Lexicography in Language Technology (LT)

Bolette Sandford Pedersen, University of Copenhagen

1. Introduction

A general intuition is that lexical resources used for LT applications need not differ radically from conventional dictionaries. The required linguistic information is basically the same - even if computer systems require that linguistic features be spelled out to a larger extent. So where dictionaries rely heavily on human pragmatic knowledge and the language-user's ability to make assumptions, computers generally call for information which is highly explicit and consistent.

Another difference that could be pointed out is the fact that machines generally require that the resources be maximally descriptive (i.e. have a high coverage) in order to perform well, whereas humans tend to focus on the degree of normalization and generalization of the dictionary. This is an interesting difference which pinpoints why lexical resources in LT will always direct themselves towards and be more dependent on large, continuously updated corpora. In fact, parts of the LT community prefer to rely rather on corpus data alone than on lexical resources, a tendency that has grown in recent years with improved unsupervised and semi-supervised machine learning techniques. For instance, Google Translate claims not to use dictionaries, but rather to learn bilingual equivalents from aligned parallel corpora. This being said, there is a general acceptance in the LT community of the fact that high-quality, large-coverage lexicons *do* improve the output quality of the applications that integrate them. To this end, lexicography is still a central issue in LT, of which the growing interest in semantic lexical resources such as wordnets and the use of Wikipedia as a lexical resource can be taken as clear signs.

While LT has evolved towards applying more and more statistical approaches, lexicography has changed equally together with modern corpus and compilation facilities. Automatically generated word profiles built on the basis of statistically processed corpus data now enable the lexicographer to deduce the meaning of words from words in use in a much more efficient and reliable way. This job was previously done by

manually going through large amounts of corpus concordances and grouping them on a more or less idiosyncratic basis.

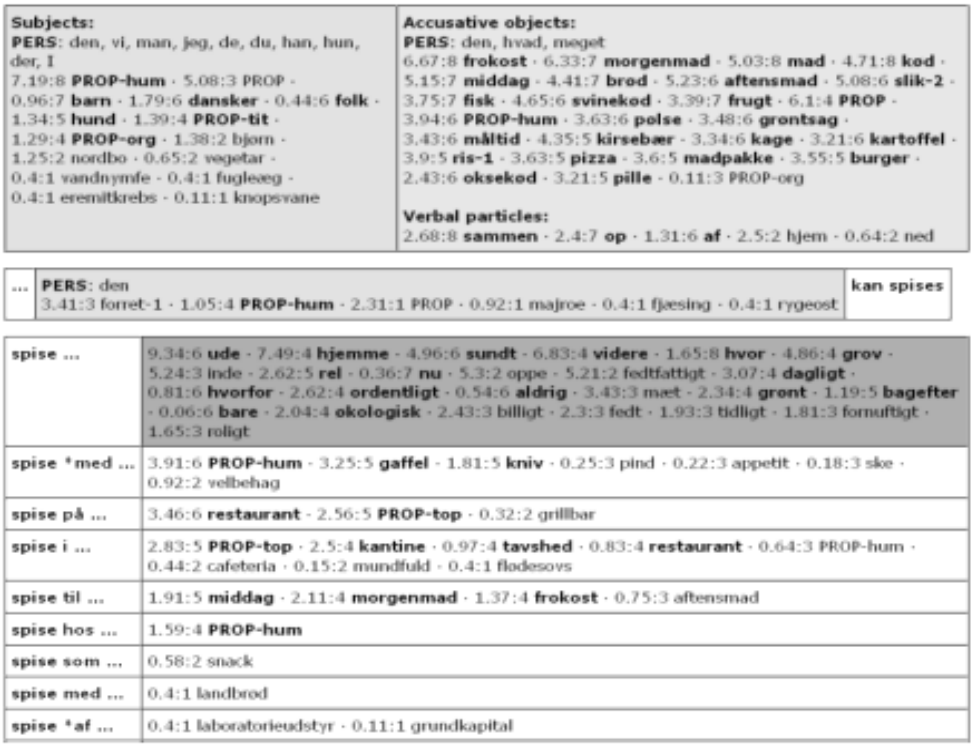
With help from these new corpus technologies, dictionaries for humans and lexical resources for machines are by and by changing into relying more and more on a common lexical core based on large amounts of corpus data. Thus, we now see several examples of large collections of lexical information from where appropriate excerpts can be extracted for different concrete purposes – be it paper dictionaries or resources as input to specific technological applications.

