NEALT Proceedings
Northern European Association for Language Technology

Proceedings of the Workshop on
Lexical Semantic Resources for NLP
NODALIDA 2013

May 22-24, 2013 • Oslo, Norway

Linköping Electronic Conference Proceedings

Proceedings of the workshop on

lexical semantic resources for NLP

at NODALIDA 2013

edited by

Lars Borin
Ruth Vatvedt Fjeld
Markus Forsberg
Sanni Nimb
Pierre Nugues
Bolette Sandford Pedersen

NEALT Proceedings Series 19 • ISBN 978-91-7519-586-5
Linköping Electronic Conference Proceedings 88
ISSN 1650-3740 (Online) • ISSN 1650-3686 (Print) 2013

Preface

High-quality lexical semantic resources with sufficiently large vocabularies still make up a serious bottleneck not only in purely rule-based NLP applications but also in supervised corpus-based approaches. The oldest widely-known lexical semantic resource, Princeton WordNet (PWN), has been around for over two decades. While PWN and the numerous wordnet projects for other languages that it has inspired adhere fairly closely to the traditional dictionary in their conception and organization, there are also lexical-semantic resources where a closer integration of lexical data information and corpus data is attempted. Such resources can be seen either as extremely richly exemplified lexicons or extremely deeply annotated corpora, depending on your outlook. Berkeley FrameNet, VerbNet, PropBank and several others can be mentioned in this connection. A recent trend in the wake of the increased awareness of the importance of standardization and interoperability of language resources, is the development towards large-scale integration of lexical resources (variously referred to as "lexical cores", "lexical macroresources", "lexical resource networks", and the like) both within and across languages, the ultimate expression of which is at the moment the linked open data in linguistics movement.

For largely extraneous reasons, English-language resources tend to receive most attention in the LT literature, but there is an increasing number of lexical semantic resources under development for many other languages, including Nordic, Baltic and other languages of the NEALT area.

In parallel to this development of new lexical semantic resources, much effort is put into exploring how such resources and formal ontologies can be made to work together in knowledge-based systems. The workshop – a follow-up on the succesful Nodalida 2009 workshop where the focus was on wordnets – was intended to bring together researchers involved in building and integrating lexical semantic resources for NLP as well as researchers that are more theoretically interested in investigating the interplay between lexical semantics, lexicography, terminology and formal ontologies.

We invited papers presenting original research relating to lexical semantic resources for NLP on topics such as:

- representation of lexical-semantic knowledge for computational use
- the interplay between formal ontologies and lexical resources
- corpus-based approaches to lexical semantic resources
- terminology and lexical semantics: concept-based vs lexical semantic approaches
- monolingual vs. multilingual approaches to lexical-semantic resources and ontologies
- word-space models for building and expanding ontologies
- domain-specific classification: taxonomy and ontology – computational aspects
- quality assessment of lexical-semantic resources: criteria, methods
- computational use of lexical-semantic resources (information retrieval, semantic tagging of corpora, MT, etc.)
- traditional lexicography and NLP lexicons: re-use and differences
- cognitive aspects: computational lexical models as opposed to the 'mental lexicon'

Out of the six submissions received, four were accepted for presentation at the workshop and inclusion in this proceedings volume after a thorough review procedure and subsequent revision by the authors of the papers, Each submission was reviewed by three (anonymous) members of the program committee:

- Lars Borin, University of Gothenburg, Sweden
- Ruth Vatvedt Fjeld, University of Oslo, Norway
- Markus Forsberg, University of Gothenburg, Sweden
- Karin Friberg Heppin, University of Gothenburg, Sweden
- Richard Johansson, University of Gothenburg, Sweden
- Rune Lain Knudsen, University of Oslo, Norway
- Dimitrios Kokkinakis, University of Gothenburg, Sweden
- André Lynum, University of Oslo, Norway
- Sanni Nimb, Association for Danish Language and Literature, Denmark
- Pierre Nugues, Lund University, Sweden
- Bolette Sandford Pedersen, University of Copenhagen, Denmark
- Joel Priestley, University of Oslo, Norway

The invited speaker at the workshop, Graeme Hirst (University of Toronto), presented some of his recent work on lexical semantic resources for NLP under the title *Ontologies versus lexical semantics*.

The workshop organizers:
Lars Borin
Ruth Vatvedt Fjeld
Markus Forsberg
Sanni Nimb
Pierre Nugues
Bolette Sandford Pedersen

WS website: `http://spraakbanken.gu.se/eng/nodalida-lexsem-ws-2013`

Acknowledgements: Financial support for the organization of the workshop has come in part from the Swedish Research Council (the project *Swedish FrameNet++*, contract no. 2010-6013), and in part from the University of Gothenburg, through its support of the *Centre for Language Technology*: `http://www.clt.gu.se`

Contents

Ontologies versus Lexical Semantics

Graeme Hirst

Department of Computer Science, University of Toronto, Toronto, Ontario, Canada M5S 3G4

gh@cs.toronto.edu

ABSTRACT

Ontologies and semantic lexicons enjoy a complex relationship. Although words denote concepts and concepts make up ontologies, a lexicon is at best an ersatz ontology, and an ontology is too impoverished to function as a semantic lexicon. There is no clear mapping from the word senses and sense relationships of a semantic lexicon to the concepts and concept relationships of an ontology.

The reasons for this include the following: Word senses overlap in complex ways; many concepts are not lexicalized in some or all languages; and languages make semantic distinctions that are not ontological, but which are nonetheless reflected in surface-form aspects such as classifier-words, diathesis alternations, and the count / mass distinction. Nonetheless, a lexicon can sometimes be the basis for the development of a practical ontology.

KEYWORDS: Ontologies, lexical-semantic resources, word senses, classifier-words, diathesis alternations.

Graeme Hirst 2013. Ontologies versus lexical semantics. *Proceedings of the workshop on lexical semantic resources for NLP at NODALIDA 2013.* NEALT Proceedings Series 19 / Linköping Electronic Conference Proceedings 88: 1.

Automatic identification of construction candidates for a Swedish constructicon

Linnéa Bäckström, Lars Borin, Markus Forsberg, Benjamin Lyngfelt, Julia Prentice, and Emma Sköldberg

Dept. of Swedish, University of Gothenburg, Sweden

{linnea.backstrom, lars.borin, markus.forsberg,
benjamin.lyngfelt, julia.prentice, emma.skoldberg}@svenska.gu.se

Abstract

We present an experiment designed for extracting construction candidates for a Swedish constructicon from text corpora. We have explored the use of hybrid n-grams with the practical goal to discover previously undescribed partially schematic constructions. The experiment was successful, in that quite a few new constructions were discovered. The precision is low, but as a push-button tool for construction discovery, it has proven a valuable tool for the work on a Swedish constructicon.

Keywords: hybrid n-gram, Swedish, constructions, constructicon.

Linnéa Bäckström, Lars Borin, Markus Forsberg, Benjamin Lyngfelt, Julia Prentice and Emma Sköldberg 2013. Automatic identification of construction candidates for a Swedish constructicon. *Proceedings of the workshop on lexical semantic resources for NLP at NODALIDA 2013.* NEALT Proceedings Series 19 / Linköping Electronic Conference Proceedings 88: 2–11.

1 Introduction

The research within the project *A Swedish constructicon* (see section 2) is targeted at Swedish constructions that are collected, analyzed, described, and published in a freely available resource.[1] Since no exhaustive construction description has ever existed for Swedish – or, to our knowledge, for any language – an important methodological question for the project is how to discover those constructions that have not been recognized as such before.

At least in the initial experiment presented here, where we have explored the use of hybrid n-grams (see section 3) as a tool for construction discovery, the search for construction candidates is restricted to partially schematic patterns, i.e. structures where at least one component is lexically fixed and at least one component is schematic, i.e., a morphosyntactic category. In this way, we target patterns with both lexical and grammatical properties, which neither purely lexical nor purely grammatical tools can capture.

2 Swedish constructicon

The Swedish constructicon (SweCxn; Lyngfelt et al., 2012) is a collection of variable multi-word units, based on principles of Construction Grammar and designed as an addition to the Swedish FrameNet (Borin et al., 2010).[2] It is still in its early stages, but the intention is for it to be developed into a large-scale, freely available resource for linguistics and language technology. At present, SweCxn consists of around 100 Swedish constructions, and is growing continually. A major concern of SweCxn is to account for linguistic patterns that are too specific to count as general rules of grammar and too general to be attributed to individual lexical units. Such constructions are peripheral both from a grammatical and a lexical perspective, and are therefore easily overlooked and neglected in grammars and lexical resources alike. Special attention is given to constructions deemed problematic for L2 acquisition. The project is a collaboration between grammarians, language technologists, lexicographers, phraseologists, semanticists, and L2 researchers.

One of the goals of SweCxn is to develop tools for automatic identification of constructions in authentic texts. This is a highly desirable research objective in itself, with potential uses in a number of NLP applications. In addition, the same methods provide the project with a heuristic tool. By automatically extracting various kinds of regularities in texts, we may discover patterns that might otherwise have been overlooked. This especially concerns seemingly insignificant constructions that do not stand out against the context the way spectacular idioms do. The resulting findings are treated as construction candidates, a subset of which may be considered actual constructions after manual evaluation.

3 Experiment setup

The general setting for our experiment is the resource infrastructure of *Språkbanken* (the Swedish Language Bank),[3] a modular set of resources and tools in the form of web services for accessing, browsing, editing and automatically annotating resources. The two facets of the infrastructure most relevant for the present purposes are the corpus infrastructure *Korp* (Borin et al., 2012b) and the lexicon infrastructure *Karp* (Borin et al., 2012a). Together, these provide a set of interoperable web services and downloadable resources which enable experiments like the one described here to be quickly set up and executed.

[1] The Swedish constructicon is accessible from here: <http://spraakbanken.gu.se/resurs/konstruktikon>

[2] See <http://spraakbanken.gu.se/swefn>.

[3] <http://spraakbanken.gu.se>

word	msd	lemma
Hur	HA	hur
är	VB. PRS. AKT	vara
det	PN. NEU. SIN. DEF. SUB+OBJ	den
då	AB	då
i	PP	i
Mellanöstern	PM. NOM	Mellanöstern
?	MAD	

Figure 1: SUC 2.0 annotations

The data source for the experiment is SUC 2.0 (Ejerhed and Källgren, 1997; Ejerhed et al., 1992), a balanced text corpus for Swedish consisting of 1.17M tokens that have been manually annotated with lemmas and MSDs (morphosyntactic description). A random example sentence from SUC 2.0 is given in Fig. 1: *Hur är det då i Mellanöstern?* 'What about the Middle East?'. The first part of the MSD is the part-of-speech, e.g., *VB* for *är* 'is'.

SUC was selected in order to avoid annotation errors confounding the experiment results, but the experiment can (and has been) run on any of the more than hundred corpora of Språkbanken that have been automatically annotated with the same information.

The experiment is based on the work on StringNet (Tsao and Wible, 2009; Wible and Tsao, 2010, 2011), where the notion of *hybrid n-gram* plays a central role. A hybrid n-gram is a generalization of an n-gram where not only the word forms are included in the process, but also the information from the annotation layers. If we limit ourselves to lemmas and part-of-speech, which is the case for this experiment, then the 2-gram *Hur är* 'How is' would generate four construction candidates: *hur vara* 'how be', *hur VB* 'how VB', *HA vara* 'HA be', and *HA VB*.

Since the aim is to capture partially schematic constructions, we discard all candidates that are fully schematic or fully lexical, i.e., consisting of only PoS tags (e.g., *HA VB*) or lemmas (e.g., *hur vara* 'how be'). Moreover, we remove all hybrid n-grams containing punctuation marks and/or words marked as foreign. They are not necessarily uninteresting, but since they did introduce a lot of noise in the candidate list, we decided to remove them. For SUC 2.0 with 2-, 3-, and 4-grams we ended up with 16M hybrid n-grams of which 8.8M were unique.

The next step is to rank all hybrid n-grams, which can be done with a wide range of association measures. We have followed StringNet in using point-wise mutual information (PMI). PMI has a known shortcoming in these kinds of experiments – it has a preference for the low-frequency items – which can be remedied by multiplying PMI with the absolute frequency. This does not solve another problem, however, which is boilerplate text, e.g., "For subscription enquiries e-mail:...". But with a small modification – instead of counting hybrid n-grams, we count UIF (unique instance frequency), which is the number of unique n-grams underlying the target hybrid n-gram – we can counteract that problem too. In sum, we end up with the following formula:

$$\text{PMI-UIF(H)} = UIF * log_2\left(\frac{P(H)}{\prod_{x \in H} P(x)}\right)$$

There is still one more problem that needs to be solved: since the bulk of the hybrid n-grams are subsets of other hybrid n-grams, we arrive at a ranking list with massive redundancy. This is solved, in the same spirit as StringNet's vertical/horizontal pruning (Tsao and Wible, 2009; Wible and

$vara_{VB}$ ute_{AB} och_{KN} VB	*är ute och letar (3)*	15	0.93	52.24
$vara_{VB}$ JJ $för_{PP}$ att_{IE}	*är viktiga för att (2)*	26	1.61	52.83
$stänga_{VB}$ av_{PL} NN	*stängt av motorn (1)*	11	0.68	52.25

Figure 2: Some example hybrid n-grams from SUC 2.0 ranked by PMI-UIF

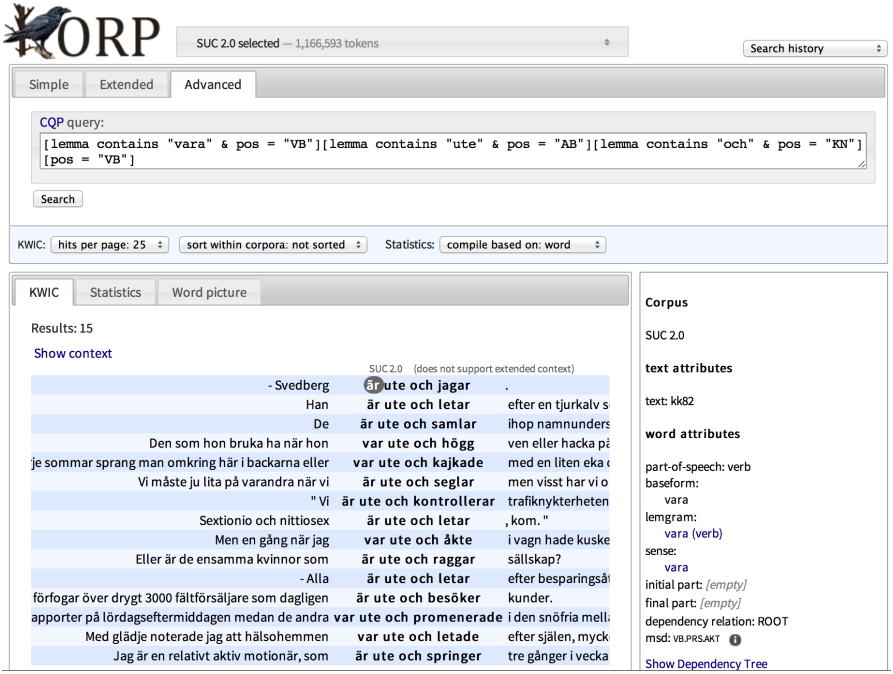


Figure 3: The instances of $vara_{VB}$ ute_{AB} och_{KN} VB

Tsao, 2010), by removing all hybrid n-grams that are subsets of other hybrid n-grams with a higher PMI-UIF. A hybrid n-gram is considered a subset of another if it occurs as a subsequence that are either equal or consisting of non-conflicting items sharing the same part-of-speech; e.g, $vara_{VB}$ is considered equal to VB.

