Lexicography between NLP and Linguistics: Aspects of Theory and Practice

Lars Trap-Jensen
Society for Danish Language and Literature
E-mail: ltj@dsl.dk

Abstract

Over the last hundred years, lexicography has witnessed three major revolutions: a descriptive revolution at the turn of the 20th century, a corpus revolution in the second half of the 20th century, and the digital revolution which is happening right now. Finding ourselves in the middle of a radical change, most of us have difficulties orienting ourselves and knowing where all this is leading. I don't pretend to know the answers but one thing is clear: we cannot ignore it and carry on as normal. In this article, I will discuss how lexicography and natural language processing can mutually benefit from each other and how lexicography could meet some of the needs that NLP has. I suggest that lexicographers shift their focus from a single dictionary towards the lexical database behind it.

Keywords: e-lexicography, NLP, interoperability, data structure

1 Introduction

Let me start with a confession: In my career, I have never been much concerned with natural language processing from a theoretical point of view. The reason is simple: My main interest lies with how natural language works, and I find that formal models of language have so far been unable to come up with convincing explanations of the way language works. This applies both to formal linguistic theories such as generative grammar in the Chomskyan tradition and to computational linguistic models such as the ones used in NLP. In my experience, NLP scholars are not concerned with linguistic theory either, they are, and I apologize for the generalisation, more like engineers: interested in making things work. Coming from computer science, their theoretical interests primarily lie in the mathematical models, algorithms and methods. This is both understandable and quite legitimate. For the same reason, NLP people are excited about big data because it makes their models work. General linguists, on the other hand, are less concerned because big data does not tell us much about how language works. However, when it comes to practical applications, much of the outcome of natural language processing is highly valuable indeed to lexicographers and linguists alike, irrespective of their theoretical positions.

Between NLP and general linguistics is computer linguistics. Computer linguists resemble NLP people in their interest in big data as attested in language corpora and the patterns you can find in them. But some computer linguists are also like corpus linguists interested in the empirical facts about language, in actual language performance as opposed to the more basic underlying linguistic system.

If we relate this to lexicography, we may perhaps draw an analogy to the relation that exists between meta-lexicography and concrete lexicographical products. In the UK, there has always been a strong empiricist tradition, and perhaps this explains why "the British are poor at lexicographic theorizing, but make the best dictionaries" (H. Bergenholtz, personal communication). In my view, a preference for linguistically observational facts is a healthy point of departure, no matter how intuitively it came about in the first place. Another but related aspect is caring about the user's needs. My experience tells me that this is something that everyone has always claimed to take seriously, but there is a huge

difference between deducing user needs from a theoretical starting point and by observing user behaviour. If you take the former standpoint, the user's need is what your theory tells you, or what you can deduce from it. This is why some people from the first camp are so fond of quoting Henry Ford: "If we had asked people want they wanted, they would have said faster horses" (e.g. Tarp 2009: 28), because people do not know what is best for them (a motor car). If you take an empiricist position, what people think and, most importantly, do does matter. To take just one example: if there are words in your dictionary that are never looked up, maybe the time is better spent revising entries that are looked up all the time.

In this paper, I will discuss why it is important to provide for both NLP and human users' needs in our lexicographical practice and propose possible ways in which it could be done. In order to so, I will first place the current situation in the historical development of lexicography over the last hundred years and characterize three major developments in the field. From this follows that the current conditions for making lexicographical products are fundamentally different from what they were only 25 years ago when I entered the field. Most importantly, it also means that we cannot make dictionaries the way we used to, and towards the end I will discuss how we can go about this.

2 Background: Three Revolutions in Lexicography

2.1 The First Revolution: the Descriptive Paradigm

Going back about a hundred years in time, it is not difficult to spot the differences in the dictionary production process. This is around the time when the large national monolingual dictionaries emerged, created with the help of boxes full of excerpts and carried out by an army of hard-working lexicographers who tirelessly produced volume after volume, such as we know it from *Oxford English Dictionary* (OED) in England, the Dutch *Woordenboek van der Nederlandsche Taal*, *Deutsches Wörterbuch* by the Grimm brothers in Germany and *Svenska Akademiens Ordbok* (SAOB) in Sweden. These dictionaries are today widely known and classic works whose acronyms are familiar far beyond the narrow circle of lexicographers. But even though they are old projects and the methods may occur simple from a present-day point of view, they do not differ very much in nature from what we know and use today: descriptions of language and language usage based on empirical analysis – corpus-based as we would call it today. In those days, the corpus consisted of excerpts of language samples stored in file boxes and analyzed for each word and meaning.