In the following, we give examples of several such emerging resources, and we present in more detail aspects of the Danish wordnet, DanNet, which is a resource designed particularly for LT applications, but which shares sense IDs and semantic information with two other Danish lexical resources, namely The Danish Dictionary (DDO) and the currently developed Danish Thesaurus (DT). In this core of lexical information, semantic knowledge is being shared and extended among the three different practical resources enriching by and by the common core so that it includes now a rich base of ontological, syntagmatic as well as thematic information. We further claim that this composite core constitutes the ideal basis for a well-founded, semi-automatic clustering of senses for use in a particularly hard LT task, namely that of automatic *word sense disambiguation*.

2. Examples of lexical cores

2.1. *Word profiles as a way to achieve lexical cores*

During the last decade, several statistically based corpus technologies have been developed which enable us to automatically generate word profiles on the basis of grammatically analyzed corpora. A general approach in these systems is to use the co-occurrences between mother-daughter dependency pairs to compile the lexical profiles of typical collocations and valency patterns, such as subject verb, object verb, verb + particle etc. Examples of such systems are SketchEngine (Kilgarriff and Rundell 2002), Worschatz (Biemann et al. 2004) and DeepDict (Bick 2009). DeepDict further applies semantic prototypes in order to generalize over semantically similar types of words, such as HUM and ORG for humans and organizations, respectively.



Subjects:	Accusative objects:
PERS: den, vi, man, jeg, de, du, han, hun, der, I 7.19:8 **PROP-hum** · 5.08:3 PROP · 0.96:7 **barn** · 1.79:6 **dansker** · 0.44:6 **folk** · 1.34:5 **hund** · 1.39:4 **PROP-tit** · 1.29:4 **PROP-org** · 1.38:2 bjørn · 1.25:2 nordbo · 0.65:2 vegetar · 0.4:1 vandnymfe · 0.4:1 fugleæg · 0.4:1 eremitkrebs · 0.11:1 knopsvane	**PERS**: den, hvad, meget 6.67:8 **frokost** · 6.33:7 **morgenmad** · 5.03:8 **mad** · 4.71:8 **kød** · 5.15:7 **middag** · 4.41:7 **brød** · 5.23:6 **aftensmad** · 5.08:6 **slik-2** · 3.75:7 **fisk** · 4.65:6 **svinekød** · 3.39:7 **frugt** · 6.1:4 **PROP** · 3.94:6 **PROP-hum** · 3.63:6 **pølse** · 3.48:6 **grøntsag** · 3.43:6 **måltid** · 4.35:5 **kirsebær** · 3.34:6 **kage** · 3.21:6 **kartoffel** · 3.9:5 **ris-1** · 3.63:5 **pizza** · 3.6:5 **madpakke** · 3.55:5 **burger** · 2.43:6 **oksekød** · 3.21:5 **pille** · 0.11:3 PROP-org **Verbal particles:** 2.68:8 **sammen** · 2.4:7 **op** · 1.31:6 **af** · 2.5:2 hjem · 0.64:2 ned

...	**PERS**: den 3.41:3 fornet-1 · 1.05:4 **PROP-hum** · 2.31:1 PROP · 0.92:1 majroe · 0.4:1 fjæsing · 0.4:1 rygeost	**kan spises**

spise ...	9.34:6 **ude** · 7.49:4 **hjemme** · 4.96:6 **sundt** · 6.83:4 **videre** · 1.65:8 **hvor** · 4.86:4 **grov** · 5.24:3 inde · 2.62:5 **rel** · 0.36:7 **nu** · 5.3:2 oppe · 5.21:2 fedtfattigt · 3.07:4 **dagligt** · 0.81:6 **hvorfor** · 2.62:4 **ordentligt** · 0.54:6 **aldrig** · 3.43:3 mæt · 2.34:4 **grønt** · 1.19:5 **bagefter** · 0.06:6 **bare** · 2.04:4 **økologisk** · 2.43:3 billigt · 2.3:3 fedt · 1.93:3 tidligt · 1.81:3 fornuftigt · 1.65:3 roligt
spise *med ...	3.91:6 **PROP-hum** · 3.25:5 **gaffel** · 1.81:5 **kniv** · 0.25:3 pind · 0.22:3 appetit · 0.18:3 ske · 0.92:2 velbehag
spise på ...	3.46:6 **restaurant** · 2.56:5 **PROP-top** · 0.32:2 grillbar
spise i ...	2.83:5 **PROP-top** · 2.5:4 **kantine** · 0.97:4 **tavshed** · 0.83:4 **restaurant** · 0.64:3 PROP-hum · 0.44:2 cafeteria · 0.15:2 mundfuld · 0.4:1 flødesovs
spise til ...	1.91:5 **middag** · 2.11:4 **morgenmad** · 1.37:4 **frokost** · 0.75:3 aftensmad
spise hos ...	1.59:4 **PROP-hum**
spise som ...	0.58:2 snack
spise med ...	0.4:1 landbrød
spise *af ...	0.4:1 laboratorieudstyr · 0.11:1 grundkapital

