:

     $J(A,B) = |A \cap B| / |A \cup B|,$
     $D(A,B) = 2|A \cap B| / (|A| + |B|,$
     $C(A,B) = |A \cap B| / min(|A|, |B|).$
   - **Smoothed Jaccard:** this metric is an improved formulation of the Jaccard coefficient that makes the optimization possible and can be adjusted to distinguish matches on shorter texts (McCrae et al. 2017). It is defined as follows:

     $$J_\sigma(A,B) = \frac{\sigma(|A \cap B|)}{\sigma(|A|) + \sigma(|B|) - \sigma(|A| \cup |B|)}$$

     where $\sigma(x) = 1 - exp(-\alpha x)$ and $\alpha$is a constant.
2. **Word Embeddings:** with the current progress in the field of NLP, representing words within vector-spaces has been widely used and is proved to be beneficial in various applications. To evaluate the usability of word embeddings in the task of WSA, we also train a model based on ODS and DDO data using the Global Vectors for Word Representation (GloVe) model (Pennington et al. 2014).
   We took as a starting point the word embeddings model trained at DSL in the DDO project using a corpus of approximately one billion running words of modern Danish. The model is trained with 500 features, a window size of 5 and a minimum occurrence of 5 (any types below this threshold are discarded), and used the Skip-Gram version of the model. See Sørensen & Nimb (2018) for details about the model[3].
   Given V_A and V_B, the corresponding vector representations of each word in our word embeddings for senses A and B, we calculate the similarity between vectors using the cosine similarity as follows:

   Similarity based on the word embeddings: $cos(\theta) = \frac{V_A . V_B}{|V_A||V_B|}$

   where θ is the angle between two vectors projected in a multi-dimensional plane.
3. **Automatic feature extraction:** In this model, we automatically extract useful features from the input data in such a way that the performance of the extracted features is maximal among the whole combination of features.

Once the similarity scores are extracted, we automatically align senses in a bijective and greedy approach where the sense pairs are ordered based on the similarity score and then aligned in such a way that a sense is linked to only one other sense in the other resource. Although this bijective constraint ignores polysemous senses, it yields a more diverse combination of sense matches.

4 Evaluation

We evaluated the performance of the models using NAISC (McCrae & Buitelaar 2018). NAISC[4] is a tool for automatic alignment of lexical and ontological data which can be configured based on various semantic similarity extraction techniques including the ones described in Section 3.3. We use precision, recall and F-measure as our evaluation metrics as described by Nakache et al. (2005).

As discussed in Section 2, senses in ODS are long and unstructured. Therefore, in addition to the original ODS data, we create three other datasets where the number of space-separated tokens is limited to 15, 20 and 25. The performance of our similarity detection models with respect to each dataset is provided in Table 2.

Although the precision of the models in automatically detecting the similarity of two senses varies in a close range of 50.3% (All-auto) and 66.7% (15-Word embeddings), there is more significant difference between the recall of each dataset and so, in F-measure. The lowest recall appears in aligning DDO with ODS with its original senses. In other terms, when senses with all the composing parts, such as usage examples and idioms, are aligned with DDO, all the three models can predict a link over 50% correctly. However, they only succeed in less than 10% of cases to retrieve relevant senses. Truncating senses from 25 tokens to 15 significantly improves both the precision and recall, proving our initial observation of the noisiness of senses in ODS. Figure 7 illustrates the correlation of senses sizes with F-measures in all the models.

ODS sense size	Model	Precision	Recall	F-measure

[3] The paper describes the training of a previous version of the model. However, the only differences are that corpus material for 2018 and 2019 have been added and that the skip-gram version is chosen instead of CBOW.

[4] The tool is openly available at https://github.com/insight-centre/naisc

15	String metrics	65.3%	**48.1%**	55.4%
	Word Embeddings	**66.7%**	48.0%	**55.8%**
	Auto	64.0%	46.6%	54.0%
20	String metrics	61.5%	44.3%	51.5%
	Word Embeddings	64.7%	46.7%	54.3%
	Auto	63.3%	45.8%	53.2%
25	String metrics	57.5%	21.9%	31.7%
	Word Embeddings	55.9%	21.2%	30.8%
	Auto	58.5%	22.2%	32.1%
All	String metrics	54.7%	9.8%	16.7%
	Word Embeddings	50.7%	9.7%	16.3%
	Auto	50.3%	9.4%	15.8%

Table 2: The performance of our similarity detection models for automatic alignment of DDO and ODS within a specific limit of space-separated tokens (15, 20, 25 and all tokens).

The highest F-measure of 55.8% belongs to the ODS dataset with a maximum of 15 tokens and trained with the word embeddings model. In comparison to the baselines presented by Kernerman et al. (2020) where an F-measure of 4.3% is reported, such an improvement is promising.