In my own country, a national dictionary, *Ordbog over det danske Sprog* (ODS, Dictionary of the Danish Language), was compiled in a similar way, and the method used constituted a real break, a paradigm shift in the classic, Kuhnian sense: ODS broke off from the 19th century tradition and insisted on the descriptive approach. We can call this the descriptive revolution. The tradition prior to the ODS was characterized by what was called 'the academy principle' with reference to the practice typical of the French Academy dictionary. In accordance with this principle, the Danish dictionaries before ODS were prescriptive, and the motivation for their compilation educational: a wish to educate the common people and teach them good and exemplary language through beautiful examples by admired authors and linguistic role models. One of the founding fathers of the *Dictionary of the Royal Danish Academy of Sciences and Letters* (1793-1905), Jacob Langebek, expressed the opinion that the dictionary should only accept words that were "good, pure, generally usable and unmistakably Danish" words (ODS 1918: the preface), whereas he saw no room for:

> All coarse, plump and horny words and words that strive against decency ... for they do not need to be known to those who do not pay heed to them, and those who want to learn them will get to know them anyway" (ODS 1918: the preface).

Consequently, the dictionaries must be selective and include only the good words, whereas the 'coarse, plump and horny' ones have no place in the dictionary.

We find the same way of thinking with another 19th-century Danish lexicographer, Christian Molbech, the author of the most widespread monolingual dictionary of Danish in those days. He said:

> Even the most frequent use of a newly formed word, especially in spoken language, does not yield it any authority, and proves nothing for its usefulness in the pure language and good style, or for its acceptance in a dictionary, if it offends an ear cultivated towards fine language. (Chr. Molbech, the preface from 2nd edition 1859, here quoted from Dahlerup 1907).

The dictionaries should contain only 'good' words, 'the most beautiful flowers in the language'.

> It should thus be an honour for a word to be included in the dictionary, as it is an honour for a piece of art to be admitted into a national arts collection (Dahlerup 1907: 68).

This is how Molbech's principle is expressed by Verner Dahlerup, the founder of the ODS. And Molbech himself writes that his dictionary should be "an interpreter of the proper use of the pure, educated written language of our present time".

A final example of the academy principle comes from the first Danish slang dictionary, *Dictionary of the Vulgar Tongue and so-called Daily Speech*. The author was V. Kristiansen, a pseudonym for Viggo Fausbøll, a professor of Indian and Eastern Philology at the University of Copenhagen. In the preface to the dictionary, he writes that the purpose of his dictionary is not so much explanatory as it is a warning against vulgar language:

> In recent years, this vulgar tongue ... is threatening to force its way into the families ... having collected some of what belongs to it, I have, apart from a purely linguistic aim, in addition wanted to draw attention to the danger and tried to provoke resistance against the same and I assume that once people have opened their eyes to the indecent crossing of the line, all educated people will agree to ban the vulgar tongue from good society and leave it to the guttersnipes and the adherents of Grundtvig in whose taste it may fall (V. Kristiansen 1866).

As a consequence, the most vulgar words were typeset in Greek letters in the first edition so as to prevent common people from exposure to words they were better off not knowing.

Another dictionary that appeared at the same time as ODS began was Dahl and Hammer's *Danish Dictionary for the People*. It is well known for its puristic approach motivated by a desire to educate common people and spare them from seeing unwanted words, whether foreign words, slang, dialect or other types of language considered vulgar or non-standard.

The merit of ODS is that it broke away from this prescriptive tradition. Its founder, Verner Dahlerup, wanted ODS to be a science-based practical tool for language understanding, and to him it did not make sense to exclude words because they were thought to be dubious or destructive:

> First of all, I cannot ask, "should this or that word be used?". I ask instead: "is it used or has it been in use?" If so, I will include the word in so far as it falls within the scope of the dictionary. (Dahlerup 1907: 71).

This quote is taken from an article in the journal *Danske Studier* (Danish Studies) in which Dahlerup explains his ideas about the new dictionary. Dahlerup was not the only one who believed that actual language usage was the key to semantic description. It was a topical belief at the time among the Neogrammarians who had turned to the study of modern languages in their attempt to demonstrate the universality of the phonetic laws. In the latter half of the 19th century, this was an advanced and controversial position in linguistics but one that would eventually pave the way for the new structuralist

paradigm that came to replace comparative linguistics. The principle of empirical analysis as the basis for linguistic description prevails even today, more than a hundred years later. But the basis for empirical analysis and description has changed dramatically.

2.2 The Second Revolution: The Corpus Revolution

The Corpus Revolution is the next major leap that completely changed the way we make dictionaries. In terms of theory and method, excerpted language material and filing boxes are not fundamentally different from corpus concordances and semantic annotation; in a way, it is just a difference between analogue and digital methods. However, with the technical development that took place from the 1960s onwards, a substantial part of the work could be automated with the help of computers, resulting in a substantial expansion of the descriptive basis while at the same time editorial work could be carried out more efficiently.

> The proper way to describe a word is to identify the grammatical constructions in which it participates and to characterize all of the obligatory and optional types of companions (complements, modifiers, adjuncts, etc.) which the word can have in such constructions, in so far as the occurrence of such accompanying elements is dependent in some way on the meaning of the word being described. (Fillmore 1995).

The potential of the use of a corpus is enormous if you take seriously what Fillmore says here, that exhaustive word description implies the investigation of all connections with other words in which a word engages to see if they condition the meaning of that word in one way or another. Fillmore has done so himself in the FrameNet project, and in lexicography we have also seen some projects along similar lines. We will come back to this later.