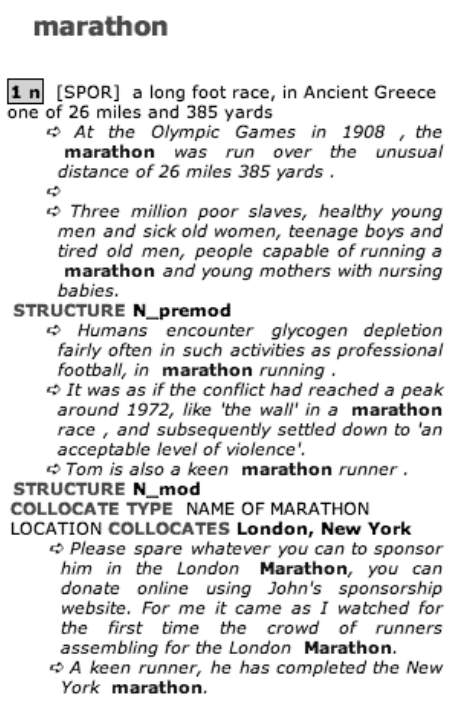
Figure 1: The verb *spise* (to eat) as compiled by DeepDict

Consider in Figure 1 the lexical profile of the Danish verb *spise* (to eat) where we are informed about the prototypical personal pronouns that function as subject and object, as well as about the semantic prototype of the most typical subject (HUM). Further we are given the typical particles (such as *spise op* – to finish) and prepositional phrases that collocate with *spise* such as *med* (with), *i* (in) etc. Typically the automatically generated profiles require human inspection in order to spot bias to the prototypical findings. Such are for instance collocations and undetected named entities which materialize very substantially in the statistics. An example shown in Figure 1 is the Danish film title "I Kina spiser de hunde" ('In China they eat dogs') which by incident is highly represented in the corpus and therefore biases the profile. By clicking on the actual corpus occurrences, however, such deviations from the prototypical distribution are generally easily identified. Since the corpus examples are still directly viewable in these tools, but just systematically

processed, lexicographers generally report such profile tools to be extremely useful in the dictionary making process.

2.2. *Examples of lexical cores in English*

The DANTE database (Atkins 2010) is a lexical database which provides a fine-grained, corpus-based description of the core vocabulary of English. The database is constructed by using the above mentioned Sketch Engine for corpus-querying. Rundell (2011) reports that the database contains semantic, grammatical, combinatorial, and text-type characteristics of more than 42,000 (single) words, 23,000 compounds and phrasal verbs as well as a number of idioms and phrases (27,000). Even if the DANTE project was initiated with a particular dictionary in mind, its focus on the methodological innovations in lexicographical work is expected to raise interest also for other uses, among these LT applications. Figure 2 shows the entry for *marathon* in DANTE with definitions, syntactical information, information on collocations, and a rich supplement of corpus examples.


marathon

1 n [SPOR] a long foot race, in Ancient Greece one of 26 miles and 385 yards

- ➪ *At the Olympic Games in 1908 , the* **marathon** *was run over the unusual distance of 26 miles 385 yards .*
- ➪
- ➪ *Three million poor slaves, healthy young men and sick old women, teenage boys and tired old men, people capable of running a* **marathon** *and young mothers with nursing babies.*

STRUCTURE N_premod

- ➪ *Humans encounter glycogen depletion fairly often in such activities as professional football, in* **marathon** *running .*
- ➪ *It was as if the conflict had reached a peak around 1972, like 'the wall' in a* **marathon** *race , and subsequently settled down to 'an acceptable level of violence'.*
- ➪ *Tom is also a keen* **marathon** *runner .*

STRUCTURE N_mod
COLLOCATE TYPE NAME OF MARATHON
LOCATION **COLLOCATES London, New York**

- ➪ *Please spare whatever you can to sponsor him in the London* **Marathon**, *you can donate online using John's sponsorship website. For me it came as I watched for the first time the crowd of runners assembling for the London* **Marathon**.
- ➪ *A keen runner, he has completed the New York* **marathon**.

Figure 2: Lexical entry in DANTE database

Comparable to this resource although much smaller is the *Corpus Pattern Analysis* initiated by Patrick Hanks as seen at nlp.fi.muni.cz/projects/cpa/. This resource of currently 720 verbs has a basic principle to discover how exactly meanings arise from patterns of usage (words in context), rather than treating words as isolable building blocks in a compositional structure.