Some sample candidates are given in Fig. 2. The hybrid n-grams are linked to the Korp interface to enable inspection of their instances (see Fig. 3). We also see the most frequent instance, followed by the absolute frequency, relative frequency, and the PMI-UIF. The full output of a top-2500 list is accessible from here: <http://spraakbanken.gu.se/eng/resource/konstruktikon/candidates> (may be subject to change). Here you will find other materials as well that have been annotated automatically using the Korp pipeline. More specifically for this experiment, we use the Swedish Hunpos tagger (Megyesi, 2009) for the part-of-speech tags, and the lexical analysis based on SALDO (Borin et al., 2008; Borin and Forsberg, 2009) for the lemmatization.

4 Data analysis

The construction candidate list makes it possible to go through a large amount of examples quickly, since every hybrid n-gram is directly linked to the instances in the corpus. However, it was a difficult task to draw the line between relevant and non-relevant constructions and this is still an ongoing matter of discussion in the project group. Of the 2500 items included in the list 50 constructions were decided to be relevant construction candidates according to our criteria, i.e., that they are partially schematic and productive multiword units that are "too general to be attributed to individual words but too specific to be considered general rules" (Lyngfelt et al., 2012).

The final list of 50 relevant constructions was extracted in several steps. First one project member went through the whole list extracting a list of 143 interesting candidates (approximately a day's work). This list was then, in consultation with the other members of the project group, gradually reduced and the final result of this process were, as mentioned above, 50 constructions that were found relevant for entries in the SweCxn. As the main goal was to discover constructions that are difficult to find with other methods the result of 50 is not the whole story – a construction candidate can also inspire descriptions of other similar constructions, which is a question of the researchers' capacity for creative thinking at a given moment in time.

The instances of the construction candidates display different properties regarding the form-function structure. The results represent patterns of lexical, idiomatic and syntactic character. A strong indication that the method identifies the correct items is that some of the qualifying constructions are already present in the SweCxn. One of these examples is:

(1) *RG NN* per_{PP} *NN*

The structure in (1) is realized in the corpus as, e.g., *en gång per dygn* 'once in 24 hours' and *500 kronor per månad* '500 Swedish Crowns per month'. This construction that can be regarded as a Swedish equivalent to a construction in the Berkeley English Cxn (Fillmore et al., 2012), the so-called Rate construction. Another construction already accounted for in the SweCxn is (2) below:

(2) den_{DT} *RO NN*

Instances of this structure found in the corpus are date expressions like *den 1 juli* 'the 1st of July' and *den tionde mars* 'the tenth of March'. Fillmore (2008) discusses this type of time expressions referring to dates or days of the week, which in English have a conventionalized structure with the preposition *on*, i.e., on_{PP} *NN RO* (*on June 17th*) and, hence, deviates from other time expressions like *in March*, *in the morning* and *at noon* (Fillmore, 2008). The Swedish date expression in (2) occurs without a preceding preposition, and differs in this respect from both its English counterpart and from the typical pattern of time expressions in Swedish as well as in many other languages, e.g., *i mars* 'in March', *på morgonen* 'in the morning', or *på eftermiddagen* 'in the afternoon'. This property makes the construction a challenge for language learners (Prentice, 2011).

A construction that is not previously included in the SweCxn is exemplified here:

(3) *RG* $år_{NN_{GEN}}$ $ålder_{NN}$

The genitive construction in (3) is realized in the corpus as, e.g., *vid sju års ålder* 'at the age of seven' and *från 17 års ålder* 'from 17 years of age'. The construction is not described in Swedish dictionaries in a sufficient way despite the fact that it can hardly be seen as completely transparent (cf. Köhler and Messelius, 2001). In, e.g., English and German the same content is expressed with a different kind of prepositional phrase (Eng. *at the age of...*; Germ. *im Alter von...*) (cf. Källström, 2012).

Other relevant candidates included in the list are the comparing constructions in (4)–(6) below:

(4) $varken_{KN}$ *NN* $eller_{KN}$ *NN*

(5) $vara_{VB}$ sig_{PN} *NN* $eller_{KN}$ *(NN)*

(6) $vara_{VB}$ sig_{PN} *PN VB* ($eller_{KN}$ $inte_{AB}$)

Examples of these structures from the corpus are (4) *varken uppehållstillstånd eller arbetstillstånd* 'neither residence permit nor work permit', (5) *vare sig fotboll eller ishockey* 'neither football or ice hockey', and (6) *vare sig vi vill eller inte* 'whether we want to or not'.[4] As we can see, (4) and (5) are used synonymously (in the sense of 'neither'), whereas (6) can be substituted by *oavsett (om)* 'regardless if'. The traditional normative rule is that *vare sig* in (5) requires explicit negation whereas negation is seen as part of the meaning of *varken*. The structure in (6) is obviously similar to (4) and (5) but here *vare sig* functions as a subjunction and does not require negation (Svenska språknämnden, 2005). In actual language use, however, both *vare sig* and *varken* are used with and without negation and, e.g., the usage notes in Svenska Akademien (2009) indicate that the candidates in (4)–(6) are subject of an ongoing language planning discussion. The potential for contamination – for both native and non-native speakers – is quite obvious, which is also reflected in the corpus. Considering the similarities and differences between the structures, as well as their discontinuity, the cluster can cause problems in relation to, e.g., language technology, lexicography, and language learning, which makes these structures excellent candidates for the SweCxn (cf. Lyngfelt et al., 2012).

Another interesting example occurring in the list is:

(7) $vara_{VB}$ ute_{AB} och_{KN} *VB*

The corpus samples linked to (7) mostly contains instances of the construction with the literal meaning 'being out doing something one typically does outside', e.g., *vara ute och jaga* 'being out hunting'. These instances are not particularly relevant as candidates for the SweCxn since they can be referred to as a general syntactical pattern. However, a search for this general structure in a wider range of corpora provides metaphorical instances of the pattern, implying a certain 'disorientation', 'confusion', or 'lack of knowledge' on the part of the agent. One of the most conventionalized examples is *vara ute och cykla* (lit. 'being out biking'), meaning 'being mistaken' or 'not knowing what one is talking about' (cf. *talk through one's hat*). Other realizations are *vara ute och segla* (lit. 'being out sailing'), and *vara ute och snurra* (lit. 'being out spinning'), which both are used synonymously with *vara ute och cykla* in a metaphorical sense. Here, the form-function structure is by no means obvious, which also makes the construction relevant from an L2-perspective. The word combination *vara ute och cykla* is included in printed dictionaries. However, the productivity of the construction is far from evident; an information that can be straightforwardly described in the SweCxn format.

As mentioned before, a majority of the items generated by the method are not relevant candidates for entries in the SweCxn, or at least of no priority in the current state of the project. One of those structures is exemplified in (8), where the candidate is a noun phrase followed by a finite verb, thus a general pattern that can be described according to syntactic rules:

(8) den_{DT} *JJ NN VB*

[4] It is doubtful whether the first component of *vare sig* should be synchronically analyzed as a form of *vara* 'be' as it has been in examples (5) and (6). Rather, the sequence *vare sig* should probably be treated as an unanalyzed whole.

Examples from the corpus are *de senaste månaderna har* 'the last months have' and *de nordiska länderna är* 'the Nordic countries are'. Another example is the sequence in (9) below:

(9) *SN PN VB* en_{DT}

Instances of this sequence are, e.g., *att det var en* 'that it was a' (in *Så fort han hörde att det var en kvinnoröst...* 'as soon as he heard that it was a women's voice...') and *Om du är en* 'if you are a' (in *Om du är en mördare...* 'If you are a killer...').

The sequence in (9) exemplifies another problem with the method, namely that it generates construction fragments – the sequence in (9) is not a recognized linguistic unit of any kind – due to the fact that the method is based on 2-, 3-, and 4-grams. In fact, the sequence in (9) is similar to the *lexical bundles* of Biber and Conrad (1999), a term that they use to refer to high-frequency (word) n-grams. However, distinct to the work reported here, the only criterion used for recognizing lexical bundles is their frequency. No collocation co-ocurrence measures or other means of ranking or filtering the results are used. Instead, fixed-length text word n-grams are sorted according to frequency and the resulting lists manually inspected for interesting results.[5] Lexical bundles are said to differ from other kinds of multi-word units in three major aspects: first, they are extremely frequent, second, they have no idiomatic meaning and last, they are not perceptually striking in themselves. Another characteristic of lexical bundles is that they often transcend structural boundaries.[6]

Biber and Barbieri (2007) ascribe lexical bundles a pre-fabricated or formulaic status, solely on the basis of their high frequency. However, this view has not escaped criticism. Nekrasova (2009) maintains that high-frequent sequences are of different strengths: A bundle should be described in terms of its place on a continuum from more holistic to more compositional units. In the NLP literature it has been observed that although frequency certainly is a strong indicator of MWE-hood (termhood, collocational strength), much can actually be done – and has been done – to improve on frequency alone (Wermter and Hahn, 2006; Pecina, 2010).

The algorithm also generates sequences like the one in (10):

(10) $till_{AB}$ och_{KN} med_{AB} *VB*

An example from the corpus is *till och med börja* 'even start' (in the context *skulle jag till med och börja dricka igen* 'would I even start drinking again'). As in example (9), the phrase has been cut off in an inadequate way. In addition, the structure in (10) contains the fixed phrase *till och med*, and can be compared with other examples from the list:

(11) i_{PP} all_{DT} $fall_{NN}$ *VB*

(12) $över_{PP}$ $huvud_{NN}$ $tagen_{PC}$ *VB*

The main parts of the items in (11)–(12) constitute lexically filled fixed phrases, *i alla fall* and *över huvud taget*, which makes them more suitable as candidates for a dictionary, and indeed, they are well covered in Swedish dictionaries of today.

In some cases the results represent a structure which at first sight does not seem very interesting or relevant according to our criteria. In some of those cases, however, the link to the corpus sample leads to interesting examples of subtypes of a general structure. (13) is one of those cases:

[5]This is a bit like attempting to discover the words in un-word segmented text by looking at the frequency of, e.g., four-character sequences, which seems to be an exercise of doubtful value.

[6]However, it is not very obvious what can be concluded about the language system or the mental lexicon of the language user from the attested high text frequency of a sequence like the example *in the case of the* cited by Biber and Conrad (1999).

(13) $komma_{VB}$ *IE VB*

The structure in (13) ('come IE VB') reflects the general valence relation between the fixed elements *komma att* 'come to' (which is used to form a periphrastic future, among other things) and the variable *VB*. These types of relations are generally well described in dictionaries and therefore not a main priority for SweCxn. However, a search in the corpus for the structure in example (13) highlights a more specific form-function structure:

(14) $komma_{VB_{PAST,SUP}}$ *IE VB*

The pattern in example (14) indicates that the action described by the verb was accidentally initiated, as in *Det var så jag kom att lösa det urgamla filosofiska problemet* 'that was how I came to solve the ancient philosophical problem'. The specific meaning associated with the verb forms is difficult to describe in a classical dictionary format, and the partially schematic structure make this construction – which is also quite productive – a relevant candidate for SweCxn. Looking at the construction in (14) from an L2-perspective, one can assume that it can cause problems, e.g., in relation to the more general structure in (13). How is the learner, who has never met the structure in (14), to know that *jag kom att lösa det urgamla filosofiska problemet* is not simply the past tense of *jag kommer att lösa det urgamla filosofiska problemet* 'I will solve the ancient philosophical problem'? And even once the learner has analyzed the difference between structures in (13) and (14), a certain potential for contamination on a semantic-pragmatic level can be expected (cf. Prentice and Sköldberg, 2011).

5 Conclusions and future work

From a methodological standpoint the experiment has been a success, in that we have been able to discover quite a few previously undescribed partially schematic constructions. The precision is low, but as a push-button tool for construction discovery, it has proven a valuable tool for the work on a Swedish constructicon.

The main issues with the construction candidates are that they often end up being too syntactic (e.g., a candidate may correspond to a regular NP pattern), too lexical (e.g., because of internal inflection of a multi-word unit), or fragmented (due to the nature of n-grams). Planned future work includes the exploration of whether a combination of existing Swedish lexical resources together with the syntactic analysis from the MALT parser (Nivre et al., 2007), accompanied by a more flexible notion of candidates than pure hybrid n-grams, can be used to counteract these issues.

An alternative approach we intend to explore is using MDL (Minimum Description Length) for the construction extraction, an approach that has been previously explored by Lagus et al. (2009).

Acknowledgments

The research presented here was supported by the Swedish Research Council (grant agreement 2010-6013), by the Bank of Sweden Tercentenary Foundation (grant agreement P12-0076:1), by the University of Gothenburg through its support of the Centre for Language Technology and of Språkbanken, and by Swedish Academy Fellowships for Benjamin Lyngfelt and Emma Sköldberg, sponsored by the Knut and Alice Wallenberg Foundation.

References

Biber, D. and Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, 26:263–286.

Biber, D. and Conrad, S. (1999). Lexical bundles in conversation and academic prose. In Hasselgard, H. and Oksefjell, S., editors, *Out of corpora: Studies in honor of Stig Johansson*, pages 77–85. Rodopi, Amsterdam.

Borin, L., Danélls, D., Forsberg, M., Kokkinakis, D., and Toporowska Gronostaj, M. (2010). The past meets the present in Swedish FrameNet++. In *14th EURALEX International Congress*, pages 269–281, Leeuwarden. EURALEX.

Borin, L. and Forsberg, M. (2009). All in the family: A comparison of SALDO and WordNet. In *Proceedings of the Nodalida 2009 Workshop on WordNets and other Lexical Semantic Resources – between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*, Odense. NEALT.

Borin, L., Forsberg, M., and Lönngren, L. (2008). The hunting of the BLARK – SALDO, a freely available lexical database for Swedish language technology. In Nivre, J., Dahllöf, M., and Megyesi, B., editors, *Resourceful language technology. Festschrift in honor of Anna Sågvall Hein*, number 7 in Acta Universitatis Upsaliensis: Studia Linguistica Upsaliensia, pages 21–32. Uppsala University, Department of Linguistics and Philology, Uppsala.

Borin, L., Forsberg, M., Olsson, L.-J., and Uppström, J. (2012a). The open lexical infrastructure of Språkbanken. In *Proceedings of LREC 2012*, pages 3598–3602, Istanbul. ELRA.

Borin, L., Forsberg, M., and Roxendal, J. (2012b). Korp – the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012*, pages 474–478, Istanbul. ELRA.

Ejerhed, E. and Källgren, G. (1997). Stockholm Umeå corpus 1.0. Produced by Department of Linguistics, Umeå University and Department of Linguistics, Stockholm University. ISBN 91-7191-348-3.

Ejerhed, E., Källgren, G., Wennstedt, O., and Åström, M. (1992). The linguistic annotation system of the Stockholm-Umeå corpus project - description and guidelines. Technical report, Department of Linguistics, Umeå University.

Fillmore, C., Lee-Goldman, R., and Rhomieux, R. (2012). The framenet constructicon. In Boas, H. and Sag, I., editors, *Sign-Based Construction Grammar*, pages 309–372. CSLI, Stanford.

Fillmore, C. J. (2008). Border conflicts: FrameNet meets Construction Grammar. In Bernal, E. and DeCesaris, J., editors, *Proceedings of the XIII EURALEX International Congress*, pages 49–68, Barcelona. Universitat Pompeu Fabra, Universitat Pompeu Fabra.

Köhler, P. O. and Messelius, U. (2001). *Natur och Kulturs svenska ordbok*. Bokförlaget Natur och Kultur, Stockholm.

Källström, R. (2012). *Svenska i kontrast. Tvärspråkliga perspektiv på svensk grammatik.* Studentlitteratur, Lund.

Lagus, K., Kohonen, O., and Virpioja, S. (2009). Towards unsupervised learning of constructions from text. In Sahlgren, M. and Knutsson, O., editors, *Proceedings of the Workshop on Extracting and Using Constructions in NLP of 17th Nordic Conference on Computational Linguistics, NODALIDA*. SICS Technical Report T2009:10.