In lexicography, the corpus revolution took off in earnest in the 1980s with the COBUILD project. At first, the greatest advantage of a corpus was that the sample material multiplied in comparison with the traditional procedure involving language slips and filing boxes. Measured in size, corpora have roughly increased tenfold every 10-15 years. In the 1960s and 1970s, the first corpora had about 1 million words – this is the size of the Brown corpus and also of the Swedish corpus Press65. COBUILD's corpus from 1985 had about 18 million words (today it has developed into the Bank of English with about 650 million words). The British National Corpus reached 100 million words in the mid-1990s. Among the largest corpora today are the German COSMAS (or DeReKo: Deutsches Referenzkorpus) from IDS in Mannheim with more than 5 billion running words, Collins Corpus with over 4.5 billion words, not to mention Google's corpus of texts that have been scanned for the Google Books project comprising more than 500 million books from 1500 to today for a number of the major languages. Different estimates exist, but it is not really important whether it contains 155, 175 or 200 billion words. To most people, the number is incomprehensible anyway or as good as infinite in size.

In my own part of the world, the Nordic region, we cannot quite match the size of the large English and German corpora, but even so the Language Bank in Gothenburg contains more than 1 billion words, and the situation is similar in Norway and Denmark: the Norwegian newspaper corpus contains more than 1 billion words, and the corpus developed at my own institution, the Danish Society for Language and Literature, has also reached a billion words.

The development in the use of corpora for dictionary work is parallel to the development in corpus size. In the early years of the corpus era, the advantage was access to larger sample material than was possible with the help of index cards. For the lexicographer, the task was to browse through concordances and arrange the tokens according to homographs and senses, quite similar to working with

index cards. However, anyone who has worked with concordances will know that even in a moderate corpus, this task becomes overwhelming when you are analysing the common words of a language because the words belonging to the core vocabulary are also frequent and, as a result, give a huge number of concordance lines.

That is why it was a step forward when it became possible to have annotated corpora. Restricting a search to, for example, verbal instances of a homograph or to particular inflectional forms, the lexicographer could skip a lot of unwanted and irrelevant tokens and thereby speed up the work process.

Around the turn of the century, we saw the first syntactically marked-up corpora, which further helped to find distinctive syntactic patterns in the texts, for example in the form of Word Sketches (Kilgarriff et al. 2004) or other forms of lexical profiles. And while the corpora grew in volume, it became increasingly necessary to do something to handle the overwhelming amount of information that came with it. The solution tends to be more and more preprocessing of the material, using techniques that pre-analyze the texts according to different parameters that allow the lexicographers to find just the instances they need in the relevant phase of the editing process. Let us look at some of the opportunities that are explored.

Sorting out corpus instances by homographs and senses is a key element in the daily routine of any lexicographer, and even though we do not yet, to my knowledge, have an operative technique for automatically sorting out concordances semantically, it needs to be mentioned first as this is the ultimate goal. So far, sense annotation is something that editors must do more or less manually. In the current state of corpus linguistics, lexical profiles can help to uncover some meanings automatically, as there is obviously some relation between semantic meaning and, for example, valency patterns or subject domains, both of which can be detected by means of corpus linguistic methods.

Using a corpus as a tool for lemma selection is another obvious possibility. Corpus frequency is one of the parameters used to determine which words should be included in a dictionary.

Statistical methods such as Mutual Information, T-score, log-likelihood etc. are methods that are suitable for demonstrating how words attract each other. Corpus linguistic techniques are therefore likely to be useful in finding and analyzing patterns within the field of phraseology. Valency patterns and lexical profiles can be found in a similar way, although it is not the direct attraction between words that is measured, but the linguistic material of syntactic categories and phrases in a syntactically marked-up corpus.

A fairly new and therefore less well-known application is the use of a corpus to monitor diachronic language development. The general idea is to compare any subset, the so-called focus corpus, against the entire corpus, in this context known as the reference corpus (see Cook et al. 2013). Significant features of the focus corpus can be said to be characteristic of that particular subset. If the focus corpus consists of a single year and we find a number of words and phrases that occur only here and not in the reference corpus, a reasonable hypothesis would be that we have come across neologisms and, possibly, lemma candidates for the dictionary. This is probably the most obvious use of monitor corpora for lexicographical purposes, but in principle anything could be investigated in similar ways. The focus corpus could be a domain specific corpus, and if one finds that common language words suddenly appear in the domain-specific corpus, it could be an indication that something interesting is going on that deserves closer inspection. An example would be if we found words like *mouse, cloud, worm* and *virus* in a computer focus corpus with above-normal frequency. This could be an indication that these common words had undergone some linguistic change in this domain-specific context, an observation that we know, in hindsight, to be true.

The development can also go in the opposite direction: words from a specific domain turn op in the general-language texts. This is the case when we find words and expressions from the world of sports

in general-language texts as a reflection of sports as a productive source of new metaphors: *below the belt, slam dunk, saved by the bell, checkmate, bullseye, the ball's in your court*.