89% **[[Human | Animal]] yawn [NO OBJ]**
[[Human | Animal]] involuntarily opens their mouth wide and inhale deeply
[[Human | Animal]] does this typically due to tiredness or boredom

7% **[[Human 1]] yawn [NO OBJ] {at [[Human 2]] | at [[Concept]] | at [[Artifact]] | at [[Speech Act]]}**
[[Human 1]] regards [[Human 2 | Concept | Document | Speech Act]] as pointless and boring
[[Human 1]] shows that [[Human 2 | Concept | Document | Speech Act]] is a nuisance by opening mouth

Figure 3: The patterns of *yawn* as described in Corpus Pattern Analysis

In Corpus Pattern Analysis, meanings of words are not identified as a word in isolation. Instead, meanings are claimed to be associated with prototypical sentence contexts. To this end, concordances are grouped into semantically motivated syntagmatic patterns. Associating the meanings with each pattern is a secondary step, carried out in close coordination with the assignment of concordance lines to patterns. Figure 3 illustrates the approach with the verb *to yarn* where two meanings are deduced from the corpus patterns.

Finally, initiatives of linking lexical resources developed in the LT community in previous years should be mentioned since some of them are currently taken the shape of lexical cores. One example is the Unified Verb Index (verbs.colorado.edu/verb-index/index.php) which co-indexes the resources VerbNet, FrameNet, PropBank and OntoNotes sense groupings. Other initiatives consider the alignment of Princeton WordNet with FrameNet as seen in Ferrández et al. (2010). Figure 4 shows an example from VerbNet with formalized patterns for roles and compositional semantics.

Class Hit-18.1			
Roles and Restrictions: Agent[+int_control] Patient[+concrete] Instrument[+concrete]			
Members: bang, bash, hit, kick, ...			
Frames:			
Name	Example	Syntax	Semantics
Basic Transitive	Paula hit the ball	Agent V Patient	cause(Agent, E)manner(during(E), directedmotion, Agent) !contact(during(E), Agent, Patient) manner(end(E),forceful, Agent) contact(end(E), Agent, Patient)

Figure 4: VerbNet: Patterns for verbs of the *hit*-class

2.3. *Examples of Swedish and Dutch lexical cores: SALDO and Cornetto*

SALDO (Borin and Forsberg 2009, spraakbanken.gu.se/resurs/saldo) is a Swedish semantic and morphological lexical resource primarily intended for use in LT applications, which however, is closely entangled with two paper dictionaries. Thus, SALDO is produced from Svenskt Associationslexikon and was further enriched from Svensk Ordbok. The SALDO editors consider the resource as a basic lexical resource for a Swedish BLARK (Basic LAnguage Resource Kit). The recent compilation of the Swedish wordnet, Swesaurus, using SALDO as the lexicographical core illustrates this function (Pedersen et al. 2012). The SALDO resource currently comprises more than 100,000 entries and is thereby one of the largest Swedish lexical resources. Figure 5 shows the verb *parkera* (to park) and its related concepts such as *parkering* (parking) and parkerbar (‘parkable’).

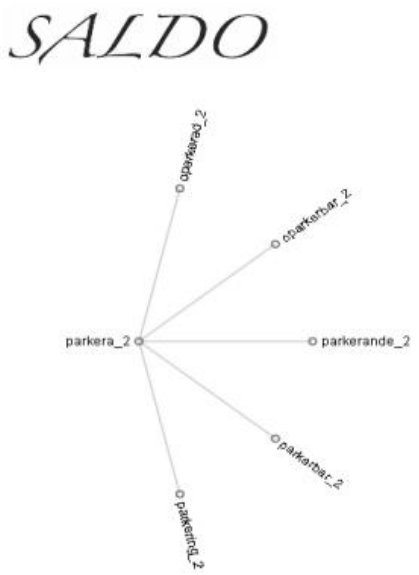


Figure 5: Visualization of the verb *parkera* (to park)

Similar to SALDO, Cornetto stands for Combinatorial and Relational Network as Toolkit for Dutch Language Technology and is a lexical semantic database that combines a wordnet with framenet-like

information for Dutch (cf. www2.let.vu.nl/oz/cltl/cornetto, Vossen et al. 2008). The combination of the two lexical resources (the Dutch wordnet and the Referentie Bestand Nederlands) is claimed to provide a richer relational database to be used in LT, such as word sense disambiguation and language-generation systems. The database is mapped to a formal ontology to provide a more solid semantic backbone. The database represents different traditions and perspectives of semantic organization; therefore, the concepts are aligned across the resources.