Lyngfelt, B., Borin, L., Forsberg, M., Prentice, J., Rydstedt, R., Sköldberg, E., and Tingsell, S. (2012). Adding a construction to the swedish resource network of Språkbanken. In *Proceedings of KONVENS 2012 (LexSem 2012 workshop)*, pages 452–461, Vienna.

Megyesi, B. (2009). The open source tagger HunPoS for Swedish. In Jokinen, K. and Bick, E., editors, *Proceedings of the Nordic Conference on Computational Linguistics (Nodalida)*, volume 4 of *NEALT Proceedings Series*, pages 239–241, Odense, Denmark.

Nekrasova, T. M. (2009). English l1 and l2 speakers' knowledge of lexical bundles. *Language Learning*, 59(3):647–686.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryiğit, G., Kübler, S., Marinov, S., and Marsi, E. (2007). Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

Pecina, P. (2010). Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44:137–158.

Prentice, J. (2011). "jag är född på andra november" konventionaliserade tidsuttryck som konstruktioner – ur ett andraspråksperspektiv. Technical report, Institutionen för svenska språket, Göteborgs universitet.

Prentice, J. and Sköldberg, E. (2011). Figurative word combinations in texts written by adolescents in multilingual school environments. In Källström, R. and Lindberg, I., editors, *Young urban Swedish. Variation and change in multilingual settings*. University of Gothenburg.

Svenska Akademien (2009). *Svensk ordbok*. Norstedts, Stockholm.

Svenska språknämnden (2005). *Språkriktighetsboken*. Norstedts Akademiska Förlag, Stockholm.

Tsao, N.-L. and Wible, D. (2009). A method for unsupervised broad-coverage lexical error detection and correction. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 51–54, Boulder. ACL.

Wermter, J. and Hahn, U. (2006). You can't beat frequency (unless you use linguistic knowledge) – A qualitative evaluation of association measures for collocation and term extraction. In *Proceedings of COLING-ACL 2006*, pages 785–792, Sydney. ACL.

Wible, D. and Tsao, N.-L. (2010). StringNet as a computational resource for discovering and investigating linguistic constructions. In *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics*, pages 25–31, Los Angeles. ACL.

Wible, D. and Tsao, N.-L. (2011). The StringNet lexico-grammatical knowledgebase and its applications. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 128–130, Portland. ACL.

LBK2013: A Balanced, Annotated National Corpus for Norwegian Bokmål

Rune Lain Knudsen & Ruth E. Vatvedt Fjeld

Institute of Linguistic and Nordic Studies, University of Oslo

r.l.knudsen@iln.uio.no, r.e.v.fjeld@iln.uio.no

ABSTRACT

At the Department of Linguistics and Scandinavian Studies (ILN) and the University of Oslo, the task of assembling a balanced corpus representing modern Norwegian Bokmål has reached a significant milestone. The Corpus for Bokmål Lexicography (LBK) now consists of more than 100,000,000 words. These documents have been selected based on a statistical analysis of reading habits in the general population of Norway. The documents have been subject to both manual bibliographic annotation, as well as automatic morphological annotation for each document. LBK will play a central part of a set of interconnected lexical resources, the aim of which is to provide an extensive documentation of Norwegian Bokmål that covers lexical and other linguistic/lexico-syntactic aspects. This paper presents LBK2013, a subset of LBK that we consider to be an accurate and comprehensive representation of modern written Norwegian Bokmål. A description of the corpus, as well as a number of related projects are described.

KEYWORDS: NoDaLiDa 2013, Speech and Language Technologies, Northern Europe, Corpora, Lexicography, Lexical Semantics.

Rune Lain Knudsen and Ruth Vatvedt Fjeld 2013. LBK2013: A balanced, annotated national corpus for Norwegian Bokmål. *Proceedings of the workshop on lexical semantic resources for NLP at NODALIDA 2013.* NEALT Proceedings Series 19 / Linköping Electronic Conference Proceedings 88: 12–20.

1 Introduction

The LBK project was initiated in 1999 in an effort to create a corpus similar to KorpusDK, the Danish corpus that has served as the foundation for Den Danske Ordbog (DDO). In addition we wanted to use the corpus for statistical studies, which made proper balancing a key requirement. In order to account for modern Bokmål specifically, the applicable texts for inclusion in the corpus were restricted to a timespan ranging from the year 1985 and onwards. This year was chosen also to avoid digitizing earlier published texts as digitized texts became more common from around this time. This also gave a reasonable time span for modern Bokmål texts.

2 Related Work

Two balanced corpora for other languages have served as points of reference and inspiration for LBK. This section presents two corpora that have provided guidance for the decisions made for LBK: KorpusDK[1] and The British National Corpus[2].

KorpusDK is perhaps the most relevant reference point as it represents Danish, a language that has a close relationship and shares a common ancestry with Norwegian Bokmål. It has also been an integral part of projects similar to the ones currently being in development at ILN. KorpusDK consists of a total of 56 million words spanning the years 1983-2002. Each text is automatically annotated with morphological information, as well as bibliographic information and genre.

The British National Corpus (BNC) contains over 100 million words. 90% of the material is written language. The construction of BNC started in 1991 and the first official version was released in 1995. BNC has served as a main point of reference for the task of deciding upon an appropriate category distribution for written text, as it is similar to LBK2013 both in terms of size and goals. However, LBK2013 differs somewhat from BNC with regards to the chosen categories for a given text. The rationale for this is partly based on differences between the literary environments and language cultures of both countries. Due to limited finances, we also had to consider what was possible to get hold of from authors and publishers. In 1999 several writers were still afraid of their texts being misused/abused or in other ways losing some of their copyrights by allowing for inclusion into a research corpus. Over the last ten years, this fear has more or less vanished.

3 Architecture for LBK

LBK makes use of IMS Open Corpus Workbench (CWB), a widely used framework for managing and querying large text corpora (Evert and Hardie, 2011). It is available for researchers through Glossa (Nygaard et al., 2008), a web-based interface for corpora developed at the Text Laboratory, ILN. This interface enables a user to perform advanced searches within a corpus, and to specify subcorpora within which to restrict searches according to various kinds of metadata.

The documents in LBK are POS-tagged with the Oslo-Bergen tagger (Johannessen et al., 2012), and annotated with bibliographic and ethnographic annotation wherever applicable. Each text is assigned a mandatory text category: fiction, non-fiction, newspapers, subtitles or non-standardized, each of these categories divided into subcategories such as magazines, biographies, scholarly texts, etc. Each text is optionally given one or more topics such as sports,

[1] `http://ordnet.dk/korpusdk`
[2] `http://www.natcorp.ox.ac.uk`

law, medicine, ecology, music, etc. Additionally, each text is given information about the author and publishing information, wherever applicable.

LBK has received material from a large number of text suppliers over the years. One of the major bottlenecks in the workflow for LBK has thus been the handling of various document types that need to be converted to an appropriate format for the CWB framework. To enable an efficient and consistent workflow for managing the conversion and annotation of documents, a desktop application called *LBKTexts* was developed during autumn 2012. This application has simplified the workflow, increased the growth rate of LBK and secured the overall data consistency. LBKTexts is written in Java and will be made freely available under an open-source license during autumn 2013.

4 LBK2013

LBK2013 is a subset of the total amount of texts in LBK, selected with the aforementioned balance requirements in mind. The overall category distribution for LBK2013 is shown in Table 1.

Newspapers	10%
Non-fiction	45%
Fiction	35%
Subtitles	5%
Non-standardized	5%

Table 1: Overall distribution over the main categories of LBK2013

LBK was initially designed with a set of text categories, subcategories and topics, along with distributional guidelines based on observations done for similar corpora and the Norwegian Media Barometer from 2003, an annual statistical survey about the use of mass media in Norway. In 2013, this picture has changed quite drastically in some respects, especially related to internet usage. The distinction between standardized and non-standardized texts is not as easy to define anymore as much of what people read is published online without necessarily being subject to traditional proofreading. The categories and subcategories have thus been somewhat revised for LBK2013 in order to account for these changes, in particular limiting the non-standardized texts to the types of texts that without doubt can be said to be non-standardized (e.g. blogs, online forums, usenet discussions etc.).

5 Interconnectivity with a Dictionary

The need for a link between dictionaries and corpora for lexicographic work has emerged during the last 20 years, and can now be said to be the norm. A central phenomenon when analyzing corpora is the degree of word-sense ambiguity and the consequences arising from this phenomenon. Word-sense disambiguation (WSD) is thus an attractive element of a corpus analysis toolbox.

There are a number of methods one can employ when disambiguating word-senses in corpora, one popular approach being a distributional approach where the surrounding context of a word is subject to statistical analyses (cf. Kilgarriff and Tugwell (2002)). Establishing a strong link between a corpus and a dictionary should be valuable to such analyses. By using information from a dictionary as a machine-readable knowledge base, context analyses making use of e.g.

the collection of Lesk algorithms (Agirre and Edmonds, 2007) can be applied. This will also provide feedback on the consistency and completeness of the dictionary material.

By augmenting concordance results with definitions from a dictionary, users will have access to sense information for each word, including the written definition, usage examples, and so forth. In addition, users will be able to do corpus queries in which the desired subcorpora are restricted not only to morphological and bibliographic attributes, but semantic attributes as well. The dictionary will also have access to specific tokens and their context in the corpus, simplifying the task of empirical studies on dictionary lookups and related phenomena.

5.1 Establishing the Connection

A sense inventory with an index over all lemmas represented in the dictionary is made, mapping the lemma to one or more senses, each sense represented by a unique identifier either extracted from the database or generated automatically. An initial linking step is subsequently done by performing a lookup for each lemmatized word in the corpus, storing the set of possible sense id's as a special attribute for the word in question. This results in a relation from each token in the corpus to one or more senses in the dictionary whenever a match is found.

At the time of writing this paper, a test run of the linking step was done for a subset of LBK containing roughly 91.4 million words. The necessary dictionary information was extracted from Bokmålsordboka[3], a dictionary developed by ILN. A total of 59.8 million words were linked to matching dictionary entries. For these words, the average ambiguity (i.e. number of possible senses for the word) was 7.56, with a standard deviation of 1.82. There were a total of 5.93 million unambiguous words (i.e. pointing to one and only one possible sense for the word).

The ambiguous relations do provide useful lexicographic information to the user for words in the corpus matching a dictionary entry. In addition, the words that have no senses assigned indicates possible areas of study for future improvements of the dictionary. A more sophisticated approach is however needed in order to make the linked dictionary valuable for computational purposes.

5.1.1 Preliminary Disambiguation Experiments

To generalize the linking step, a sense-tagger meant to be used as an add-on for the Oslo-Bergen tagger is currently under development. Some preliminary experiments have been performed to test the functionality of the tagger. A set of sentences have been randomly extracted from LBK for the purpose of making a test set. These will be subject to manual sense-annotation using the sense inventory extracted from BOB as reference material. At the time of writing this paper, 68 sentences have been annotated manually by three annotators. A total of 907 words were assigned a total of 1254 senses. This might indicate a somewhat unneccesary degree of sense-granularity in some parts of the sense inventory. This suspicion is strenghtened by the low amount of fully agreed sense assignments done by the three annotators, as can be seen in Table 2.

The set of annotated sentences is too small to give any conclusive evidence one way or the other. However, some observations done during the annotation process was made. A large part of the cause of disagreement is due to the fine-grained sense categories. Many of the senses for a word

[3] `http://www.nob-ordbok.uio.no`

Senses	1254	100.00%
Full Agr.	552	44.01%
Major Agr.	268	21.37%
Minor Agr.	434	34.61%

Table 2: Summary of sense annotation results. **Full Agr.** is the number of senses assigned by all three annotators, **Major Agr.** is the number of senses assigned by two annotators, and **Minor Agr.** is the number of senses assigned by only one annotator.

tend to overlap, as has been observed in a number of other related projects and experiments ((Kilgarriff and Rosenzweig, 2000)). This especially holds for prepositions and auxilliary verbs. Whether or not a disambiguation at such a fine-grained level is actually neccessary has been subject to discussion in similar experiments (Palmer et al., 2006).

The sense-tagger initially assigns all possible senses for the words in the randomly selected sentences. It then walks through a series of disambiguation steps. Currently, only two disambiguation steps are actually in use: a Part-of-Speech (PoS) disambiguation, followed by an implementation of the Simplified Lesk (Kilarriff and Rosenzweig, 2000) algorithm. The PoS disambiguation removes all senses where there is a mismatch between the grammatical class that the definition belongs to, and the grammatical class assigned to the word by the Oslo-Bergen tagger. The Simplified Lesk algorithm ranks the possible remaining senses for each word by the overlap score of examples and glosses over sentence context. To increase the amount of possible overlapping words, all words subject to the overlap process are stemmed using the Snowball stemming framework[4], a set of programming libraries and tools containing improved versions of the Porter stemming algorithm for several languages. We considered lemmatizing the examples and definitions in order to do the overlap analysis on lemmas, but earlier experiments in tagging short, compressed sentences have shown that the lack of context makes automatic PoS inference unreliable. After ranking the sense candidates, all but the highest scoring candidates are discarded.

We intend to use the resulting scores from this tagger as a guideline for establishing a baseline for evaluating future versions of the sense-tagger. Currently, our test set of 68 sentences is not sufficient to make any concrete assertions. Some cautious remarks can nonetheless be made when looking at the present scores (see Table 3 for precision, recall and f-measures). Disambiguating by PoS did not improve the scores substantially, and it would be interesting to compare a larger experiment with similar experiments in other languages to see if something can be observed regarding homographs spanning more than one grammatical category. The Lesk algorithm increases the score, albeit a low one even for such a simple algorithm. The high degree of granularity for the sense inventory in use is one of the probable causes for this.

6 Interconnectivity With Other Resources

We are currently developing a framework for linking a selection of resources for lexicography and language technology. The framework will provide a communication layer that enables a given resource to query the other resources in order to supplement its own data, enabling synergetic properties to arise from the combined resources.

[4] `http://http://snowball.tartarus.org`

	None	PoS	Lesk
Precision	0.122	0.142	0.370
Recall	0.996	0.906	0.498
F-measure	0.218	0.245	0.424

Table 3: Precision, Recall and F-measures for the sense-tagger prototype under development, using no disambiguation (**None**), disambiguation by PoS (**PoS**), and further disambiguation by the Simplified Lesk algorithm (**Lesk**).

The dictionary used for the pilot project is already implicitly linked to Norsk Ordbank[5], as well as NorNet (Fjeld and Nygaard, 2009) by using common unique identifiers for senses. NorNet, a prototype for a Norwegian wordnet developed at ILN, was created by analyzing the dictionary entries in BOB, establishing a link between word-senses and synsets in NorNet and the dictionary entries in BOB. This means that the link between LBK2013 and BOB will also result in linkage between LBK2013, Norsk Ordbank and NorNet.

We wish to perform the same experiments using other lexical resources as the source of additional sense inventories. A full-scale wordnet for Norwegian Bokmål and Nynorsk is now available. This wordnet is developed by Kaldera Språkteknologi[6]. The Kaldera wordnet is based on a translation of the the Danish wordnet (Pedersen et al., 2009), and they both contain links to Princeton Wordnet (Fellbaum, 1998) via relations denoting synset-equivalence across separate wordnets. If the glossary for the Kaldera wordnet proves to be enough material for an adequate sense inventory, new possibilities for interconnectivity will arise, extending the lexical resources to cover several languages.

Glossa provides a common interface to several corpora. We plan to make use of the functionality of this interface to investigate potential candidates for interconnectivity between LBK and other corpora. The first corpus subject to such an investigation will be the Nordic Dialect Corpus. This corpus includes information such as phonetic annotation as well as transcribed audio. The aim is thus to connect tokens and sentences in LBK to spoken equivalents in the Nordic Dialect Corpus, as this information is absent in BOB.