The technique can be used similarly as a tool to mark senses with usage and domain labels. If a comparison reveals that a word or expression is overrepresented in texts from a particular domain, it is likely to be a domain-specific meaning. The method can be applied to any type of variable relating to speaker (age, gender, regional distribution, occupation etc.) or text type (style (formal/informal), channel (spoken/written), public/private, genre etc.) as long as the corpus has been marked up with the relevant metadata.

Finding good language examples is another area where pre-processing has proved successful. This basically means that only the best-suited examples are shown as the result of a corpus query or, alternatively, they are shown at the top of the concordance list. Finding good examples manually is a time-consuming and therefore expensive task, so there is much to be gained if an automatic procedure can find all and only the best examples. What is considered a good example may vary from dictionary to dictionary, depending on the intended user group and the dictionary's distinctive style but they probably have a few characteristics in common: a good example consists of a whole sentence, neither too long nor too short (about 15-20 words), it should not contain proper nouns (because they require cultural knowledge, are short-lived and may be a breach of people's privacy), it should not contain deictic expressions or pronouns referring to something outside the sentence; and it should preferably contain a typical collocation or some other idiomatic pattern, whereas difficult or rare words should be avoided. Criteria like these can be determined in advance and the query result sorted in such a way that the lexicographer will be presented with examples that meet all criteria first and can find an example that is a good candidate for inclusion in the dictionary. Such a feature is integrated with the SketchEngine corpus query tool (GDEX, see Kilgarriff et al. 2008), and a similar system has been developed at the Berlin-Brandenburg Academy of Wissenschaften for German (Didakowski et al. 2012).

To summarize, the corpus revolution can hardly be described as a paradigm shift in the sense of Kuhn. But it brought the descriptive tradition, starting from the beginning of the 20th century, a giant step forward and took lexicography much closer to the goal: the dictionary description as a mirror of language in all its diversity. In the corpus era, description shifted in focus from the exemplary language of classic writers towards capturing more and more language varieties in an attempt to embrace and reflect the entire language. In the 1990s, many believed that representativeness was at least as important as volume, but that idea has now more or less been abandoned, probably due to practical rather than theoretical reasons: balanced corpora are much more difficult and expensive to develop, and none of today's mega-corpora are particularly well-balanced. Instead, they are either dominated by journalistic texts or have been harvested from the web. The keyword is accessibility.

Viewed from the users' perspective, the corpus revolution has hardly been noticed by many outside the lexicographic world because it did not change dictionaries and lexicographical products radically. But for lexicographers, it had a great impact as it improved the descriptional basis and enabled us to make better dictionaries.

The growing size of corpora also changed the working conditions for practitioners in the field. A hundred years ago, the editors would spend most of their time editing an entry, starting from scratch until there was a complete article ready for publication. The only pre-processed material used in the process was the box with excerpts, the rest was the editor's responsibility. This has changed dramatically. In the early days of corpus lexicography, the editor was still to a large degree in charge of the entire process: the corpus was a mere tool, while the editor's responsibility was to read through concordances, sorting and selecting from them. But when one is working with huge corpora, as we

do today, this task becomes increasingly insurmountable. The material is simply too vast. The solution is pre-processing: the editor is presented with semi-manufactured elements, suggestions which the computer has analyzed in advance: spelling, inflection, valency patterns, collocations, idioms, morphological and syntactic restrictions, perhaps quotes and linguistic labels, just to name the most obvious options. The editor's role has changed into one of choosing, checking and validating from the pre-processed material presented to them. Consequently, the skills needed to become a qualified editor have changed: certain skills are no longer required, while others are becoming more important.

2.3 The Digital Revolution

Finally, I am getting to the point with the third major development that can be identified in the last century. It is perhaps also the most difficult to describe as we are still in the middle of it. It is of course what has been called the electronic or digital revolution: the development that has changed the dictionary from an analogous paper product and turned it into something digital, a webpage, an app or an embedded feature somewhere in cyberspace. The development is gradual and has been going on for more than 30 years. The first digitization projects were launched in the early 1980s (SAOB in Sweden, OED in the UK). In the 1990s, CDs and PDAs became popular, offering better search facilities, while content-wise remaining basically the same. During the 2000s, any dictionary with self-respect has either migrated from paper to screen or has gone out of business. Especially in the last 10 years, things have developed rapidly, not least due to the spread of smartphones and tablet computers. Young people today grow up in a world where communication, reading and learning are dominated by computers, and especially on small mobile devices of some sort.

If the influence of the corpus revolution was mostly an internal affair that affected the lexicographic community, the digital revolution has not only been evident to dictionary users but has in effect changed the daily life for every one of us. But it has of course also changed the way we make dictionaries.

In the time of paper dictionaries, there was a close relationship between the contents written and edited by the lexicographers and the finished product. The printed work was the main product, the output created by the lexicographer was an intermediate stage in the process, no matter if it was written on a sheet of paper, in a word processing program or in a database.

Today, we are much more aware of the fact that the database is the central element of the work, and that the database structure must be well organized and sufficiently flexible so as to publish in different media and for different platforms.