2.4. *A Danish lexical core: A dictionary, a thesaurus and a wordnet*

What we label here 'a Danish lexical core' consists of three interlinked resources with common sense IDs, as is the case of Cornetto for Dutch. Two of these resources are meant for human use and are compiled at the Danish Society for Language and Literature (DSL), whereas the wordnet, DanNet, is a collaborate project between DSL and the University of Copenhagen. DanNet is expressed in OWL (Ontology Web Language) and compiled for LT purposes. The transfer of data between the resources is performed in a triangle-like fashion, illustrated by Figure 6, where information was recompiled from DDO when constructing the Danish wordnet, DanNet (cf. Pedersen et al. 2009), and where the DT, which is now in progress, was initiated on the basis of the taxonomical structure of DanNet. Currently, thematic and syntagmatic information is transferred from DT to DanNet (Nimb and Pedersen 2012, Nimb and Pedersen submitted) enriching thereby the original wordnet with valuable information as described further below.

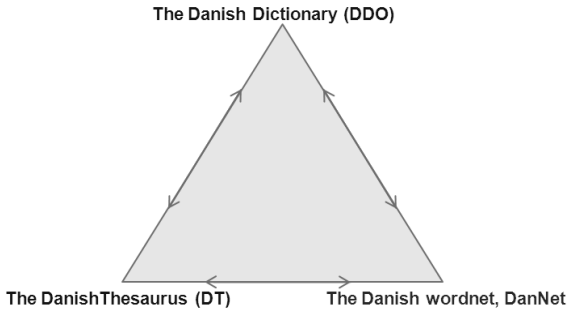


Figure 6: Danish lexical core triangle

3. Information transfer applying the Danish lexical core

3.1. *Transfer of thematic and syntagmatic information*

As indicated above, DDO formed the initial source for the Danish wordnet, DanNet, since all genus specifications in DDO were taken over in DanNet after a manual adjustment. This adjustment focused primarily on the disambiguation of ambiguous genus lemmas and harmonization of inadequate hyponymies. For instance, a lemma like *sauterpande* (sauté pan) would have the following definition: *pande med høje kanter og evt. låg til at sautere kød, grøntsager m.m. i.* (pan with high edged and a lid.) with the lemma *pande* as the genus proximum. However, the genus proximum in DDO is *not* a unique reference to a sense, so the original DDO excerpt seen in Figure 7 would require disambiguation in order to organize the hyponyms correctly, meaning to link pan hyponyms to pande 1_1 (frying pan) and not to pande 2_1 (forehead).

	Expanded le	Pos	GenProx	Definition
1	grillpande_1	sb.	pande	pande af smedejern hvis steg
2	pande,1_1	sb.	redskab	redskab til at stege med, mec
3	pande,2_1	sb.	del	den øverste del af hovedets fo
4	sauterpande_	sb.	pande	pande med høje kanter og evt
5	teflonpande_	sb.	pande	pande med teflonbelægning
6	æbleskivepa	sb.	pande	pande med halvkugleformede

Figure 7: Ambiguous genus proximus in DDO (*pande* – frying pan and forehead)

A further task when compiling the wordnet from DDO was to add supplementary information in terms of relations not specified for in the definitions since they were generally seen as obvious to the reader. Examples of such relations were the *used_for* relation in DanNet, which for the more general concepts were not accounted for in the dictionary (cf. Lorentzen and Nimb 2009 and Nimb 2009). As further stated in Pedersen et al. 2009:272:

> For example, nothing is generally said about the human user when DDO describes the use of instruments and buildings since it is

> obvious to the reader. Only when the user belongs to a very restricted group it is mentioned in the definition. (…)Interestingly enough, inheritance can facilitate the manual enrichment of semantic information. The inheritance mechanism ensures that relations are added systematically to all hyponyms (to be restricted or blocked if necessary). For example all hyponyms of *butik* (shop) inherit the involved agent *handlende* (shopkeeper). Thus, the DanNet editor is prompted to identify the involved agent of the more restricted hyponym: that the shopkeeper of a pharmacy is a pharmacist, the shopkeeper of a bakery is a baker and so on. Such information is only rarely specified in DDO definitions (although sometimes provided implicitly as examples of word formation).

When the compilation of the DT was initiated in 2011, it was decided to use DanNet as a starting point since the taxonomical structure had now been carefully ordered and supplemented with missing semantic information. In this respect it was better suited as a basis for a concept dictionary than the original DDO. When developing DT further, it became clear, however, that the additional thematic and syntagmatic information provided now in DT would be valuable also for the wordnet, cf. Nimb and Pedersen (2012), Nimb and Pedersen (subm.). An example of such supplementary information is seen in Figure 8 where events of crying are listed together with references to crying persons as well as the properties of crying person.