7 Future Work

LBK2013 will be of value to a number of projects that require a balanced corpus of a substantial size. This section presents some of the remaining work to be done for LBK2013, as well as a selection of planned projects that will make use of LBK.

As shown in section 5.1.1, the relations between words in the corpus and dictionary entries need to be disambiguated further. Experiments on reducing this ambiguity automatically is in progress. We will experiment with other variants of the Lesk algorithm as well as more sophisticated algorithms. We will also attempt to make use of the metadata available in LBK2013 to see whether this can improve the disambiguation somewhat.

[5] `http://www.edd.uio.no/prosjekt/ordbanken`
[6] `http://www.kaldera.no/`

7.1 Development of a Database for Multiword Expressions

LBK will serve as the main dataset for a statistical analysis designed to assist the discovery of multi-word expressions (MWE's) and collocations. This type of analysis was done on the 40-million version of LBK in 2008 (Fjeld et al., 2010). The result of this analysis will serve as the foundation for both a lexical database designed to aid language research, and a new human-readable dictionary for MWE's and collocations. We also plan to compare the analysis done for the 40-million version with the 100-million version in order to document the benefits of enlarging a corpora. The analysis can be refined to investigate subcorpora, which will make it possible to document the phraseology of different subject fields, differences in phrases used by certain age groups, geographic locations, and more.

7.2 Lexical/Linguistic Resources

By using a large balanced corpus for lexicographic studies, hypotheses can be verified or falsified empirically. The use of corpora in lexicographic research has become the standard approach for modern lexicographic work, and it is one of the cornerstones for the continuous development of theoretical foundations for lexicography.

LBK is available as reference material for investigating existing lexical resources. It can be used as evidence for suggested revisions of existing dictionaries, such as removal of rare lemmas and/or inclusion of new lemmas based on frequency information extracted from LBK. Due to the bibliographic annotations, this can be further refined to specific category domains, for example the most frequent lemmas within the domain of law.

LBK2013 will also prove useful in the field of language standardization by providing frequency analyses of variant forms, either on a general level or for specific areas like age, sex, place of birth, even for specific sets of authors.

LBK2013 will constitute as an integral part of the BRO-project (Bokmålets og riksmålets ordbase), a collaborative effort initiated by ILN and Det Norske Akademi for Sprog og Litteratur, the aim of which is to provide an extensive lexical resource for Norwegian Bokmål based on the combined lexical resources of both parties. This will in part be based on the work done on enabling interconnectivity between a number of lexical databases at ILN. By linking data from Norwegian wordnets, the MWE database currently in development, dictionaries and LBK, we plan to create an extensive resource for Norwegian Bokmål, both machine and human readable.

LBK and the future MWE database will provide material for research and development of other lexical resources like wordnets, framenets etc. Extraction of domain knowledge should prove feasible due to the topic annotations (e.g. economy, law, medicine, computer science). We will conduct experiments on word-sense disambiguation using both statistical analyses and dictionary- and knowledge-based methods. Based on the interconnectivity framework previously described, we want to conduct experiments on relations between definitions in a dictionary and tokens in LBK, as well as information from wordnets.

Various resources useful for language technology will be made available. Frequency lists and n-gram statistics will be made available, both for the corpus as a whole and user specified subcorpora. Upon completion of the interconnectivity efforts, semantic annotations from the dictionary, wordnet and MWE will be available as source data for language technology and related research.

Acknowledgements

We would like to thank Johanne Wictorsen Kola for her meticulous work on collecting and processing texts for LBK in this important period leading up to LBK2013. We would also like to thank Kjersti Wictorsen Kola for her supervision and collaboration on the LBK material, her bibliographic knowledge, and for taking part in the sense-annotation work used in this paper. Finally we would like to extend our gratitude to the Text Laboratory at the University of Oslo for hosting LBK, providing important technical support over the years, and for developing Glossa.

References

Agirre, E. and Edmonds, P., editors (2007). *Word Sense Disambiguation - Algorithms and Applications*, chapter 5, pages 107–131. Springer.

Evert, S. and Hardie, A. (2011). Twenty-first century corpus workbench: Updating a query architecture for the new millenium. In *Proceedings of the Corpus Linguistics 2011 Conference*. University of Birmingham.

Fellbaum, C., editor (1998). *WordNet - An Electronic Lexical Database*. MIT Press.

Fjeld, R. V. and Nygaard, L. (2009). NorNet - a monolingual wordnet of modern norwegian. In *NODALIDA 2009 workshop: WordNets and other Lexical Semantic Resources - between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*, volume 7 of *NEALT Proceedings Series*, pages 13–16.

Fjeld, R. V., Nygaard, L., and Bick, E. (2010). Semi-automatic retrieval of phraseological units in a corpus of modern norwegian. In *Korpora, Web und Datenbanken. Computergestützte Methoden in der modernen Phraseologie und Lexicographie*, volume 25.

Johannessen, J. B., Hagen, K., Lynum, A., and Nøklestad, A. (2012). OBT+Stat: A combined rule-based and statistical tagger. In *Exploring Newpaper Language*, volume 49 of *Studies in Corpus Linguistics*, pages 51–65. John Benjamins.

Kilarriff, A. and Rosenzweig, J. (2000). English SENSEVAL: Report and results. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*.

Kilgarriff, A. and Rosenzweig, J. (2000). Framework and results for english SENSEVAL. In *Computers and the Humanities*, volume 34, pages 15–48. fd.

Kilgarriff, A. and Tugwell, D. (2002). Sketching words. In *Lexicography and Natural Language Processing*. Euralex.

Nygaard, L., Priestley, J., Nøklestad, A., and Johannessen, J. B. (2008). Glossa: a multilingual, multimodal, configurable user interface. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., and Tapias, D., editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008.

Palmer, M., Fellbaum, C., and Dang, H. T. (2006). Making fine-grained and coarse-grained sense distinctions, both manually and automatically. In *Natural Language Engineering*, volume 12.

Pedersen, B., Nimb, S., Asmussen, J., Sørensen, N., Trap-Jensen, L., and Lorentzen, H. (2009). Dannet: the challenge of compiling a wordnet for danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, 43(3):269–299.

Clustering word senses from semantic mirroring data

Hampus Lilliehöök, Magnus Merkel

Department of Computer and Information Science, Linköping University

hamli515@student.liu.se, magnus.merkel@liu.se

ABSTRACT

In this article we describe work on creating word clusters in two steps. First, a graph-based approach to semantic mirroring is used to create primary synonym clusters from a bilingual lexicon. Secondly, the data is represented by vectors in a large vector space and a resource of synonym clusters is then constructed by performing K-means centroid-based clustering on the vectors. We evaluate the results automatically against WordNet and evaluate a sample of word clusters manually. Prospects and applications of the approach are also discussed.

KEYWORDS: word senses, clustering, semantic mirroring.

Hampus Lilliehöök and Magnus Merkel 2013. Clustering word senses from semantic mirroring data. *Proceedings of the workshop on lexical semantic resources for NLP at NODALIDA 2013.* NEALT Proceedings Series 19 / Linköping Electronic Conference Proceedings 88: 21–35.

1 Introduction

The lack of access to lexical resources of high quality can be seen as a bottleneck for the development of language technology applications of many kinds. Many lexical resources, such as bilingual lexicons and thesauri, contain linguistic data that could be even more useful and interesting if the different sources of data could be combined in a controlled manner.

In this work we look at a way of combining the semantic mirroring method with a vector space-based word clustering approach. The idea is that while semantic mirroring is an excellent starting point for building lexical resources, given a bilingual lexicon or word-aligned parallel corpus, semantic mirroring data could be refined using more standard clustering techniques.

Semantic mirroring means using bilingual resources, such as lexicons or parallel corpora, for identifying semantic relations (e.g. synonymy and hypernymy) between words in a common language. Several experiments have been done, yielding clusters of words related in this manner. Those of importance to this article are Dyvik (2004) and Eldén et al. (2013).

Eldén et al. (2013) developed an application in MATLAB® for carrying out a graph-based semantic mirroring procedure.[1] A simplified description of how it is done is as follows: A *seed word* from language *a* is looked up in a dictionary, or a *translation matrix*, yielding a set of words in language *b*. Each of these are then translated back, generating a set of words in *a*, including the *seed word*. Each of these, except the seed word, are translated back and forth once again in the same manner, ignoring words that were not in the first translation set – with the number of *translation paths*[2] being counted – yielding a *weighted adjacency matrix* representing the number of paths between words in the translation matrix. Using measures of connectedness derived from the data generated in the prior step – and spectral graph theory – the adjacency matrix is decomposed into smaller word clusters. By assumption, these clusters represent different *word senses* – often, but not always related to that of the seed word.

The method has shown promising results, but the effects of combining the results from several seed words have not been examined in detail. There are several reasons why one wants to do this, perhaps most clearly illustrated by the technical limitations of applying the procedure only once. Firstly, the seed word is lost in some clusters from the GBSM application where it should be kept; by unifying similar clusters we aim to retain the seed word. Secondly, it is only when the procedure is repeated that a word can be associated with *several* senses.[3]

The aim of this work is to extend the GBSM application used by Eldén et al. (2013) to generate the word clusters for *all* available seed words, and to unify the results using a *vector space model*. In effect this means combining the word clusters generated from all seed words into a *unified word sense resource*, the result of which is evaluated in two steps:

1. Against WordNet (Miller, 1995): by measuring cosine similarity between output clusters and existing WordNet *synsets*.

[1] The work described in Eldén et al. (2013) is an extension of the implementation in Fagerlund et al. (2010). From here on the graph-based semantic mirroring application is referred to as *the GBSM application*.

[2] A translation path links together two words in some language via a shared translation word in another language.

[3] Which is desired every time two or more words are homonyms.

2. Manually: by examining a randomly selected subset of the output data.

2 Semantic mirroring and graph-based semantic mirroring (GBSM) application

The extraction of semantic data from bilingual resources is motivated linguistically by assumptions regarding the nature of translation. The term "mirroring" refers to the view of a translation as a semantic mirror of the source (Dyvik, 2004). The translation and source could be for example a bilingual dictionary, or an aligned parallel corpus. Semantically related words tend to have overlapping sets of translations in other languages. *Words in one language are nodes joining together words in another language through translation*. Given that the assumptions are valid, and that the resource is of adequate quality, translation paths between any two nodes (words) in a language indicate that there is some semantic relation between them; they may be synonyms, have similar meanings, or one word may be a subordinate to the other (hyponym). They may also share accidental translation paths, e.g. due to word strings having multiple meanings (homonyms and/or homographs).

Below is an example to give some idea of how the method can be applied. This example uses nouns, but is in other aspects similar to how it has been used in the GBSM application:

The Swedish seed word *rätt* (noun) is looked up in a dictionary, and the following translations are found:

{*course, dish, meal, justice, law, court}*.

Among the words in this set, we want to find translation paths via words in the Swedish language, so they are all translated back and forth. *course* has the Swedish translation set

{*lopp, kurs, lärokurs, flöde, fat, gång, stråt, väg}*

of which *kurs* and *fat* have translations in the initial translation set;
kurs translates to

{*tack,* **course***, class*},

and *fat* translates to

{*bowl, saucer, plate, barrel,* **course***,* **dish***, platter}*.

Thus, it turns out that *course* has a translation path to *dish*, and by assumption they have similar meanings, hence they qualify as a cluster.[4] Further on, it turns out that there is a whole structure of interconnectedness in the initial translation set. The task of the GBSM application is to identify this structure and mathematically determine where its links are weakest, and to decompose the interconnected structure into smaller clusters – each representing a semantic meaning. The translation matrix used by the GBSM application was generated from a subset of Norstedts Swedish – English lexicon, containing only adjectives (Norstedts, 2000).

[4]It also turns out that *dish* has a translation path to *course*. This is common, but not necessarily the case at all times.

3 Clustering

By clustering we mean the task of grouping sets of objects into new groups (or clusters) so that the objects in the same cluster are more similar to each other than to those in other clusters.

How clustering is done thus depends largely on the type of data, and for what end the data is being clustered. In most cases the data can be represented as points in some metric space, with data points (or items) having a similarity (or dissimilarity) measurable as a distance between each other in this space (Witten et al., 2011).

Two large categories of clustering are *hierarchical clustering* and *centroid-based clustering*. Hierarchical clustering can be either *divisive* (roughly: *top-down*) or *agglomerative* (roughly: *bottom-up*): Divisive clustering means that the algorithm starts with all data items belonging to a single cluster, with the goal of decomposing it into several clusters. Agglomerative clustering, starts with each data item as its own cluster, successively merging them together. A *linkage criterion* is used for determining which data items are involved in agglomeration or division (Witten et al., 2011). Three common criteria are single-linkage, complete-linkage and average-linkage. Single-linkage may cause clustering operations to be done when any pair of data items have a high similarity metric. With complete-linkage, all items in a cluster are taken into account; for example maximizing the sum of similarities for each data item in a cluster, against some candidate. With average-linkage, similarity is measured against some mean value of a cluster. When hierarchical clustering is done, a *dendrogram* is often generated, representing the points in time at which clusters were merged (agglomerated) or decomposed (divided). Selecting a point yields the cluster space at that time.

Centroid-based clustering[5] starts with an assumption about how many clusters the data is to be grouped into (this value is often called K). An initial clustering is done, either randomly or according to some heuristic. Next, the *centroid* items of each cluster are calculated. This can be done arbitrarily but a mean value is commonly used. The centroid items may or may not be actual members of the cluster. Following this, all members of all clusters are associated with their nearest, or most similar centroid, according to a selected similarity metric (for example Euclidean distance). Each data item is then moved to the cluster containing that centroid, and the process is repeated – including recalculation of the centroids – until no more items need to be moved. The cluster space present at this stage is the result of the algorithm.

Clustering algorithms in general suffer from some drawbacks. First of all, they are computationally hard – especially noticeable when dealing with large and/or high-dimensional datasets. Secondly, the numerous configurations that are possible makes them unlikely to produce satisfying results without some supervision, testing and tweaking.

There are also drawbacks related specifically to the different categories of clustering algorithms: For hierarchical clustering, it is hard to algorithmically decide at what point in time to stop the agglomeration or division. For centroid-based clustering, it is hard to know in advance what a good value for K is. Also K needs to be kept relatively small due to time complexity. Further on, the outcome of centroid-based clustering is very sensitive to the state of the initial clustering – the solution is guaranteed to be optimal with regard to this state – but there may exist better, but unreachable global solutions.

[5] One configuration of this algorithm is commonly known as *K-means* or *Lloyd's algorithm*.

4 Extending the GBSM application

The first practical step in this work has been to compile all the required input data. As mentioned earlier, the graph-based semantic mirroring application (GBSM) developed by Eldén et al. (2013) for their experiment has been used for this.

The program has been modified to generate all data at once, in an unsupervised manner; the program body has been included in a loop statement, which terminates after all words in the dictionary have been processed. The interactive features of the application have been replaced by statically or dynamically declared values where appropriate.

One main feature of the the program is a loop that divides larger groups (at this stage they are represented as graphs) into smaller ones, until a condition is met. A threshold *fiedler value* of 0.2 has been used for this. The fiedler value is a measurement for the connectedness of a graph; if the fiedler value is below the threshold the graph is too loosely connected and a cut must be made to separate the graph into two. Calculating the cut that yields two as connected subgraphs as possible is computationally difficult[6] and the fiedler value is a heuristic for this problem.[7]

In effect, sprawly, or loosely connected groups are decomposed into smaller, more tightly connected groups. Groups having a fiedler value equal to, or above the threshold level, are considered good enough, and the decomposition loop terminates.

Output data are all clusters of English adjectives, derived from graphs with three or more nodes[8] associated with the seed words that were used to retrieve them.