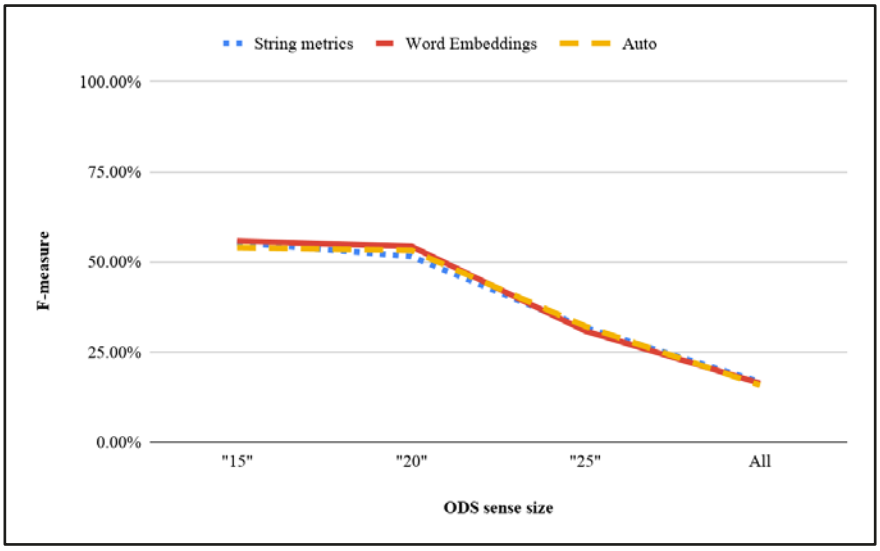

Figure 7: The correlation of sense sizes in ODS with F-measure using various methods.

5 Conclusion

In this paper, we studied the automatic alignment of senses across two Danish dictionaries, ODS and DDO. We demonstrate

that basic string similarity metrics along with word embeddings and automatic feature extraction models can be efficiently used to align senses of identical lemmas across these two resources. Converting printed historical dictionaries into structured electronic forms is an expensive and burdensome task. As future work, we are interested in exploring unsupervised methods to detect sense boundaries in dictionaries such as ODS. Moreover, we would like to explore further methods to automatically detect the type of the semantic relationship that may exist between two senses, also of non-identical lemmas, and study to which degree manual markups of the metainformation in the ODS improve the method, as well..

6 References

Ahmadi, S., McCrae, J. P., Nimb, S., Khan, F., Monachini, M., Pedersen, B. S. & Troelsgård, T. (2020). A multilingual evaluation dataset for monolingual word sense alignment. In *Proceedings of the 12th Language Resource and Evaluation Conference (LREC 2020)*, volume 1, 3232–3242.

Bjerring-Hansen, J., Jelsbak, T., Sørensen, N. H. & Fischer, F. (2019). 'Nodes and Edges in Literary History: Modelling 19th Century Literary Landscapes', *Digital Humanities*, Utrecht, 9-12 July, 2019, Accessed at: https://georgbrandes.dk/research/3explorations/brandes-poster-dh2019-utrecht.pdf [30/05/2020]

Buitelaar, P. (2000). 'Reducing Lexical Semantic Complexity with Systematic Polysemous Classes and Underspecification'. In *Proceedings of the ANLP2000: Workshop on Syntactic and Semantic Complexity in Natural Language Processing Systems*. Seattle, USA. Accessed at: http://dfki.de/~paulb/anlp00.html [30.03.2010].

Dahlerup, V. (1918-54). *Ordbog over det danske sprog, volume 1-28*; *Supplement til Ordbog over det danske Sprog, volume 1-5* (1992-2005). Det danske Sprog- og Litteraturselskab/Gyldendal, Copenhagen. Online: ordnet.dk/ods

Hjorth, E., Kristensen, K. (2003-2005). Den Danske Ordbog, volume 1-6, Det danske Sprog- og Litteraturselskab/Gyldendal, Copenhagen. Online: ordnet.dk/ddo

Jacobsen, L., Juul-Jensen H. (1918). Indledning til Bind 1 in *Ordbog over det danske Sprog, volume 1*, 1918, Det Danske Sprog- og Litteraturselskab, Copenhagen. Accessed at: https://ordnet.dk/ods/tekster-fra-den-trykte-ordbog [30/05/2020]

Kernerman, I., Krek, S., McCrae, J. P., Gracia, J., Ahmadi, S. & Kabashi B. (eds) (2020). Globalex Workshop on Linked Lexicography. *European Language Resources Association (ELRA) - LREC 2020 Workshop Language Resources and Evaluation Conference*, volume 1, 115.

Kilgarriff, A. (2003). "I don't believe in word senses". In B. Nerlich, D. D. Clarke, Z. Todd & V. Herman (eds.), *Polysemy - Flexible Patterns of Meaning in Mind and Language*, Series: Trends in Linguistics. Studies and Monographs [TiLSM], 142, De Gruyter Mouton, pp. 361–392.