The digital revolution has changed dictionaries and dictionary-making in several ways: the business model, improved search possibilities and assistance to language learners and insecure spellers, user involvement and crowd-sourcing, to name but a few of the present challenges. In this context, we are principally concerned with certain aspects: the possibility to access, link and share data with others.

3 The Digital Era

In recent years, lexicography and the NLP community have been brought closer together, in particular for two reasons: the Internet and the use of hyperlinks to connect data and websites have made it possible, and the digital development has made it necessary: seeking information online has taught the users to be impatient. They want their questions answered quickly and they want the answers for free. Dictionaries are no longer the golden eggs they used to be for publishing companies, so they are forced to change their business models if they want to stay in the business while newcomers have entered the scene hoping to get their share.

Where does that leave lexicography in the current situation? On the one hand, existing dictionary providers have migrated from print to screen, having improved data structure and search facilities, perhaps added new information in the form of audio and video clips, in the hope that users will use and appreciate the new products. On the other hand, their efforts seem to have had limited success. Some people are quite pessimistic in their assessment:

> the biggest problem of lexicography is that lexicographic products are no longer perceived as relevant for the vast majority of people. Most people, in fact, do not use dictionaries, and if they need to find help when communicating or when looking for data, they simply use the Internet instead (Simonsen, 2017: 409).

Undoubtedly, this is a generalization that ignores a lot of variation in both behaviour and experience that people have across languages and resources. But it is probably true that young people today are less willing to use dictionaries than the generation before them, and therefore dictionaries are not used as much as they could (or should) be. The answer why this is so, is complex and several factors are involved but the explanation made by Simonsen certainly plays a role:

> why do not people use online or mobile dictionaries? Obviously, there are a number of reasons, but I would argue that the most important reason is that most lexicographic resources are not tool-integrated and not specifically related to the user's job tasks (Simonsen, 2017: 409f.).

When you are in the middle of a transition, standing at the crossroads, it is difficult to predict in which direction the future is pointing. Even so, some tendencies can be traced that we need to take seriously.

Traditional lexicography has been challenged by newcomers from natural language processing offering computationally developed lexicographical resources, either meant for computers, such as lexicons like WordNet, FrameNet and VerbNet and various lexical knowledge bases and ontologies used in the Semantic Web, or resources for human users obtained by combining already existing web resources in new ways, such as BabelNet, TheFreeDictionary.com and others. Also, new resources have been created through collaborative efforts by dedicated volunteers, with Wikipedia and Wiktionary as the best-known examples.

While most people agree that traditional lexicographical resources provide high-quality semantic descriptions of languages, it does not follow from this that traditional resources are the ones that are used in computational lexicography and NLP. In fact, they are not, and the explanation is simple: most traditional dictionaries were developed as self-contained entities, encoded in data structures known alone to the publishing company or the institution responsible and kept by them as a secret.

The world of the digital era, with Semantic Web and linked open data as the current buzz-words, is altogether different. Here the keywords are accessibility and interoperability. Even if we accept the fact that users are becoming increasingly reluctant to look up words in a dictionary, there is no reason to believe that their needs for language assistance have decreased. But we may have to meet their needs in new ways, most obviously by embedding language tools directly in the applications and other computer software.

In general, few users are interested in learning the totality of meanings of a particular word when they are looking up. More likely, they are interested in the meaning of the word in a specific context – the one that caused them to look up. So, if embedding is important, it is equally important to develop techniques to pick just the right lexical unit in a given context.

It has been well known for a long time that word sense ambiguation is a major challenge for natural language processing and computational lexicography. For a computer, it is very difficult to determine whether two words, or rather lexical units, are similar. Consider the following definitions of the adjective *kind*:

- (1) behaving in a way that shows you care about other people and want to help them (Macmillan)
- (2) generous, helpful, and thinking about other people's feelings (Cambridge English Dictionary)
- (3) caring about others; gentle, friendly and generous (OALD)
- (4) saying or doing things that show that you care about other people and want to help them or make them happy (LDOCE)

Taken individually, these definitions are quite acceptable ways of explaining what it means to be kind, and we, as lexicographers, make it a point of honour not to copy a definition from others but phrase it in a style that is characteristic of our dictionary. For a computer, however, they are too different for it to work out that they are different ways of explaining the same word meaning.

For comparison, consider the following definitions:

- (1) saying kind things to someone who has problems and behaving in a way that shows you care about them (LDOCE, *sympathetic*)
- (2) kind, helpful, and sympathetic towards other people (Macmillan, *caring*)
- (3) behaving in a pleasant, kind way towards someone (Cambridge, *friendly*)
- (4) (of a person) kind, friendly and sympathetic (OALD, *warm-hearted*)

Likewise, it is difficult or even impossible for a computer to tell that these explanations differ from the previous ones and are in fact explanations of four different but semantically related words: *sympathetic, caring, friendly* and *warm-hearted,* respectively.