Python was used to implement the extended GBSM application, for its ease of use with text data and availability of relevant packages: NumPy, part of the SciPy library (Jones et al., 2001), iPython (Pérez and Granger, 2007), and WordNet (Miller, 1995), used here as part of the NLTK Python library (Bird et al., 2009).[9]

4.1 Dataset

As mentioned earlier, the dataset, or input data, are clusters of English adjectives, some of which can be seen in figure 1. Each section separated by ## represents the result of one iteration in the main loop of the modified GBSM application. The first two items of numerical data are only labels for the clusters and of no significance. What follows afterwards is the seed word followed by lines of data, or clusters, derived from the seed word, each including its fiedler value and word strings.

One important clarification on terminology: The input data consists of *word clusters* that are to be *clustered* together by comparing their vectors. The term *cluster* in this text can thus mean either a word cluster (a set of items, or a vector representing it), or a cluster of clusters (a set of sets, or a set of vectors). In the context of input data analysis or reduction of the dataset, clusters refer solely to word clusters that are present in the input data. In the context of K-means clustering, on the other hand, the *data items* to be clustered are in fact

[6] This is an instance of the *minimization problem* (Eldén et al., 2013).

[7] For details about how the fiedler value is calculated, see Eldén et al. (2013)

[8] Nodes are words in a common language, that share some translation path(s).

[9] WordNet is often used as a gold standard for evaluation of automated extraction of word sense data ((Cicurel et al., 2006), (Bansal et al., 2012))

```
13133:35:interchangeable
1:convertible,
0.66667:exchangeable,renewable,commutable,permutable,
##
13059:36:arctic
0.80357:northern,north,
0.31515:Hyperborean,boreal,northward,north-polar,
##
13007:37:cannonproof
1:shotproof,
1:bulletproof,
##
```

Figure 1: An excerpt of the dataset generated by the GBSM application, used as input for clustering.

$$\{w_1, w_2, \ldots, w_n\} \rightarrow \begin{array}{c} \{c_a = \{w_{a_1}, \ldots, w_{a_z}\}, \\ c_b = \{w_{b_1}, \ldots, w_{b_z}\}, \\ \ldots \\ c_j = \{w_{j_1}, \ldots, w_{j_z}\}, \\ c_k = \{w_{k_1}, \ldots, w_{k_z}\}, \\ \ldots \\ c_m = \{w_{m_1}, \ldots, w_{m_z}\}\} \end{array}$$

Figure 2: A set of possibly overlapping word clusters is yielded from a set of seed words.

clusters themselves, so clustering of data items means the grouping of word clusters into sets.

The final product of this application are *word clusters* that are yielded from merging similar *word cluster clusters*. This is an important property of this project; while the GBSM application groups words into sets, the Python application groups sets of words into sets of sets – which are then merged (see figures 2 and 3).

4.1.1 Reductions in the dataset

Reductions in the dataset have been made, in order to both decrease its size and increase its quality. The upper threshold, discarding input data clusters of larger size, has been set to 15. It is reasonable to argue that word clusters should be no larger; first of all, they are to be merged later on, and are thus to grow even larger. Also, it is assumed that most of the larger ones are large because of unfortunate properties of the GBSM application, rather than valid

$$\{c_a, c_b, \ldots, c_m\} \rightarrow \{\{c_a, \ldots, c_i\}, \{c_j, \ldots, c_k\}, \ldots, \{c_l, \ldots, c_m\}\}$$

$$\rightarrow \begin{array}{c} \{\{w_{a_1}, \ldots, w_{a_z}, \ldots, w_{i_1}, \ldots, w_{i_z}\}, \\ \{w_{j_1}, \ldots, w_{j_z}, \ldots, w_{k_1}, \ldots, w_{k_z}\}, \\ \ldots \\ \{w_{l_1}, \ldots, w_{l_z}, \ldots, w_{m_1}, \ldots, w_{m_z}\}\} \end{array}$$

Figure 3: Word clusters from the GBSM application are clustered into sets of word clusters. The unique word set of the clustered clusters are considered the result.

semantic properties. Clusters of sizes above 200 are not uncommon. By comparison, the largest synset of adjectives in WordNet is of size 23, and the average size is ∼1.65.

We have also chosen to filter duplicate clusters. Duplicate clusters in this sense are word clusters containing the same set of words – i.e. regardless of the seed word, fiedler value or order (they are automatically sorted upon creation). The set of words is the only attribute of a cluster that plays any role in the clustering algorithm (see section 4.3). It should be noted, however, that if duplicate word clusters are present in the clustering algorithm, they *do* affect it, but they will not add any qualitative features – only quantity to features already present.

4.2 Vector space model

Each unique word in the input data is assigned an index in the range $[0, n-1]$ of natural numbers, n being the number of unique words in the input data. Each word cluster is assigned in the same manner for the range $[0, m-1]$, m being the number of word clusters in the input data. For each of the clusters, a vector v of length n is created. $v_{i,j} = 1$ if word with index j is a member of the cluster with index i, $v_{i,j} = 0$ otherwise.

$$V_{m,n} = \begin{bmatrix} \begin{bmatrix} v_{0,0} & v_{0,1} & \cdots & v_{0,n-1} \end{bmatrix} \\ \begin{bmatrix} v_{1,0} & v_{1,1} & \cdots & v_{1,n-1} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} v_{m-1,0} & v_{m-1,1} & \cdots & v_{m-1,n-1} \end{bmatrix} \end{bmatrix}$$

Using the reduced input data: $n = 8832, m = 13691$. The graphic style of the matrix bears some resemblance to the data structures used for its representation. While the matrix is an array of dimensions $m * n$, each row vector is in its own an array of dimension $1 * n$, and as an element in the matrix array they can be accessed in constant time. For all partitioning and clustering purposes, lists of indices referring to these vectors are the data items, and various lookup functions are used for comparison.

Clustering in general is computationally hard in large and/or high-dimensional datasets. Depending on the configuration (see section 4.1) – some parameters determine which input data to be accepted or discarded – the number of data items, or word clusters in this application is in the range of 13691 – 21093. With an average cluster size in the range of 4.70 – 6.75, we could expect K to be in the order of around 2200 – 4500. Further on, the dimensionality (in this case equal to the number of unique words, which is the capacity each vector needs to hold) is in the order of around 8800 – 10000. Using K-means directly on this set quickly proved to be intractable.

Some observations were made, however. We found that some sets of word clusters would always be disconnected from the rest. Formally, this means that some sets of vectors can be found, with each vector being orthogonal to all other vectors in the vector space except at least one within its own set.

To give an example, consider the following matrix M as our vector space:

$$M = \begin{array}{c} \\ \mathbf{u} \\ \mathbf{v} \\ \mathbf{w} \\ \mathbf{x} \\ \mathbf{y} \end{array} \begin{array}{c} \begin{array}{ccccc} a & b & c & d & e \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

Analogous to the vector space (section 4.2), indices a, b, c, d and e represent words that may or may not be members of clusters **u**, **v**, **w**, **x** and **y**. The value 1 at $M_{u,a}$ means that word a is a member of vector u. In this case, vectors **u** and **v** are both orthogonal to vectors **w**, **x** and **y**, but not to each other. Vectors **w** and **y** are orthogonal to each other, but neither of them is orthogonal to vector **x**.

Any way of disjoining {**w**, **x**, **y**} into two or more sets would entail non-orthogonality between some pair of these sets, hence there is no way of separating them further.

As a result, M can be reduced into the two matrices M_1' and M_2':

$$M_1' = \begin{array}{c} \\ \mathbf{u} \\ \mathbf{v} \end{array} \begin{array}{c} \begin{array}{ccccc} a & b & c & d & e \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \end{pmatrix} \end{array} \quad M_2' = \begin{array}{c} \\ \mathbf{w} \\ \mathbf{x} \\ \mathbf{y} \end{array} \begin{array}{c} \begin{array}{ccccc} a & b & c & d & e \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

Or even simpler:

$$M_1'' = \begin{array}{c} \\ \mathbf{u} \\ \mathbf{v} \end{array} \begin{array}{c} \begin{array}{cc} a & b \end{array} \\ \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \end{array} \quad M_2'' = \begin{array}{c} \\ \mathbf{w} \\ \mathbf{x} \\ \mathbf{y} \end{array} \begin{array}{c} \begin{array}{ccc} c & d & e \end{array} \\ \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \end{array}$$

The hope was to reduce the vector space into a few roughly equally large subspaces by this principle, each of which would then be input to its own instance of the clustering algorithm, thus reducing the factors of time complexity. [10] Unfortunately, this was only a moderate success. The partitioning of the vector space by this criterion instead yielded one very large subspace of size 13085 and dimensionality 8127 – out of the complete vector space of size 13691 and dimensionality 8832 – plus many smaller ones.

On the positive side, it had a very useful side effect with regard to qualitative, or semantic aspects. Many of the disjoint spaces would in fact be found by the clustering algorithm, although it is not guaranteed. Additionally, since cosine similarity is used as the metric for cluster- or vector comparison, the partitioning of the vector space as described above yields the most dissimilar sets of clusters there are. This operation thus proves to be not only reducing the computational time of the clustering algorithm, but also guarantees improving the result.

The large connected subspace has the size of 13085 out of 15328 vectors (85.34%) when duplicates are allowed, and 13085 out of 13691 vectors (95.57%) when duplicates are forbidden (as in the case described above). Hence, all duplicate clusters appear to occur outside of this connected subspace.

[10] Explained in section 4.3.

4.3 The K-means algorithm

What remains before evaluation is to perform clustering on the data that can be reduced no further using disjunctive partitioning. The K-means algorithm has been chosen for this.

With hierarchical clustering there is an uncertainty of when to stop the division or agglomeration, while with centroid-based clustering (K-means being of this family), the value K, or the expected number of clusters, is pre-determined, and results from different configurations may need to be compared.

Initially, the value K is set, and two arrays of size K are created:

1. Array C for the centroid vectors, which is initially empty
2. Array D containing the partitions of vectors obtained from making $K-1$ cuts in the data. This is the initial clustering.

Next, the centroid vectors for each of the initial clusters in D are calculated. For each data item, or vector, in each cluster of D, its nearest centroid vector in C is calculated by measuring each centroid against the item using cosine, and the item is moved to the cluster having that centroid vector. If no nearer centroid vector is found, then the item is not moved. Next, the centroids are re-calculated and the procedure is repeated until no nearer centroid vector is available for any item. The state of D at this point is the result of the algorithm.

4.3.1 Weaknesses

The complexity of the K-means algorithm increases as a function of the following three factors: 1) the size N of D; 2) the value K, and 3) the number of iterations I needed to reach a solution.

The computational time complexity of the algorithm is in $\mathcal{O}(N \times K \times I)$. Further on, the computation time for each item-centroid comparison depends on the dimensionality of the vectors, what metric is used and how it is implemented.

4.3.2 Implementation in a word sense context

The initial idea was to use K-means directly on the vector space. The reductions in the input data were done partially for reducing complexity-causing factors, and partially for qualitative reasons. The partitioning of the cluster space did provide further reductions in data size, although not as much as was needed. K-means thus served two purposes:

1. To reduce the large non-disconnected subspace into more manageable subspaces.
2. To perform clustering on each subspace.

In practice this meant that the large interconnected subspace of size 13085 was used as input to the K-means algorithm, with $K = 15$, generating 15 clusters of sizes in the order of around 600 – 1100.[11] Next, K-means was applied on each subspace, with K derived from the number of unique words in that subspace.[12] Figure 4 gives an illustration of this.

[11] $\frac{N}{K} = \frac{13085}{15} \simeq 872,3$

[12] $K_{subspace} = \frac{number\ of\ unique\ words\ in\ subspace}{average\ cluster\ size\ in\ vector\ space}$

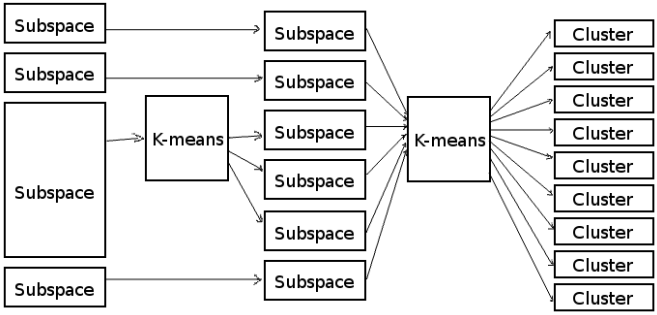


Figure 4: The K-means algorithm is applied in two steps.

4.4 WordNet as gold standard

Using WordNet as a gold standard is not ideal, but it is an extensive resource, performing well at approximating results, and it provides a foothold where the set of options otherwise would be small.

Cicurel et al. (2006) and Bansal et al. (2012) both evaluate word sense clustering results against WordNet. How this is done – i.e. how corresponding WordNet synsets are chosen, what metric is used, and whether WordNet's hierarchical relations are taken into account – is subject to variety. We have chosen to evaluate our data against WordNet as well, using a method similar to one presented by Cicurel et al. (2006) in their article; *one-to-one association*. As the name suggests, one WordNet synset is associated and compared to one word cluster.

Our evaluation procedure is as follows: Fifty clusters were selected at random on five intervals of cluster sizes, aiming at an even distribution.

For each of the clusters that contain at least one word that is present also in WordNet, we fetch every WordNet synset of adjectives having at least one word in common with that cluster. The similarities between the word cluster and the synsets are then determined using cosine. We do this in three ways:

1. With words not in WordNet removed from the cluster, plus words not in our resource's word space removed from each WordNet comparison synset
2. With words not in WordNet removed from the cluster
3. Without removing anything

By doing (1) and (2), we simulate a so-called projection of our clustering onto WordNet's set of words. [13] In effect, the clusters and synsets that are compared in (1) will always be subsets of the same word space, while in (2) and (3), this property is gradually relaxed, allowing for more dissimilarities between the word spaces of the cluster and the synset.

[13] This is partly inspired by (Bansal et al., 2012), who *project* WordNet's synsets onto their word resource to create reference clusters to evaluate against. Intuitively, one can think of the projection as the answer to "How would WordNet cluster this set of words?"

What one can expect from this is that the similarity score will be best in (1), followed by a gradual decline in (2) and (3).

Regardless of method, the word cluster and the synset are both represented as vectors with dimension equal to the number of words in their union. Differences in case are ignored, and the frequent separation characters "-" and "_" are replaced by spaces, as to normalize inconsistencies between the two resources. For each of the three methods, the synset having the highest similarity with the word cluster is chosen as the associated synset for that cluster.[14]

$$similarity(\{horse, apple, banana\}, \{apple, banana, aircraft\}) =$$

$$cos(\theta) = \frac{(1,1,1,0) \cdot (0,1,1,1)}{\|(1,1,1,0)\|\|(0,1,1,1)\|} = \frac{2}{3}$$

Figure 5: The two sets are compared using cosine, in a vector space having the dimension of their union.

Finally, the average similarity scores for the associated synsets are calculated. For all three sets of similarity scores described above, average similarity is calculated not only for the sets respectively, but also for the intervals of cluster sizes. Large clusters are suspected to score less, and this helps illustrate the suggested correlation. The average similarity scores for the entire output dataset are also calculated and presented.

4.5 Manual evaluation

As a second step in the evaluation, the authors evaluated the set of fifty word clusters manually. The clusters are the same as in the previous step. In this step, "linguistic intuition" was complemented by the use of dictionaries for verification.[15] The clusters were investigated independently by each evaluator, the predominant senses were identified, and then a precision score was calculated, based on the proportion of suggested words sharing the same meaning in relation to the number of words in the cluster. The interpretations and scores were then compared and discussed, and a consensus conclusion was reached.