Figure 8: Excerpt from DT (in progress) illustrating information types on theme and arguments

Where DT is ordered in thematically based chapters and subchapters, the primary structuring principle in DanNet is the hyponymic backbone which is also resembled in the EuroWordNet ontology (Vossen et al. 1999). However, this construct does not necessarily account for thematic resemblances, as the ones seen in Figure 8. This problem is often referred to as *the tennis problem* in the wordnet community (cf. Sampson 2000) pertaining the fact that wordnets traditionally do not account for the relatedness of concepts such as *tennis*, *tennis player*, *ball*, *racquet* and *net* or, as exemplified in Figure 8, for the relatedness of *tudeprins* (crybaby) and *grådkvalt* (tearful). For LT applications that require some level of "deep" understanding such as information retrieval, question answering, text navigation and text mining this lack of information in wordnets remains problematic. So even if hyponymy may include some very basic aspects of the way we organize and conceive concepts, and even if this structuring principle is convenient for computers for its inheritance properties, it is far from sufficient to account for the central relatedness

Travel
pas_1_1 (passport)
rejse_1_1_1 (travel)
ankomme_1 (to arrive)
valuta_1_1 (currency)
rejselyst_1 (wanderlust)
billet_1 (ticket)

Figure 9: Travel concepts related via DT

between concepts. Figure 9 shows how concepts related to travel have been related in DanNet via thematic information from DT. An issue closely related to the tennis problem is the co-called ISA overload, i.e. the situation where sets of unequal hyponyms are grouped as sisters under the same hypernym. Thus, hyponyms subsumed under the same synset may share some very general dimension of functionality or form, but they belong to all sorts of domains and would, in a thesaurus, basically be categorized in a completely different way.

Finally, DanNet as well as most other wordnets lack information on arguments of valency bearing words. Typical subjects and objects of specific verbs are information types frequently asked for in applications. Therefore, we are experimenting with the integration of syntagmatic information from DT, which encodes a large range of selectional restrictions on verbs and deverbal nouns. Figure 10 shows a sample of verbs in DT which are encoded with the relation involved agent=person which we plan to integrate as an enrichment of DanNet.

Involved_agent relation
klæde_om_1 (to dress) **involved_agent** person_1 (person)
kræve_ind_1 (to make demands) **involved_agent** person_1 (person)
tyde_1 (to interpret) **involved_agent** person_1 (person)
powernappe_1 (to powernap) **involved_agent** person_1 (person)
sexliv_1 (sex life) **involved_agent** person_1 (person)

Figure 10: Persons as involved agents extracted from DT (Nimb et al. subm).

Likewise, properties constitute a specific syntagmatic case where information on the external argument is not prototypically given in wordnets. Figure 11 shows an excerpt from DT where the event *koge* (to cook) is supplemented with a set of properties related to the food that is being prepared in the event, such as 'al dente', 'tender', 'steamed' etc. Via information transfer to DanNet, this information is encoded as potential properties of foods. For a full account of these experiments and their validation, we refer to Nimb and Pedersen (2012) and Nimb et al. (subm).

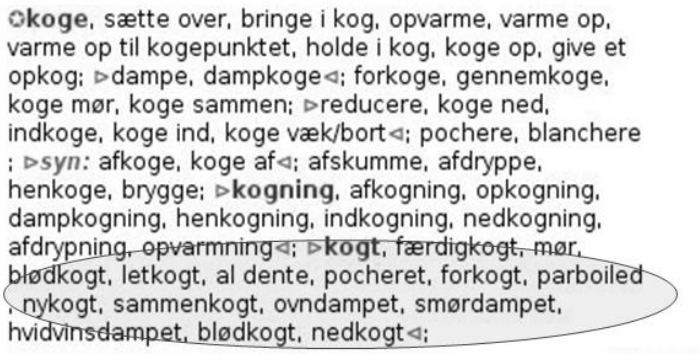


Figure 11: Properties of cooking in DT (in progress)

3.2. *Building appropriate sense inventories for word sense disambiguation – a challenge in LT*

An LT task that proves to constitute a fundamental problem in most applications, is word sense disambiguation. It seems that choosing the right meaning of a word in a specific context is one of the tasks that really challenges the lexicographical setup. Expressed in another way, when performing automatic word sense disambiguation we are thrown directly into an old, still ongoing lexicographical combat between so-called lumbers and splitters and between corpus linguists who claim that word senses are simply a lexicographer's construct in order to impose order to a words role in language (Kilgarriff 2007). Kilgarriff states further that (2007:29), "the trouble with word sense disambiguation is word senses. There is no decisive way of identifying where one sense of a word ends and the next begins". In so-called supervised word sense disambiguation – where the computer is trained to disambiguate on the basis of a gold standard of annotated data – the sense inventory becomes a very crucial issue. Our claim is that in order to establish a suitable sense inventory for sense annotation, we can gain from exploiting composite information from several parts of the lexical core.