If a computer should identify and link words like these in, say, a new and automatically generated resource, it needs a helping hand. That it is why resources like WordNet and FrameNet have become popular in natural language processing, because they do just this: label the semantic relations between words in an explicit way. In WordNet, *kind*, *sympathetic*, *caring*, *friendly* and *warm-hearted* can easily be identified as synonyms or near-synonyms if they belong to the same or neighbouring synsets.

Seen from the computer's point of view, the solution would be to give all words that are labelled synonyms identical definitions. In a computer lexicon, this would make sense: if meanings are synonymous, they should have the same denotation and so their definitions should reflect this fact.

Luckily (for the human user at least), this is to my knowledge not carried into effect in any existing dictionary. Few words, if any, are totally congruent in meaning anyway, and in a dictionary for human users, the most important thing is that the definition is well-phrased and is functioning in itself.

A similar case which is frequently mentioned as a problem for NLP exploitation of lexicographical data is Apresjan's notion of regular polysemy: the fact that the same meanings can regularly be identified for whole groups of words. For instance, words like *hospital*, *school*, *office*, *supermarket* etc. can all have the senses 1) a building ('she went into the office'), 2) the people working there ('the hospital decided to close the clinic') and 3) an institution or business ('the highest-ranking schools in the country'). Or a food container can refer to either the physical object ('a glass', 'a bottle') or its content ('they had two glasses and left'). In a computer lexicon, this would imply that these senses would have to be present for all hyponyms, or subordinate terms, of school, including *pre-school*, *high school*, *secondary school* and *summer school*. In real life – which in this connection means in concrete dictionaries – they are typically not, either accidentally if the lexicographer estimates that they need not be elaborated (*secondary school*: a school for children between the ages of 11 and 16 or 18, Macmillan), or deliberately so if some of the meanings happen to be too infrequent in the underlying corpus for them to be described in the dictionary.

The solution is, as I see it, in neither case to bring the dictionary data in harmony with the demands of the computer. If a definition works fine in itself and is helpful to its user, there is no reason to

standardize it just to make it compliant with the definition of a synonym. And there is no need to describe an infrequent sense just because it can be inferred from a superordinate term. Instead, if lexicographers want to make their data applicable for further exploitation and enrichment by the NLP community, there are several other ways of proceeding, and probably even more than a non-expert such as myself can imagine.

However, over the last 25 years I have seen what (my colleagues') foresight can bring, and if others can learn from that, I am happy to pass it on. Let us look at a few examples.

In the 1990s, during the preparation of The Danish Dictionary, we decided to create a separate element in the data structure devoted to the genus proximum of a word. Here the lexicographer would enter the nearest hyperonym (or superordinate term) of the lexical unit in question. At the time, we only had vague ideas about the future use of such an element, and it was of course not visible in the final printed dictionary.

Similarly, a systematic domain element was created where the lexicographer would assign a domain label to a lexical unit if at all possible. This element should not be confused with the traditional domain label. Whereas the domain label is visible in the dictionary and serves the function of marking a sense as a technical term, the systematic domain element was never intended for publication and simply makes explicit what part of the vocabulary a particular sense belongs to, whether used in common language or as a technical term.

These two elements turned out to be highly useful when years later we used the data from The Danish Dictionary to create the Danish WordNet, DanNet, in collaboration with Copenhagen University (Pedersen et al. 2009). It was also crucial for the decision to create the WordNet from scratch rather than translating from Princeton WordNet. Later, we used the structure of the WordNet to help us organize the meanings of The Danish Dictionary conceptually when we edited a Danish thesaurus (Nimb et al. 2014). And most recently, the data from the Danish Thesaurus have been used to develop a Frame Lexicon for Danish in another collaborative project with Copenhagen University (Nimb 2018).

4 Lexicographical Solutions

It is no coincidence that we are witnessing an increasing need for high-quality lexicographical data. Language technology and artificial intelligence are moving into a phase where the word lists and morphological lexicons developed inside the NLP environment itself are insufficient to meet the demands for developing smarter and more sophisticated products. Automatic content summaries, domain classification and virtual assistants are but a few examples of applications that require 'knowledge' or some way of handling the semantics of human language. By far the best existing semantic descriptions of language are dictionaries, and for that reason, it is obvious that existing dictionaries are interesting for developers of such applications. But it is also obvious that the structure and nature of data used for computer dictionaries are different than data for human dictionaries. The justification of a project such as ELEXIS (for example Pedersen et al. 2018) is exactly to bring these two communities together and explore how existing lexicographic descriptions in human dictionaries can be accessed and converted in such a way that they can serve as input to computers.

My point in this connection is to stress that convergence should take place in both directions and that there are several things that traditional lexicographers can do to meet the needs of language technology. First and foremost, it is necessary that lexicographers shift their focus away from the concrete end product and towards a lexical database that can serve both worlds at the same time. My expertise

does not suffice to make exhaustive suggestions what such a database should ideally look like, but my hope is that experts from both sides will join forces and agree on the minimum requirements and propose useful ideas. The recommendations could include the following – and probably many more:

4.1 Accessibility

For data to be accessible, it is not enough that data owners are willing to share them with colleagues or customers on certain conditions (that can be specified in a license), the data itself must also be recognizable when exchanged. This is why a lot of discussions seem inevitable in any infrastructure project: we need to agree on certain data formats and standards if we want to profit from the benefits of exchanging data with others. If the data do not follow any of the current standards, at least they should be capable of conversion for export and exchange purposes.