We are however aware of our limitations as non-native English speakers, and would have preferred assistance of independent and experienced lexicographers, for a more reliable account. However, we believe the results fairly reflect the overall performance. A recurring difficulty in this area is to determine what synonyms really are. Jurafsky and Martin (2009) provide a method of testing for multiple word senses: "We might consider two senses discrete if they have independent truth conditions, different syntactic behavior, and independent sense relations, or if they exhibit antagonistic meanings." By contrast, this means that *synonyms are lexemes that denote the same word sense*, and by this reasoning, substituting a word for a synonym should preserve its propositional truth value and syntactic function in a sentence. WordNet (Miller, 1995) states: "Synonyms – words that denote the same concept and are interchangeable in many contexts." Any two synonyms will, however, carry different connotations. These usually affect the interpretation of the words, and thus in what contexts one might expect to find them.

[14] A similarity $\in (0, 1]$, $\in \mathbb{R}$ is obtained, where 1 means that the compared sets of words are identical.

[15] http://www.merriam-webster.com/ and http://www.oed.com

5 Results

Out of the 13691 clusters from the GBSM application, 2727 word clusters, or senses, were distinguished by K-means. Figure 6 shows how these are distributed by the criterion *number of unique words*. 1824 words, out of our total of 8832 unique words, are not in WordNet.

Cluster size	Clusters in range
2-4	467
5-7	629
8-11	681
12-17	592
18-70	358

Figure 6: Distribution of clusters by their number of unique words.

5.1 WordNet as gold standard

Below the results from the first evaluation procedure are shown. Average similarity$_{1,2,3}$ are the three different measurements described in section 4.4. Since we suspected that clusters of different sizes would not score equally well, scores for the size ranges described earlier have been measured, in addition to the overall score.

Cluster size	Avg. sim.$_1$	Avg. sim.$_2$	Avg. sim.$_3$
2-4	0.765	0.703	0.576
5-7	0.505	0.505	0.471
8-11	0.485	0.464	0.438
12-17	0.468	0.434	0.392
18-70	0.374	0.352	0.329

Figure 7: Average similarity with nearest WordNet synsets.

Cluster size	Avg. sim.$_1$	Avg. sim.$_2$	Avg. sim.$_3$
all (50 clusters)	0.515	0.487	0.438
all (2727 clusters)	0.508	0.481	0.432

Figure 8: Average similarity with nearest WordNet synsets.

As figure 7 suggests, small word clusters tend to have a relatively high similarity with WordNet, while large clusters score lower.

Figure 8 shows that the average similarity values are slightly higher for the selected fifty clusters than for the whole set. Some of this deviation is probably caused by the loss of precision that occurs when the size ranges are selected, since they can never be uniformly distributed.[16]

5.2 Manual evaluation

In the second evaluation step, we judge the clusters intrinsically[17] and with our human notions of sense. We have used the same size ranges here. Overall, the scores are higher in this evaluation step, as figure 9 illustrates.

[16] In the ideal case, the size ranges would constitute equally large portions of the data.

[17] That is, each cluster is evaluated by its own quality, rather than having external factors affecting the score.

Cluster size	Average score
2-4	0.975
5-7	0.742
8-11	0.764
12-17	0.807
18-70	0.725
all	0.802

Figure 9: Average scores for manually evaluated clusters.

There is a tendency towards better semantic consistency in the smaller clusters. Still, many of the larger clusters hold together surprisingly well.

5.3 Clusters compared to WordNet

The following examples are all from the selection of fifty clusters, unless noted otherwise. Average similarity$_{1,2,3}$ are to the right of each synset.

Cluster: {cufic, oddball}
WordNet: -
None of the words exist as adjectives in WordNet, hence this cluster was not included in the evaluation set.

Cluster: {misbelieving, miscreant, heterodox, heretical}
WordNet: {dissident, heretical, heterodox} 0.816 0.816 0.577
This case illustrates how easily the score departs from the precious 1.0.

Cluster: {frigorific, refrigeratory}
WordNet: {frigorific} 1.0 1.0 0.707
This example illustrates the effects of having several measurements; *refrigatory* is not in WordNet, so the cluster and the synset become identical once it is removed from the cluster.

Cluster: {suffocating, stifling, smothery, sweltering}
WordNet: {sweltering, sweltry} 0.577 0.408 0.354
WordNet: {smothering, suffocating, suffocative} 0.577 0.333 0.289
WordNet is more specific in its definitions here, while the generated cluster covers both senses. However, these words could still be interchangeable in many contexts. The cluster can be considered a supersense of the two WN synsets.

Cluster: {drunken, boozy, crapulous, potatory, Bacchic, full, drunk, inebriated, inebriate}
WordNet: {intoxicated, drunk, inebriated} 0.436 0.436 0.385

Cluster: {inevitable, everyday, mundane, familiar, accustomed, wonted, middling, average, run-of-the-mill, ordinary, white-bread, habitual, routine, second-rate, indifferent, moderate, standard, customary, vanilla, mediocre, straight, regular, bog-standard}
WordNet: {accustomed, customary, habitual, wonted} 0.426 0.426 0.417
This cluster may be a bit overgrown, still there are some words that are rated as synonyms in the manual evaluation.

6 Discussion

The results point in the direction of the intended result. We have been able to generate good semantic clusters by unifying clusters from the graph-based semantic mirroring application, in accordance with our hypothesis. There is, however, still room for improvement in several aspects.

Word clusters of smaller sizes are usually better than larger ones. Large word clusters tend to spring from large vectors, rather than many smaller ones. Some vectors carry multiple senses from the input data, and can therefore never provide qualities for improvement – they can even be destructive. Some clusters with N multiple senses are successfully split into N meaningful clusters when K-means is applied on it again with $K = N$. For other clusters, however, this instead creates semantically overlapping clusters. The average word cluster size among clusters generated using GBSM is much higher than that of WordNet's synsets.[18] Manual evaluation turned out to be harder than what was first expected. One can often identify small features, or nuances, of words that disqualify synonymy. Due to the nature of translation, this method is to some extent insensitive to varying levels of specificity. Very specific words may share clusters with less specific ones.

It would be interesting to further investigate the possibility to deterministically apply K-means a third time on *some* clusters, since manually doing so has proven useful in many cases, but destructive in others. Improvements to the input data from the GBSM application would be another way to enhance results, by for example having a more dynamic setting of when to split a graph compared to the current implementation. Furthermore, deriving the value K from factors other than the number of unique words when clustering a word cluster space may improve the result, whether or not one uses selective re-clustering.

7 Conclusion

We have found a method of clustering lexical synsets into more useful word clusters. It should be pointed out that the input data needs not originate in a GBSM application at all. One could for example use the Python program presented here to unify thesauruses or other semantic resources created elsewhere. Looking forward, the area of application determines which modificational steps should be taken. The biggest changes are probably attained by modifying the input data, which in this case means tuning parameters in the GBSM application, and/or using additional resources. Tweaking the GBSM application a bit should make way for more concise input vectors, giving a final result that scores higher by all metrics.

Acknowledgements

This work has been funded by the Swedish Research Council (Vetenskapsrådet).

[18] These are not directly comparable, however; WordNet contains many synsets of size 1, while here, such small clusters are discarded before the vector space is initialized. Moreover, WordNet is relatively fine-grained in its definitions.

References

Bansal, M., DeNero, J., and Lin, D. (2012). Unsupervised translation sense clustering.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.

Cicurel, L., Bloehdorn, S., and Cimiano, P. (2006). Clustering of polysemic words. In *GfKl'06*, pages 595–602.

Dyvik, H. (2004). Translations as semantic mirrors: From parallel corpus to wordnet. *Language and Computers*, 49(1):311–326.

Eldén, L., Merkel, M., Ahrenberg, L., and Fagerlund, M. (2013). Computing semantic clusters by semantic mirroring and spectral graph partitioning. Manuscript, submitted for publication.

Fagerlund, M., Merkel, M., Eldén, L., and Ahrenberg, L. (2010). Computing word senses by semantic mirroring and spectral graph partitioning. In *Proceedings of TextGraphs-5 - 2010 Workshop on Graph-based Methods for Natural Language Processing*, pages 103–107.

Jones, E., Oliphant, T., Peterson, P., et al. (2001). SciPy: Open source scientific tools for Python.

Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing*. Pearson/Prentice Hall.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41.

Norstedts (2000). *Norstedts stora engelsk-svenska ordbok*. Norstedts.

Pérez, F. and Granger, B. E. (2007). IPython: a System for Interactive Scientific Computing. *Comput. Sci. Eng.*, 9(3):21–29.

Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques (Third Edition)*. Morgan Kaufmann.

Enriching a wordnet from a thesaurus

Sanni Nimb[1], Bolette S. Pedersen[2], Anna Braasch[2], Nicolai H. Sørensen[1], Thomas Troelsgård[1]

(1) Society for Danish Language and Literature, Denmark
(2) University of Copenhagen, Denmark

sn@dsl.dk, bspedersen@hum.ku.dk, braasch@hum.ku.dk, nhs@dsl.dk, tt@dsl.dk

ABSTRACT

Wordnets are traditionally built around synonym sets with the vertical hyponymy relations as the central structuring principle. The hyponymy relation, however, does not necessarily group concepts into synsets that are particularly close from a thematic or functional point of view, a phenomenon which is sometimes referred to as the "ISA overload", or if contemplated from a thematic view point: the "tennis problem". In this paper we present two experiments. The first one concerns a method for remedying these problems by transferring thematic information from a thesaurus to a wordnet (Danish Thesaurus to DanNet). Hereby we can automatically subdivide co-hyponyms thematically as well as relate synsets thematically across parts of speech. Since the thesaurus is not yet fully completed, the paper describes work in progress; nevertheless, with an error rate below 5% of the most coarse-grained transferred themes, the experiment appears to be very promising. Finally, the second experiment concerns extension of DanNet via the Danish Thesaurus: The thematic organisation of the thesaurus in near synonyms is further applied as a very precise method for automatically extending the lexical coverage of DanNet.

KEYWORDS: Wordnet, " tennis problem", ISA overload, thesaurus, thematic information.

Sanni Nimb, Bolette S. Pedersen, Anna Braasch, Nicolai H. Sørensen and Thomas Troelsgård 2013. Enriching a wordnet from a thesaurus. *Proceedings of the workshop on lexical semantic resources for NLP at NODALIDA 2013.* NEALT Proceedings Series 19 / Linköping Electronic Conference Proceedings 88: 36–50.

1 Wordnets, ISA overload and the "tennis problem"

Wordnets (Fellbaum 1998; Vossen 1998) are traditionally built around synonym sets with the vertical hyponymy relations as the central structuring principle. This paradigmatic structure is further supplemented by a set of horizontal relations such as antonymy and meronymy. Applying the hyponymy relation as the skeleton for word taxonomies is indeed very convenient, in particular in relation to computational applications since it first of all facilitates the strong inference mechanism of inheritance. However, wordnets as they generally stand appear to lack crucial relations among concepts if they are to be used efficiently as knowledge bases for language technology applications. These include in particular applications requiring some level of "deep" understanding such as information retrieval, question answering, text navigation and text mining.

So, even if hyponymy may include some very basic aspects of the way we organize and conceive concepts in our mental lexicon, and even if this structuring principle is convenient for computers because of its inheritance properties, it is far from sufficient to account for the central relatedness between concepts. First of all, hyponymy does not necessarily results into groups of concepts that are particularly close from a thematic or functional point of view, a phenomenon which is sometimes referred to as the "tennis problem" (cf. Fellbaum 1998, Sampson 2000), pertaining the fact that wordnets traditionally do not account for the relatedness of concepts such as tennis, ball, racquet and net.

Seen from the taxonomical perspective, this lack of expressivity relates to the so-called ISA overload, i.e. the situation where sets of unequal hyponyms are grouped as simple sister terms under the same superordinate, cf. among others Guarino (1998), Guarino & Welty (2002), Huang et al. (2008). To illustrate the problem, consider in the Danish wordnet, DanNet (cf. Pedersen et al. 2009), the hyponyms for concepts like stang ('bar', 'stick') and maske ('mask'). Stang subsumes heterogeneous sets of hyponyms like candy bars, slate pencils, candles, and rods on mens' bicicles, where maske refers to hyponyms like a mask for dressing up for a carnival, diving masks, smoke masks and facial treatments. Thus, the hyponyms subsumed by these synsets may share some very general dimension of form or functionality (i.e. covering the face), but they belong to all sorts of domains and would, in a thesaurus, basically be categorized in a completely different way. Some belong to the food domain, some to entertainment, and others to different professions. In some cases an additional hypernym which clearly refers to the domain is given to these concepts; indicating for instance that a candy bar is also a kind of candy which is again a kind of food, but this is not always possible unless you want to introduce artificial concepts.

In this paper we present two experiments of automatic information transfer from a thesaurus to DanNet: one concerned with thematic information transfer in order to remedy the problem sketched out above (Section 3.1 and 3.2) and one experiment concerned with an extension of the number of synsets on the basis of near synonyms encoded in the thesaurus (Section 3.3).

2 Related work

The suitability of wordnets in intelligent language technology applications has been examined with shifting intensity during the last two decades. In the nineties the Text Retrieval Conferences (TREC) gave rise to a series of thorough testing of Princeton WordNet (PWN) in information retrieval (Voerhees 1993, Voerhees 1994, Voerhees & Harman 1997, Mandala et al. 1998; Gonzalo et al. 1998) without, however, showing radical improvement of system performance. In

2007 the EU project KYOTO (Knowledge-Yielding Ontologies for Transition-Based Organization) was launched as an ambitious, multilingual testing of the wordnet framework meant for mining, structuring and distributing knowledge across languages (Vossen et al. 2008). The project signalled a renewed interest in the use of wordnets for advanced language technology applications such as text mining, question answering, and text retrieval.

The idea of extending standard wordnets with supplementary relations is well-known, see for instance Fellbaum & Miller (2006) for psycholinguistic experiments on associative relations or Veale & Hao's work on folk knowledge in wordnets (2008). The same goes for employing semi-automatic expansion methods from other resources. For instance, in languages with rich productive morphological derivation (such as the Slavic languages), several experiments have been performed in order to semi-automatically capture such morphological relatedness across word classes, as seen in the Czech Wordnet (Pala & Hlaváčková 2007), the Polish WordNet (Piasecki et al. 2010) but also in the Turkish WordNet (Bilgin et al. 2004). Further, the Polish WordNet 2.0 has been enriched with information about verb sub-categorization and semantic classification of aspectual verb pairs. Likewise, innovations in the Hungarian WordNet (Kuti et al. 2008) comprise both language independent and language dependent expansions to the wordnet for verbs and adjectives, and in the Portuguese WordNet (Amaro et al. 2010) an explicit description of argument and event structure is given.

Other wordnets increase their number of relations by inheriting them from the wordnets they link to. The Arabic WordNet Project (Black et al. 2006) which uses the base concept sets of BalkaNet and EuroWordNet as the starting point, obtained a number of semantic relations expressed in SUMO for the English synsets simply by transferring the links from Princeton WordNet to Arabic WordNet.

Similar to the domain-related methods that we are proposing here, are Montoyo et al. (2001) who describe a method to enrich PWN with domain information, arguing that such information provides a natural way to establish semantic relations among synsets. On this approach, wordnet senses are automatically identified in files containing arranged information within a classification system, and the domain information from the file is assigned to the synset in PWN. In addition, earlier work by Navigli and Velardi focuses on relations for domain concepts in the framework of OntoLearn (Navigli & Velardi 2002; Navigli et al. 2004). In SentiWordNet, PrincetonWordNet is semi-automatically extended with sentiment information expressed as polarity values (Baccianella et al. 2010). Further, Veale & Moueddeb (2010) exploit lexical distribution patterns in corpora and semantic similarity scores extracted from WordNet in order to gain more semantic knowledge.