Previous annotation projects support this claim since they show that wordnets are generally too fine-grained and unstructured to achieve good inter-annotator agreement, cf. Brown, Rood and Palmer (2010) who suggest a manual collapsing of senses (since inter-annotator agreement is only a little above 50% using Princeton WordNet as it is). For comparison, relatively good annotator agreement (close to 90%) is achieved in the Dutch semantic corpus (Dutch SemCor, Vossen et al. 2011) where a composite Dutch lexical resource is applied for annotation

(cf. the Cornetto Database above). Relating to the Danish lexical core, we are therefore experimenting with the semi-automatic clustering of the Danish sense inventory exploiting in this process the information types of *all its three resources*. To be more specific, we are combining i) information of the main and sub-senses in DDO with ii) the ontological typings of DanNet and iii) the thematic labeling from DT.

In this way we believe to be able to establish on automatic grounds a more suitable and sufficiently coarse-grained sense inventory specifically tuned for the task. Our initial approach is to cluster sub-senses of a word with the main sense that they belong to, unless a sub-sense has another ontological type or topic/theme than the main sense. This is typically seen with metaphorical senses, as in for instance *lys* (light) whose main sense is of the type PHYSICAL_PHENOMENON whereas the sub-sense is of the type MENTAL_PHENOMENON, as in *bringe lys og glæde* (bring light and joy). In this case both senses are kept since the metaphorical sense includes both another ontological type and another theme than the main sense. In contrast, several subtypes of *kort* (card) with the same ontological type and same theme will be collapsed, including the senses postcard and visiting card. A focus on disagreeing purposes (i.e. focusing on the telic role or the used_for relations as encoded in DanNet) *could* distinguish such senses in case the sense inventory is considered to be too coarse-grained. For verbs, which generally have many fine-grained sub-senses, a similar approach is adopted, including here, however, also the valency aspect of each particular sense. The intuitive belief is further that thematic information on surrounding words derived from DT will constitute a strong sense indicator (i.e. if other words in the context relates to food making, then *pande* is more likely to refer to a frying pan); a hypothesis that has previously been very hard to test at a large scale due to insufficient thematic information in the lexical resources (the afore mentioned "tennis problem").

4. Conclusions

In this paper we claim that the synergies between lexicography for humans and lexicography for machines have increased radically in recent years since many projects now rely on a common, corpus-derived lexicographical core from which different excerpt can be drawn. In this

respect, lexicographical cores which keep track of the basic sense inventory in terms of unique IDs prove to constitute a strong, composite resource which can be re-compiled for many different purposes, be they in the shape of a paper dictionary or realized as databases to serve as input to LT systems.

Lexical resources are applied in a large range of LT applications today spanning from morphologically based word guessing devices in mobile phones, over phonology in speech systems, syntax checking in word processing tools to semantic use in more experimental applications such as content-based information retrieval, question-answering and e-learning. We have focused in this paper on the use of the latter information type and have referred to some experiments made on the transfer of thematic and syntagmatic information in the Danish lexicographical core. In addition, we have given some ideas of how the lexical core as a whole can be exploited for a semantically founded establishment of a sense inventory which we believe is better suited for automatic word sense disambiguation than what is previously seen in the LT community.

References

A. Dictionaries

Hjorth, E. and K. Kristensen (eds.) 2005. *Den Danske Ordbog*. Copenhagen: Gyldendal & Det Danske Sprog- og Litteraturselskab. (DDO.) Online version 2012: http://ordnet.dk/ddo.

B. Other literature

Atkins, B. T. S. 2010. 'The DANTE Database: Its Contribution to English Lexical Research, and in Particular to Complementing the FrameNet Data.' In G.-M. de Schryver (ed.), *A Way with Words: Recent Advances in Lexical Theory and Analysis. A Festschrift for Patrick Hanks*. Kampala: Menha Publishers.

Bick, E. 2009. 'DeepDict — A Graphical Corpus-based Dictionary of Word Relations.' In *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009*. Northern European Association for Language Technology (NEALT).