4.2 Unique Identification

Trivial as it may sound, it is important that the central units of the database can be uniquely identified. Traditionally, headwords have been the central units of dictionaries, but headwords are far from always enough to identify all occurrences of the words in a dictionary since many words can be either homographic or polysemous or both. When we say that *pale* and *light* are synonyms, this is, strictly speaking, not a relation between these two words themselves, but rather between the two words in one of their respective meanings. And when we talk about a *pale imitation*, we mean not 'pale' and 'imitation' as lemmas, but in a specific meaning of the two words. So, if we want to be able to identify word occurrences uniquely (and for computer purposes this is necessary), we must be able to point to the combination of lemma form and meaning. This combination is also known as a lexical unit, and undoubtedly it is wise to use a unique ID number for each lexical unit in the database. In this way, it would be possible to mark up all the words used in a dictionary (including defintions and examples) with a unique ID number, something which would be of help to a computer when given data as input for further processing.

4.3 Data Structure

In the same way as hidden elements for genus proximum and systematic domain assignment were used in The Danish Dictionary, it is possible to add further elements or attributes to the database in the form of ID numbers or links to other relevant resources. This could be links to a synset in WordNet or to the proper WordNet super sense, it could be to a particular frame in FrameNet, etc. In this way, the resource would become gradually integrated with other resources, and this is indeed the whole idea behind linked open data and other data-exchanging communities.

4.4 Consistency

Computers require more consistent data than humans, and for this reason the database should contain consistent descriptions compliant with systematic polysemy, rules of inheritance throughout the ontological hierarchies, etc. However, none of this needs to be visible in an extract of the database used to publish a traditional dictionary for humans as long as the elements in the database are clearly marked and consistently used for computer purposes alone.

5 Conclusions

By no means an NLP expert myself, I am sure that others will have both more and better suggestions on what it takes to improve language technology and how lexicography can contribute. There

is, however, little doubt in my mind that much is to be gained if lexicographers are willing to accede to the demands made by computational processing. These could involve, but do not exhaust, the following steps:

- (a) Use a lexical database for your data that is independent of the particular product that you are working on. But make sure to organize the database in such a way that it permits extraction of exactly the kind of information you need for a concrete dictionary to a given target audience.
- (b) Use a standard format (such as TEI) to make your data easy to export and modify when exchanged with others.
- (c) Use ID numbers to uniquely identify the central elements of the database; most often these will be the lexical units: a headword in one of its senses. Often this is also useful for internal purposes because ID markup of all the words in the dictionary, including definitions, synonyms, collocations and fixed phrases, is required for unique links between the words.
- (d) Use elements (or attributes) in the database that could be useful for NLP purposes: genus proximum, systematic domain assignment, ontological type and super senses have been mentioned above, but the only limitation is your imagination, and the NLP community is invited to make further suggestions to their needs.
- (e) Use attributes (or elements) that make explicit the exact position or relation of a lexical unit to one or more existing external NLP resources such as WordNet, FrameNet or VerbNet.

What has been said above is by no means in itself decisive, what matters is that NLP specialists and lexicographers realize that they are dependent on each other and must work together to find new ways of making lexical resources for the next generation. One such initiative is ELEXIS, and hopefully this project will pave the way for further co-operation and mutual understanding between the two fields. It is not necessarily, or at least not only, a question of giving NLP free and open access to data developed by lexicographers. It is as much a question of lexicographers realizing that their data are not exclusively made for a specific end-product, a traditional dictionary. The data are equally important as input for NLP and must be be structured optimally to be suited for that purpose. In fact, we as lexicographers may soon come to realize that traditional dictionaries are no longer in demand. The users of tomorrow's computers may want their linguistic problems solved by simply talking to their device: "What does X means in this sentence?" or "give me another word for *kind*". In order for us to provide the answer, it is likely that a new division of labour is needed where lexicographers maintain the underlying database while other experts are responsible for the access paths and interfaces that connect the database content with the user. Lexicography and NLP must get together and work on joint solutions to meet the challenges of the digital era. This is crucial for the continued prosperity for lovers of language, words and reference works.

References

Apresjan, J. D. (1974). Regular Polysemy. In Linguistics, Volume 12, Issue 142, pp. 5–32.

Cambridge English Dictionary. Accessed at: https://dictionary.cambridge.org [30/04/2018].

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

Dahl, B.T. & Hammer, H. (1907–14): Dansk ordbog for Folket I–II (Danish Dictionary for the People). Copenhagen and Kristiania: Gyldendalske Boghandel, Nordisk Forlag.

Dahlerup, V. (1907). Principer for ordbogsarbejde (Principles of Lexicographical Work). In Danske Studier 1907, 65–78.