Lorentzen, H. (2004). "The Danish Dictionary at large: presentation, problems and perspectives". In W. Geoffrey & S. Vessier (eds.) *Proceedings of the 11th EURALEX International Congress*, Vol. 1, pp. 285-294, Université de Bretagne-Sud, Faculté des lettres et des sciences humaines, Lorient, France.

McCrae, J. P., Arcan, M. & Buitelaar, P. (2017). "Linking knowledge graphs across languages with semantic similarity and machine translation." *Foreword by the chairs*, Vol. 1, p. 31.

McCrae, J. P., Buitelaar, P. (2018). Linking datasets using semantic textual similarity.*Cybernetics and Information Technologies*, 18(1), pp. 109–123.

Miles, A., Bechhofer, S. (2009). SKOS simple knowledge organization system reference. W3C recommendation, 18:W3C. Accessed at: https://www.w3.org/TR/skos-reference/ [30/05/2020]

Moro, A., Li, H., Krause, S., Xu, F., Navigli, R. & Uszkoreit, H. (2013). Semantic rule filtering for web-scale relation extraction. In *International Semantic Web Conference*, volume 1, pp. 347–362. Springer.

Nakache, D., Metais, E. & Timsit, J. F. (2005). Evaluation and NLP. In *International Conference on Database and Expert Systems Applications,* volume 1, pp. 626-632. Springer, Berlin, Heidelberg.

Nimb, S. (2018). The Danish FrameNet Lexicon: method and lexical coverage. In *Proceedings of the International FrameNet Workshop at LREC 2018*, vol. 1, p. 51-55, Miyazaki, Japan.

Nimb. S., Lorentzen, H., Theilgaard, L. & Troelsgård, Th. (2014). *Den Danske Begrebsordbog*, Det Danske Sprog- og Litteraturselskab & Syddansk Universitetsforlag.

Nimb, S. Sørensen N. H. &Troelsgård, T. (2018). "From standalone thesaurus to integrated related words in the Danish Dictionary". In: *Proceedings from Euralex 2018*, volume 1, p. 183, Ljubliana, Slovenia.

Nimb, S. Trap-Jensen, L. & Lorentzen H. (2014). "The Danish Thesaurus: Problems and Perspectives". In: A. Abel, C. Vettori & N. Ralli (eds.), *Proceedings of the XVI EURALEX International Congress: The User in Focus*. 15-19 July 2014. Bolzano/Bozen: EURAC Research, volume 1, pp. 191-199

Palmer, M. (2009). SemLink: Linking PropBank, VerbNet and FrameNet. In *Proceedings of the generative lexicon conference*, volume 1, pp. 9–15. GenLex-09, Pisa, Italy.

Pedersen, B.S, Nimb, S., Asmussen, J. Sørensen, N., Trap-Jensen, L. & Lorentzen H. (2009). "DanNet – the challenge of compiling a WordNet for Danish by reusing a monolingual dictionary". In: *Language Resources and Evaluation, Computational Linguistics Series*, Volume 3, 269-299.

Pedersen, B. S., Nimb, S., Olsen & S, Sørensen, N. H. (2018). "Combining Dictionaries, Wordnets and other Lexical Resources - Advantages and Challenges". In *Globalex Proceedings* 2018, volume 1, p. 102-105, Miyasaki, Japan.

Pedersen, B. S., Nimb, S., Olsen, I. R. & Olsen, S. (2019). "Linking DanNet with Princeton WordNet". In *Global WordNet 2019 Proceedings*, volume 1, 10 p. Wroclaw, Poland.

Pennington, J., Socher, R. & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, volume 1, pp. 1532-1543.

Pustejovsky, J. (1995). *The Generative Lexicon*, Cambridge, MA: MIT Press.

Pustejovsky, J. (1998). "The semantics of lexical underspecification". In *Folia Linguistica 32*(?), pp. 323– 347.

Sørensen, N. H., Nimb, S. (2018). "Word2Dict–Lemma Selection and Dictionary Editing Assisted by Word Embeddings". In: *Proceedings from Euralex 2018*, volume 1, p. 148, Ljubljana, Slovenia.

Svendsen Møller, M.-M., Sørensen, N.H., Troelsgård T. (2020). "An automatically generated Danish Renaissance Dictionary. Building a period dictionary by reducing and merging relevant existing dictionary resources". In *Proceedings of the LREC 2020 Globalex Workshop on Linked Lexicography* (I. Kernerman, S. Krek, J. P. McCrae, J. Gracia, S. Ahmadi & B. Kabashi), European Language Resources Association (ELRA), volume 1, pp. 29-32, Paris, France.

Acknowledgements

This work has received funding from the EU's Horizon 2020 Research and Innovation programme through the ELEXIS project under grant agreement No. 731015. The authors would like to thank Thomas Troelsgård, Society for Danish Language and Literature (DSL) who contributed to the linking of identical lemmas between ODS and DDO and Sussi Olsen, CST, University of Copenhagen who contributed to the manual annotation task.