Finally should be mentioned three approaches for extending wordnets on the basis of encyclopedic information. Ruiz-Casado et al. (2005) enrich PWN with encyclopedic definitions from a (rather small) online encyclopedia. In Veale (2006) and Veale & Butnariu (2010) an automated system to extend the number of synsets in PWN is described, building on the extraction and morphological analysis of new words in Wikipedia texts as well as on semantic knowledge from the same text. Finally, Navigli & Ponzetto (2010) describe a similar approach to produce a large, wide-coverage multilingual semantic network. In BabelNet, concepts and relations are automatically extracted from PWN and Wikipedia. The approach involves automatic mapping of Wikipedia pages to PWN synsets after a disambiguation process of candidates from both knowledge sources.

3 Transferring information from a Danish thesaurus to the Danish wordnet, DanNet

3.1 Transferring thematic information from three different levels

Since both thesauri and wordnets arrange concept data in a logical relationship, these resources resemble similar semantic properties to a considerable extent. However, they also differ in several ways. Where wordnets use the hyponymy hierarchy as the primary organizing principle as we have seen, often with an ontology-based division of the world at the uppermost level, thesauri tend to operate with less abstract main categories and a larger number of basic thematic units that go across hyponymies and across parts of speech. In other words, where wordnets are basically ordered vertically, thesauri are rather ordered horizontally – or by themes.

In the present experiment, performed by the Society for Danish Language and Literature (developer of DT and co-developer of DanNet) and the University of Copenhagen (co-developer of DanNet) in collaboration, we exploit the fact that a new Danish thesaurus (DT), which is being compiled at the moment (2009-2013) at DSL, shares common sense IDs with DanNet. Actually, both resources are derived from a third resource, Den Danske Ordbog, a medium-sized contemporary monolingual dictionary of Danish (Hjorth & Kristensen 2005). According to the plan, DT will when it is completed by the end of 2013, contain more than 100,000 word senses compared to DanNet's only 65,000 synsets. The senses are grouped according to a set of different types of relations, and formalized information on the group is annotated in a header. Figure 1 below exemplifies the structure of DT with thematic chapters, sections, subsections and clusters.

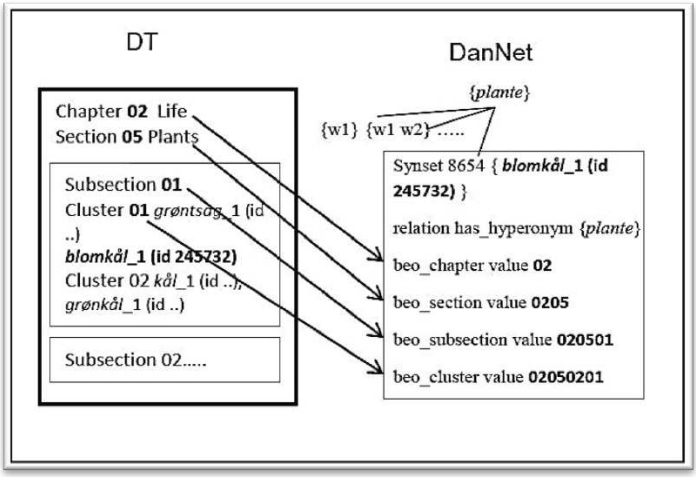
Chapter [section [subsection HEADER [word, cluster[word, word..], cluster[…], word ..] subsection..]]
Chapter2 ***Life*** [section 02.02 ***Plants*** [subsection1 HEADER: has_hyperonym *nytteplante _1* ('utility plant'), concerns *spise_1* ('eat') [cluster1 [*grøntsag_1* ('vegetable'), *blomkål_1* ('cauliflower'), *broccoli_1* ('broccoli'] cluster2 [*kål_1* ('cabbage'), *grønkål_1* ('collard')..] cluster3 [*kornsort_1* ('cereal'), *rug_1* ('rye')..]] [subsection2..]]

Figure 1: Structure of DT exemplified by the chapter 2, Life, the section 02.02 Plants, the subsection of utility plants and clusters of different types of cabbage, cereals etc. (in progress).

As briefly mentioned in the introduction, a well-known problem in the context of hyponymy is the ISA overload where heterogeneous groups of hyponyms are grouped as sisters under the same hypernym. To further illustrate this, person ('person') has more than 6,000 hyponyms in DanNet comprising both persons with inherent characteristics, as well as persons with persistent or temporary roles (Pedersen & Braasch 2009, we avoid artificial hypernyms (in contrast to e.g. PWN which operates with artificial nodes such as 'evil person' etc.). DT also groups a big part of the senses according to a common hyperonym, as the example of utility plants seen in Figure 1. But opposite DanNet, the thesaurus also allows for a flexible placement of concepts in different groups irrespective of whether a common, precise hypernym can be identified or not. In this way, thematic grouping can be provided into e.g. reference to persons having some particular feelings, persons involved in travelling, persons involved in music etc. The domain information in DT is therefore likely to be far more relevant when it comes to a subdivision of the cases of high number of co-hyponyms in DanNet.

The completed thesaurus will consist of 22 chapters divided into approx. 970 sections containing a total of more than 6,000 subsections which contain headers with different types of semantic features. We have carried out the transfer experiment when 1/3 of the thesaurus was completed, using information regarding 2,178 finished subsections (thematic level 3), their corresponding section (thematic level 2) and the chapter to which the section belongs (thematic level 1). Also information on the most detailed semantic level 4, the clusters in DT which typically group near synonyms or synonyms was transferred ; we will return to the use of this data in 3.3.

Since each "synonym" (i.e. lexical representation) in a DanNet synset share the unique id with the relevant sense or senses in DT, the transfer was carried out by assigning the level numbers to the synsets, cf. Figure 2.



Figur 2: The transfer of thematic and semantic information on synset members from DT to their corresponding synset in DanNet.

Four different level numbers (corresponding to chapter, section, subsection and cluster) were assigned the synsets in order to be able to experiment with the data in different ways. Only one member of a synset had to be represented in DT in order for the number values to be assigned to the whole synset. In case of conflicting values the level numbers for that synset were discarded. The transfer resulted in thematic assignment of level numbers to 17,816 synsets.

In the case of large groups of co-hyponyms in DanNet, we assume intuitively that the section level with its 970 divisions (corresponding to 'beo_section value 0205' in Figure 2) is quite informative when it comes to a subdivision. As an example, the completed third of DT contains 15 subsections with co-hyponyms of stang (bar, pole) distributed among 15 different sections. Pløk ('plug', 'peg') and teltstang ('tent pole') appear in the same section as the only ones and are thereby both grouped together and sorted out from the other 13 hyponyms.

In cases where a number of co-hyponyms belong to the same section but to different subsections, the subsection domain numbers (eg. the beo_subsection value 020501 in Figure 2) are the ones that reduce the ISA overload. For example, words denoting 'persons who dislike something or somebody' constitute only one of the two 'person' subsections in the same section ('Dislike') (the other one grouping different words for persons to be disliked). In such cases, the section domain information transferred from DT is not enough to delimit the relevant words in DanNet. A similar example is seen in Figure 3 for food words from the section 'Food and Dishes' in DT.

Figure 3: Hyponyms of food divided in three semantic groups in DT according to different meaning aspects such as **1) 'concerns': main ingredient**: fish dish [pasta dish, spaghetti dish] rice dish, vegetable dish [meat dish, dish of minced meat, venison dish, chicken dish] potato dish, egg dish, **2) 'concerns': consistency and quantity: cream**, goo [**puree**, mash, mousse] [blob, sprinkling], [**drop/spot**, splash] [sip, mouthful, swallow, gulp, bite, morsel] **[slice/cut**, hunk, chunk, lump, section] lump of meat, hunk of bread, [sprinkling of (chopped) parsley, sprinkling of chives, sprinkling of sugar] knob/nut of butter, blop of whipped cream] [section of a lemon, section of an apple] [**serving**, ration, bowl, bowlful, plate, plateful], and **3) 'concerns': time**[breakfast, lunch dish, dinner meal, midnight snack][summer dish, Christmas meal, Easter meal]. Bold words in DT function as keywords.

In all three subsections the hypernym is food, but the subsection information distinguishes between three semantic dimensions of food 1) the major ingredient, 2) the consistency or quantity and 3) the time when it is eaten.

Within the field of terminology a similar method to distinguish between co-hyponyms via semantic dimensions has been introduced (Madsen & Thomsen 2009; Madsen et al. 2004), but contrary to this method, the unique semantic criteria which connects the words of a certain group (e.g. concerns konsistens (consistency) in Figure 3) is not always made explicit in DT.

Furthermore, some subsection divisions are established in DT from purely thematic reasons (and marked as such in the header) when precise semantic relations, such as for example a common hypernym, are impossible to assign to the group. Actually, 12 % of the subsections in DT consists of words grouped together just for thematic reasons. E.g. in the section 1.2 Himmellegemer ('heavenly bodies') all words concerning the sun (korona ('corona'), solvind ('solar wind'), solbane ('path of the sun') etc.) are grouped together in one subsection of this type and simply assigned the relation concerns sol ('sun') in the header. Likewise, words concerning golf (golfklub ('golf club'), golfbane ('golf course'), par (par) etc.) are grouped together in the

same type of subsection and assigned concerns golf ('golf') in the header. Along the same line, Figure 4 shows examples of boxing terms from different part of speech in the same subsection.

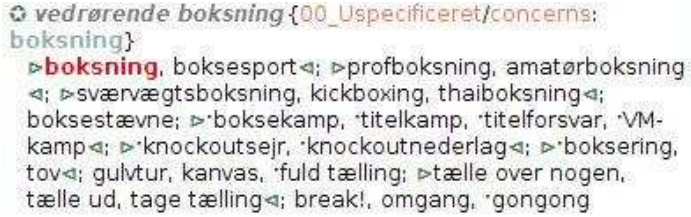


Figure 4: Boxing words in DT, such as 'boxing', professional boxing' 'amateur boxing', 'take the count', 'ring' etc.

The transferred data on subsection number for the words in these groups to their corresponding synsets in DanNet may constitute suitable answers to the tennis problem. However, in some cases it is probably more convenient to apply the more coarse-grained thematic groupings (section and chapter). To illustrate the classical group of concepts mentioned above, tennis, raquet, net, and ball, these belong to different subsections of 'Raquet sports' (badminton, tennis and squash) and are only related via the more broad section division.

To conclude, there is no doubt of the fact that the relevant information is indeed present in DT, but it should be investigated further and at a larger scale which detail of thematic information proves to be most appropriate to the majority of the cases.

3.2 Assessment of the coarse-grained divisions

At the current stage of work in progress, only the coarse-grained thematic groupings at chapter level have been systematically assessed. Two percent of these assignments were manually judged in order to see to which extent the subdivision of co-hyponyms on the basis of this raw material made sense. The result was that 4,6 % of the assigned themes were not considered fully intuitive. Among these, several, however, made sense when considering DT in more detail. For instance, all wooden materials were assigned the theme "Equipments, technology", simply because they were described as materials for producing furniture and buildings.

The transferred data on subsection number for the words in these groups to their corresponding synsets in DanNet may constitute suitable answers to the tennis problem.

All in all, the experiment showed a considerable enrichment of the data in spite of its coarse-grainedness. In Figure 5 is shown how some of the many direct hyponyms in DanNet of the synset {egenskab_1, beskaffenhed_1} ('property') are sorted in intuitively meaningful groups based on chapter information in DT. In Nimb & Pedersen (2012) we discuss in detail how information on properties in DT are transferred and used in the form of a new semantic relation in DanNet.

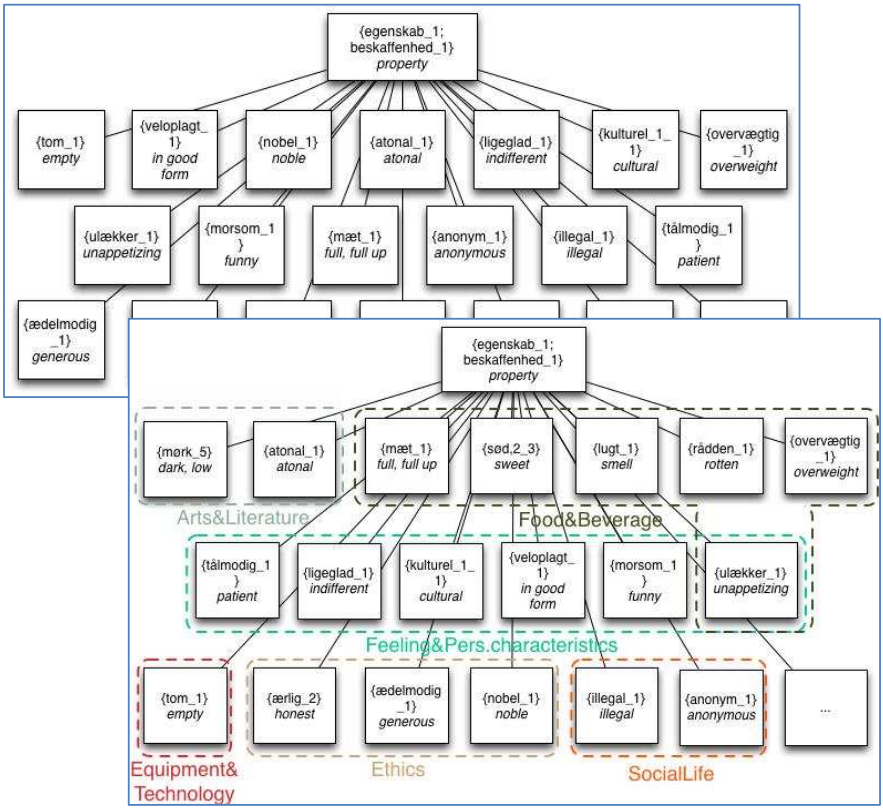


Figure 5: Hyponyms of the synset {egenskab} ('property') in DanNet, above presented without DT values, below sorted out in thematic groups based on the assigned DT chapter values.

Also Figure 6 shows some of the thematic groupings of co-hyponyms that appeared after the coarse-grained assignment based on chapter values in DT, in this case under persons, masks and bars.

Persons	Masks	Sticks/bars
Sports&leisure	**Equipment&technology**	**Food&beverage**
gæst_1 (guest)	*maske_1_1* (mask)	*stang_2* (bar)
bryllupsgæst_1 (wedding guest)	**Life**	*chokoladebar_1* (chocolate bar)
middagsgæst_1 (dinner guest)	*muddermaske_1* (mud mask)	*ostebjælke_1* (cheese stick)
sportspige_1 (sportsgirl)	*ansigtsmaske_1* (face mask)	**Location&change**
lilleput_1 (pre-teen)	**Sports&leasure**	*stang_1* (pole)
gymnastiklærer_1 (gym master)	*dykkermaske_1* (diving mask)	*bom_1* (balance beam)
Feelings&pers.characteristics	*fægtemaske_1* (fence mask)	*lossebom_1* (derrick)
ønskebarn_1 (planned child)	*fastelavnsmaske_1* (halloween mask)	*jernbanebom_1* (railway barrier)
kælledægge_1 (darling)		
møgunge_1 (kiddie)		
øjesten_1 (apple of ones eye)		
heltinde_1 (heroine)		
bøddel_1 (tormentor)		
Art&culture		
balletbarn_1 (ballet child)		
geisha_1 (geisha)		
mavedanser_1 (belly dancer)		
sanglærer_1 (song teacher)		
musikforsker_1 (musicologist)		
kubist_1 (cubist)		
skjald_1 (scald)		
Location&change		
bærer_1 (carrier)		
rumforsker_1 (space researcher)		
Equipment&technology		
håndværker_1 (workman)		
tømrer_1 (carpenter)		
murer_1 (brick layer)		
bygmester_1 (master builder)		
saddelmager_1 (saddler)		
arkitekt_1 (architect)		
Life		
skabsbøsse_1 (closet queen)		
elsker_1 (lover)		
muskelmand_1 (muscleman)		

Figure 6 : Three examples of automatic thematic groupings of synsets in DanNet with identical hyponyms (persons, masks and sticks), based on transferred chapter values from DT.