DanNet. Accessed at: http://wordnet.dk [30/04/2018].

Dansk Ordbog udgiven under Videnskabernes Selskabs Bestyrelse 1–8 (Danish Dictionary published under the Guidance of the Royal Danish Academy of Sciences) (1793–1905). Copenhagen.

Den Danske Ordbog, DDO (The Danish Dictionary). Accessed at: https://ordnet.dk/ddo [30/04/2018].
Didakowski, J., Lemnitzer, L., Geyken, A. 2012. Automatic example sentence extraction for a contemporary German dictionary. In Fjeld, R.V., Torjusen, J.M. (eds.) Proceedings of the 15th EURALEX International Congress. Department of Linguistics and Scandinavian Studies, University of Oslo, pp. 343–349.
Fillmore, C.J. (1995). The Hard Road From Verbs To Nouns. In M. Chen & O. Tzeng (eds.) In honor of William S-Y. Wang. Taipei, Taiwan: Pyramid press, 105–129.
Grimm, J. & Grimm, W. Deutsches Wörterbuch. Accessed at: http://woerterbuchnetz.de/cgi-bin/WBNetz/wbgui_py?sigle=DWB [30/04/2018].
Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. In Williams, G. & Vessier, S. (eds.) Proceedings of the Eleventh EURALEX International Congress. Lorient, France, pp. 105–115.
Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, M. (2008). GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In Bernal, E. & DeCesaris, J. (eds.) Proceedings of the XIII EURALEX International Congress. Barcelona: Universitat Pompeu Fabra, pp. 425–433.
Kristiansen, V. (1866). Bidrag til en Ordbog over Gadesproget og saakaldt Daglig Tale (Contributions to a Dictionary of the Common Language and So-Called Vulgar Tongue). Copenhagen: Boghandler H. Hagerups Forlag.
Longman Dictionary of Contemporary English, LDOCE. Accessed at: https://www.ldoceonline.com [30/04/2018].
Macmillan English Dictionary Online. Accessed at: https://www.macmillandictionary.com [30/04/2018].
Molbech, C. (1859): Molbechs ordbog 1–2. Copenhagen: Gyldendalske Boghandlings Forlag.
Nimb, S., (2018). Fra begrebsordbog til FrameNet (From Thesaurus to FrameNet). In Nielsen, J.G., Petersen, K.S. (eds.) DSL's Årsberetning 2017–2018 (Annual Report of DSL 2017-2018). Copenhagen: Society for Danish Language and Literature, pp. 80–86.
Nimb, S., Trap-Jensen, L. & Lorentzen, H. (2014). The Danish Thesaurus: Problems and Perspectives. In Abel, A., Vettori, C. & Ralli, N. (eds.) Proceedings of the XVI EURALEX International Congress: The User in Focus. 15–19 July 2014. Bolzano/Bozen: EURAC Research, pp. 191–199.
Ordbog over det danske Sprog, ODS (Dictionary of the Danish Language) (1918–1956). Copenhagen: Society for Danish Language and Literature and Gyldendal Publishers. Accessed at: https://ordnet.dk/ods [30/04/2018].
Ordbok över svenska språket utgiven av Svenska Akademien (Dictionary of the Swedish Language published by the Swedish Academy), SAOB (1898–). Accessed at: https://www.saob.se [30/4/2018].
Oxford English Dictionary, OED. Accessed at: www.oed.com [30/04/2018].
Oxford Advanced Learner's Dictionary, OALD. Accessed at: https://www.oxfordlearnersdictionaries.com/definition/ english [30/04/2018].
Pedersen, B. S., McCrae, J., Tiberius, C., & Krek, S. (2018). ELEXIS – a European infrastructure fostering cooperation and information exchange among lexicographical research communities. In Proceedings of Global WordNet Conference 2018. Singapore.
Pedersen, B. S., Nimb, S., Asmussen, J., Sørensen N.H., 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, Volume 43, Number 3, Springer Netherlands, pp. 269–299.
Simonsen, H.K. (2017). Lexicography: What is the Business Model? In Kosem, I., Tiberius, C., Jakubíček, M., Kallas J., Krek, S. & Baisa, V. (eds.) Electronic Lexicography in the 21st Century: Proceedings of eLex 2017 conference. Brno: Lexical Computing CZ s.r.o., pp. 395–415.
Tarp, S. (2009). Beyond Lexicography: New Visions and Challenges in the Information Age. In Bergenholtz, H., Nielsen, S. & Tarp, S. (eds.) Lexicography at a Crossroads. Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow. Bern: Peter Lang AG, International Academic Publishers, pp. 17–32.
Trap-Jensen, L. (2014.) Leksikografisk tradition og fornyelse: tre revolutioner på 100 år? (Lexicographic tradition and innovation: three revolutions in 100 years). In Fjeld, R.V. & Hovdenak, M. (eds.) Nordiske Studier i leksikografi 12. NFL-skrift nr. 13. Oslo: Novus Forlag, pp. 42–68.
Woordenboek der Nederlandsche taal. Accessed at: http://gtb.inl.nl/search [30/04/2018]]