Biemann, C., S. Bordag, U. Quasthoff and C. Wolff 2004. 'Language-Independent methods for compiling monolingual lexical data.' In *Computational Linguistics and Intelligent text processing*. Springer: Berlin, 217–228.

Borin, L., and M. Forsberg 2009. 'All in the Family: A Comparison of SALDO and WordNet.' In Proceedings of the NODALIDA 2009 workshop WordNets and other Lexical Semantic Resources — between Lexical Semantics, Lexicography, Terminology and Formal Ontologies. NEALT Proceedings Series, Vol. 7 (2009), 7–12.

Brown, S. W., T. Rood and M. Palmer (2010). 'Number or Nuance: Which factors restrict reliable word sense annotation?' In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner and D. Tapias (eds.), *Proceedings of the 7th International Conference on Language Resources and Evaluation*. Valletta: European Language Resources Association.

Ferrández, O., M. Ellsworth, R. Muñoz and C. Baker 2010. 'Aligning FrameNet and WordNet based on Semantic Neighborhoods.' In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner and D. Tapias (eds.), *Proceedings of the 7th International Conference on Language Resources and Evaluation*. Valletta: European Language Resources Association, 310–314.

Kilgarriff, A. 2007. 'Word Senses.' In E. Agirre and P. Edmonds (eds.), *Word Sense Disambiguation – Algorithms and Applications. Text, Speech and Language Technology*. Dordrecht: Springer Verlag, 29–47.

Kilgarriff, A. and M. Rundell 2002. 'Lexical profiling software and its lexicographic applications – a case study.' In A. Braasch and C. Povlsen (eds.), *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002, Copenhagen, Denmark, August 13–17, 2002.* Copenhagen: *CST*, 807–818.

Lorentzen, H. and S. Nimb 2010. 'Fra ordbog til wordnet. Hvordan udmøntes en ordbogsdefinition i en formaliseret wordnetbeskrivelse?' In H. Lönnroth and K. Nikula (eds.), *Nordiska studiar i lexikografi 10. Rapport från Konferencen om lexikografi i Norden, Tammerfors 3-5 juni 2009.* Tammerfors, 329–344.

Pedersen, B. S, S. Nimb, J. Asmussen, N. Sørensen, L. Trap-Jensen and H. Lorentzen 2009. 'DanNet: The challenge of compiling a WordNet for Danish by reusing a monolingual dictionary.' In *Language Resources and Evaluation, Computational Linguistics Series* 43.3, 269–299.

Pedersen, B. S., L. Borin, M. Forsberg, K. Lindén, H. Orav and E. Rögnvalssson 2012. 'Linking and Validating Nordic and Baltic

Wordnets – A Multilingual Action in META-NORD.' In *Proceedings of 6th International Global Wordnet Conference*. Matsue, Japan, 254–260.

Nimb, S. 2009. 'The Semantic Relations of Artifacts in DanNet.' In Proceedings of the NODALIDA 2009 workshop. WordNets and other Lexical Semantic Resources – between Lexical Semantics, Lexicography, Terminology and Formal Ontologies. NEALT Proceedings Series, Vol. 7.

Nimb, S. and B. S. Pedersen 2012. 'Towards a richer wordnet representation of properties – exploiting semantic and thematic information from thesauri.' In *LREC 2012 Proceedings*. Istanbul, Turkey, 3452–3456.

Nimb, S., B. S. Pedersen, A. Braasch, L. Trap-Jensen and N. Sørensen (submitted). 'Enriching wordnets from thesauri with thematic and syntagmatic relations.' Submitted to *LRE Journal Special Issue on Wordnets Relations*.

Rundell, M. 2011. Keynote talk at *eLEX2011*, Slovenia.

Sampson, G. 2000. 'Review of WordNet: An Electronic Lexical Database.' International Journal of Lexicography 13.1: 54–59.

Vossen, Piek (ed.) 1998. EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Dordrecht: Kluwer Academic Publishers.

Vossen P., I. Maks, R. Segers, H. VanderVliet and H. van Zutphen 2008. 'The Cornetto Database: the architecture and alignment issues.' In *Proceedings of the Fourth International GlobalWordNet Conference - GWC 2008*, Szeged, Hungary, January 22-25, 2008.

Vossen, P., A. Görög, F. Laan, M. van Gompel, R. Izquierdo and A. van den Bosch 2011. 'DutchSemCor: Building a semantically annotated corpus for Dutch.' In I. Kosem and K. Kosem (eds.) *Electronic lexicography in the 21st century: New applications for new users. Proceedings of eLex 2011. 10-12 November 2011. Bled, Slovenia*. Ljubljana: Trojina, 286–296.