From the opposite perspective, Figure 7 exemplifies how the coarse-grained chapter level information now relates concepts that otherwise are unrelated in DanNet like instruments used for cooking, containers for containing food, food itself, properties of food, eating and cooking events as well as persons involved in eating.

Food&Beverage
bestik_1 (cutlery)
ravioli_1 (ravioli)
tærteform_1 (baking tin)
slikmund_1 (sweet-tooth)
slubre_1 (to slurp)
gennemstegning_1 (cooking to be well-done)

Figure 7: Some of the **t**erms in DanNet which have been related via the Food&Beverage theme information from DT

3.3 Automatic compilation of new synsets in DanNet based on DT

Based on the most fine-grained thematic level in DT, where we find clusters of near synonyms, we have further experimented with the automatic compilation of new synsets in DanNet. A well-known corpus-based method for extending the coverage of a lexical-semantic resource is to examine syntactic patterns such as enumerative noun phrases and look for unknown words in the phrases. In such investigations, the semantics of new words is guessed upon with some accuracy based on the information from the already known words in the phrase (see for instance Kokkinakis et al. 2000). Our approach is conceptually similar to this method; however, we base our compilation not on enumerations in a corpus, but on the more precise near synonyms given in DT. If a new word is listed in a cluster in DT where at least two DanNet synsets are already represented, then we consider the new word to be of the same synset type (i.e. with same ontological type and the same hyperonym) given that the two known synsets have identical hypernyms and identical ontological types.

The current experiment results in 440 new synsets, an excerpt of which can be seen in Figure 8. An assessment of 10 % of these indicates that the method is indeed very precise since all the evaluated synsets were assigned a correct hypernym as well as a correct ontological type. All in all we estimate that we can automatically generate 1,500 new synsets using this method on the completed DT data. In future, however, we plan to investigate further whether we can extend the DanNet coverage also on the basis of less precise data in DT. Approx. half of the subsections in DT consist of co-hyponyms, and we consider this encoding to be of significant value for automatically generating new synsets. In accordance with the experiment described above, information on existing synsets in the same subsections will be considered. Since DT already contains more than 20,000 concepts not represented in DanNet, the amount of potentially easy-accessible material for extending DanNet is considerable.

Near synonyms	New synset(s)
espresso_1 *café au lait_1,* *cappuccino_1*	*caffé latte_1*
es_1_1 (ace) *toer_3* (two/deuce) *treer_2* (three) *firer_3* (four) *femmer_2*(five) *sekser_1* (six)	*syver_1* (seven) *otter_2* (eight) *nier_1* (nine) *tier_2* (ten) *joker_1* (joker)
jaloux_1 (jealous) *skinsyg_1* (jealous) *syg_1_1* (compulsive)	*besidderisk_1* (possessive)
kuffert_1 (suitcase) *rejsetaske_1* (travel bag) *rygsæk_1* (backpacker)	*læderkuffert_1* (leather suitcase) *håndkuffert_1* (gripsack)

Figure 8: Examples of new synsets in DanNet based on near synonyms in DT

4 Conclusions

It is not an easy task to define which semantic relations are the crucial ones for automated, intelligent information handling. Each application can be seen as having its own very particular requirements regarding relevant cognitive associations. Put to the extreme, there seems to be infinite dimensions of meaning similarity and infinite ways in which concepts can relate to each other and it is unattainable to provide a full lexical semantic network which contains all relations of potential relevance.

Nevertheless, in this paper we have argued that classical hyponymy is often underspecified with regard to some very central meaning dimensions such as thematic context and particular use. Also, we have seen that many wordnets, including DanNet, lack important relations across part of speech. Classical thesauri have obvious resemblances with wordnets, but they differ with respect to the criteria used to carve up the conceptual world. For example, as we have shown, thematic relations are in fact well-represented in these resources.

Because of the close connection between the two resources DT and DanNet, both based on the original dictionary, DDO, and both maintaining the same sense IDs, transfer from one to the other is in fact technically feasible and profitable as our experiments have indicated. The fact that all word clusters in DT are XML tagged with one or more semantic relation types makes transfer directly practicable, especially in the cases where the relations are not currently present in DanNet. In contrast, supplementary, more functionally oriented hyponymy relations based on the DT have to be introduced semi-automatically to prevent clashes with the existing taxonomies.

Our experiment has shown that DanNet can be extended profitably by adding the different thematic levels given in DT in order to be able to distinguish between high numbers of co-hyponyms and thematically to relate concepts across the hierarchy. Further, the demonstrated method of adding new synsets to DanNet on the basis of near synonyms in DT has shown the potential for further research in this area.

References

Amaro, Raquel, Sara Mendes & Palmira Marrafa (2010). Encoding Event and Argument Structures in Wordnets. TSD 2010, LNAI 6231, 21–28. Berlin Heidelberg: Springer-Verlag. DOI:10.1007/978-3-642-15760-8.

Baccianella, Stefano, Andrea Esuli & Fabrizio Sebastiano (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. Proceedings of 7th LREC - Language Resources and Evaluation. Paris: ELRA (European Language Resources Association). http://www.lrec-conf.org/proceedings/lrec2010/index.html.

Bilgin, Orhan, Özlem Cetinoglu & Kemal Oflazer (2004). Building a Wordnet for Turkish. Romanian Journal of Information, Science and Technology, 7 (1-2), 163-172. Bucarest: Editura Academiei Române.

Black, William, Sabri Elkateb, Horacio Rodriguez, Musa Alkhalifa, Piek Vossen, Adam Pease, Christiane Fellbaum (2006). Introducing the Arabic Word Net Project. Petr Sojka, Key-Sun Choi, Chritiane Fellbaum, Piek Vossen (Eds.) Proceedings of the third International WordNet Conference (GWC-06). Brno: Masaryk University. http://NLPweb.kaist.ac.kr/gwc/ pdf2006/74.pdf

Braasch, A. & B.S. Pedersen (2010). Encoding Attitude and Connotation in Wordnets . In: The 14th EURALEX International Congress, Leeuwarden , The Netherlands.

Fellbaum, Christiane (ed) (1998). WordNet – An Electronic Lexical Database. Cambridge, Massachusetts, London, England: The MIT Press.

Fellbaum, Christiane, Georg A. Miller (2006). Whither Wordnets? Zampolli Prize Presentation at LREC 2006, Genova. http://www.lrecconf.org/lrec2006/IMG/pdf/ AZPrize.Christiane%20 Fellbaum%20Presentation.LREC06.pdf.

Fellbaum, Christiane & Piek Vossen (2008). Challenges for a Global WordNet. Online Proceedings of the First International Workshop on Global Interoperability for Language Resources (ICGL 2008), 75-82. Hongkong: City University of Hongkong. http://icgl.ctl.cityu.edu.hk/2008/html/resources/~proceeding_conference.pdf.

Gonzalo, Julio, Felisa Verdejo, Carol Peters & Nicoletta Calzolari (1998). Applying EuroWordNet to Cross-Language Retrieval. Computers and the Humanities. 32 (2/3), 185-207. The Netherlands: Kluwer Academic Publishers.

Guarino, Nicola (1998). Some Ontological Principles for Designing Upper Level Lexical Resources. Proceedings from the First International Conference on Language Resources and Evaluation, 527–534. Granada.

Guarino, Nicola & Chris Welty (2002). Identity and Subsumption. Green, R., Bean, C.A. & Myaeng, S. H. (Eds.), The Semantics of Relationships: An Interdisciplinary Perspective, Information Science and Knowledge Management. Springer Verlag.

Hjorth, Ebba & Kjeld Kristensen (eds.) (2005). Den Danske Ordbog. Copenhagen: Gyldendal & Det Danske Sprog- og Litteraturselskab. Online version: http://ordnet.dk/ddo.

Huang, Chu-Ren., I-Li Su, Pei-Yi Hsiao, Xiu-Ling Ke (2008). Paranymy: Enriching Ontological Knowledge in WordNets. Proceedings of the Fourth Global WordNet Conference, 221–228. Szeged, Hungary: Juhász Press Ltd.

Kokkinakis, Dimitrios, Maria Toporowska Gronostaj, Karin Warmenius (2000). Annotating, Disambiguating & Automatically Extending the Coverage of the Swedish SIMPLE Lexicon. Proceeding LREC 2000, 1397-1403. Paris, France: ELRA

Kuti, Judit, Károly Varasdi, Ágnes Gyarmati, & Péter Vajda (2008). Language Independent and Language Dependent Innovations in the Hungarian WordNet. Proceedings of the Fourth Global WordNet Conference. 254-268. Szeged, Hungary: Juhász Press Ltd.

Madsen, Bodil Nistrup, Hanne Erdman Thomsen, & Carl Vikner (2004). Comparison of Principles Applying to Domain-Specific versus General Ontologies. Ontolex 2004, 90-95. Paris, France: ELRA.

Madsen, Bodil Nistrup & Hanne Erdman Thomsen (2009). Ontologies vs. Classification Systems. Proceedings of the NODALIDA 2009 workshop WordNets and other Lexical Semantic Resources — between Lexical Semantics, Lexicography, Terminology and Formal Ontologies. NEALT Proceedings Series 7, 27-32. Tartu: Northern European Association for Language Technology (NEALT) and Tartu University. http://dspace.utlib.ee/dspace/handle/10062/9840.

Mandala, Rila, Takenobu Tokunaga, & Hozumi Tanaka (1998). The use of WordNet in Information Retrieval. Proceedings of the COLING-ACL workshop on Usage of Wordnet in Natural Language Processing, 31– 37. Montreal, Canada: ACL / Morgan Kaufmann Publishers.

Montoyo, Andrés, Manuel Palomar and German Rigau (2001). Method for WordNet Enrichment using WSD. Matousek, V. ,P. Mautner R. Moucek and Karel Tauser (eds.) Proceeding TSD 2001 Lecture Notes in Computer Science,Volume 2166 , 180-186. Springer.

Navigli, Roberto & Simone Paolo Ponzetto (2010). BabelNet: Building a Very Large Multilingual Semantic Network. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 216-225. Uppsala, Sweden. Association for Computational Linguistics.

Navigli, Roberto & Paola Velardi (2002). Automatic Adaptation of Wordnet to Domains. Proceedings of the Third International Conference on Language Resources and Evaluation (LREC), 1499-1504. Paris, France: ELRA.

Navigli, Roberto, Paola Velardi, Alessandro Cucchiarelli & Francesca Neri (2004). Extending and Enriching WordNet with OntoLearn. Proceedings of The Second Global Wordnet Conference - GWC 2004. Brno: Masaryk University. http://www.dsi.uniroma1.it/~navigli/pubs/GCW_2004_Navigli_al.pdf.

Nimb, S. & B.S. Pedersen (2012). Towards a richer wordnet representation of properties – exploiting semantic and thematic information from thesauri. In: LREC 2012 Proceedings pp. 3452-3456. Istanbul, Turkey.

Pala, Karel & Dana Hlaváčková: Derivational Relations in Czech WordNet (2007). ACL '07 Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies. Stroudsburg, PA, USA: Association for Computational Linguistics. http://portal.acm.org/citation.cfm?id=1567559.

Pedersen, Bolette.S. & Patrizia Paggio (2004). The Danish SIMPLE Lexicon and its Application in Content-based Querying. Nordic Journal of Linguistics 27 (1), 97-127. Cambridge University Press.

Pedersen, Bolette S, Sanni Nimb, Jørg Asmussen, Nicolai Sørensen, Lars Trap-Jensen & Henrik Lorentzen (2009). DanNet: The challenge of compiling a WordNet for Danish by reusing a monolingual dictionary. Language Resources and Evaluation, Computational Linguistics Series 43 (3), 269-299, doi:10.1007/s10579-009-9092-1.

Pedersen, B.S. & A. Braasch (2009). What do we need to know about humans? A view into the DanNet Database. In: K. Jokinen and E. Bick (eds.) Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009. NEALT Proceedings Series, Vol. 4, Odense, Denmark.

Pianta, Emanuele, Luisa Bentivogli & Christian Girard (2002). MultiWordNet – Developing an aligned multilingual database. Proceedings of the First International Conference on Global WordNet, 293-302. Mysore, India.

Piasecki, Maciej, Stanislaw Szpakovicz & Bartosz Broda (2010). Toward plWordNet 2.0. Proceedings of the 5th International Conference on Global Wordnet (GWC2010), 263-270. Mumbai: Narosa Publishers.

Ruiz-Casado, Maria, Enrique Alfonesca & Pablo Castells (2005). Automatic Assignment of Wikipedia Encyclopedic Entries to WordNet Synsets. Piotr S. Szczepaniak, Janusz Kacprzyk, Adam Niewiadomski (Eds.): Advances in Web Intelligence Third International Atlantic Web IntelligenceConference, AWIC 2005, Lodz, Poland, Proceedings. Lecture Notes in Computer Science 3528. Springer

Sampson, Geoffrey (2000). Review of WordNet: An Electronic Lexical Database. In International J. of Lexicography 13.54–9, 2000.

Veale, Tony (2006). Tracking the Lexical Zeitgeist with WordNet and Wikipedia. Proceedings of the 17th European Conference on Artificial Intelligence (ECAI 2006), IOS Press, 56-60. Amsterdam, The Netherlands.

Veale, Tony & Yanfen Hao (2008). Enriching WordNet with Folk Knowledge and Stereotypes. Proceedings of the Fourth Global WordNet Conference, 453-461. Szeged, Hungary: Juhász Press Ltd.

Veale, Tony & Cristina Butnariu (2010). Harvesting and understanding on-line neologisms. Alexander Onysko, Sascha Michel (eds.) Cognitive Perspectives on Word Formation. 399-420. De Gruyter Mouton.

Veale, Tony & Mourad el Mouedden (2010). Similarity, Comparability and Analogy in WordNet: Squaring the Analogical Circle with Mondrian.

Proceedings of the 5th International Conference on Global Wordnet (GWC2010). Mumbai: Narosa Publishers. http://afflatus.ucd.ie/Papers/Mondrian%20GWC%20paper.pdf

Voorhees, E.M. (1993). Using wordnet to disambiguate word senses for text retrieval. Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, 171-180. New York, NY, USA: ACM.

Voorhees, Ellen M. (1994). Query expansion using lexical-semantic relations. Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, 61-69. New York: Springer-Verlag New York, Inc.

Voorhees, Ellen M. & Donna Harman (1997). Overview of the fifth text retrieval conference (trec-5). Proceedings of the Fifth Text Retrieval Conference, 1-28. NIST Special Publication 500-238. Gaithersburg: NIST. http://trec.nist.gov/pubs/trec5/t5_proceedings.html

Vossen, Piek, Eneko Agirre, Nicoletta Calzolari, Christiane Fellbaum, Shu-Kai Hsieh, Chu-Ren Huang, Hitoshi Isahara, Kyoko Kanzaki, Andrea Marchetti, Monica Monachini, Feririco Neri, Remo Raffaelli, German Rigau, Maurisio Tesconi & Joop CanGent (2008). KYOTO: A System for Mining, Structuring and Distributing Knowledge Across Language and Culture. Proceedings of the Fourth Global WordNet Conference, 474-484. Szeged, Hungary: Juhász Press Ltd.

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

Sponsors of
NODALIDA 2013 & NEALT

WeSearch

iness

Lingit

max manus

computas

The Center of Estonian Language Resources

DET HUMANISTISKE FAKULTET
KØBENHAVNS UNIVERSITET

GSLT

Lingsoft
LANGUAGE SOLUTIONS

Mikro Værkstedet

National Library of Norway

textUrgy

www.tungutaekni.is

Design • Joel Priestley
Photo • Arthur Sand

NEALT Proceedings Series 19 • ISBN 978-91-7519-586-5
Linköping Electronic Conference Proceedings 88
ISSN 1650-3740 (Online) • ISSN 1650-3686 (Print) 2013