NEALT PROCEEDINGS SERIES

VOL. 7

Proceedings of the NODALIDA 2009 workshop

WordNets and other Lexical Semantic Resources – between Lexical Semantics, Lexicography, Terminology and Formal Ontologies

May 14, 2009
Odense, Denmark

Editors

Bolette Sandford Pedersen
Anna Braasch
Sanni Nimb
Ruth Vatvedt Fjeld

NORTHERN EUROPEAN ASSOCIATION FOR LANGUAGE TECHNOLOGY

Proceedings of the NODALIDA 2009 workshop

WordNets and other Lexical Semantic Resources - between Lexical Semantics,
Lexicography, Terminology and Formal Ontologies

NEALT Proceedings Series, Vol. 7

ISSN 1736-6305

Published by
Northern European Association for Language
Technology (NEALT)
http://omilia.uio.no/nealt

Electronically published at
Tartu University Library (Estonia)
http://dspace.utlib.ee/dspace/handle/10062/9209

Volume Editors
Bolette Sandford Pedersen
Anna Braasch
Sanni Nimb
Ruth Vatvedt Fjeld

Contents

Preface

NODALIDA 2009 workshop

WordNets and other Lexical Semantic Resources – between Lexical Semantics, Lexicography, Terminology and Formal Ontologies

Half-day workshop held May 14 at Nodalida 2009, Odense, Denmark. The goal of the workshop was to bring together researchers involved in building and integrating lexical semantic resources as well as researchers that are more theoretically interested in investigating the interplay between lexical semantics, lexicography, terminology and formal ontologies.

The workshop organizers received 14 papers for review and each paper was reviewed by two reviewers of the Programme Committee. Borderline papers with considerable disagreement between reviewers have undergone an additional judgment by a third member of the Programme Committee. Of the 14 papers, 8 were selected for presentation and publication, 4 were rejected, and 2 were withdrawn by the authors.

Workshop organizers

Bolette Sandford Pedersen, University of Copenhagen, Denmark

Anna Braasch, University of Copenhagen, Denmark.

Ruth Vatvedt Fjeld, University of Oslo, Norway

Sanni Nimb, Society for Danish Language and Literature, Denmark.

Programme Committee

Claudia Kunze, Universität Tübingen, Germany

Helge Dyvik, Unversity of Bergen, Norway

Janne Bondi Johannesen, University of Oslo, Norway

Maria Toporowska Gronostaj, Språkdata, Göteborg University, Sweden

Tamás Váradi, Hungarian Academy of Sciences, Hungary

Heili Orav, University of Tartu, Estonia

Jørgen Fischer Nilsson, Technical Unviersity of Denmark, Denmark

Hanne Erdmann Thomsen, Copenhagen Business School, Denmark

Ruth Vatvedt Fjeld, University of Oslo, Norway

Anna Braasch, University of Copenhagen, Denmark

Bolette Sandford Pedersen, University of Copenhagen, Denmark

Sanni Nimb, Society for Danish Language and Literature, Denmark.

Lars Trap-Jensen, Society for Danish Language and Literature, Denmark

Henrik Lorentzen, Society for Danish Language and Literature, Denmark

Jørg Asmussen, Society for Danish Language and Literature, Denmark

Bodil Nistrup Madsen, Copenhagen Business School, Denmark

Automatic Extraction of Semantic Relations For Less-Resourced Languages

Anna Björk Nikulásdóttir
University of Iceland
Reykjavík, Iceland
abn@hi.is

Matthew Whelpton
University of Iceland
Reykjavík, Iceland
whelpton@hi.is

Abstract

This paper addresses the challenge of creating a network of semantic relations for languages which do not have the resources of investment and manpower which have allowed the development of resources like WordNet for English. We first present a pilot study in this area which used a well-established pattern-based method to extract semantic relations from an Icelandic monolingual dictionary. This proved to have a good success rate for ten semantic relations. We then present a newly funded project which aims to extend and adapt this methodology for use with unstructured tagged corpora. We hope that this will allow the largely automated development of the target semantic resources.

1 Introduction

Although Icelandic language technology (LT) has taken great strides forward in the last ten years (Rögnvaldsson 2008), there are as yet no specifically LT-oriented semantic resources for Icelandic. However, Iceland has a rich lexicographic tradition which provides an excellent starting point for the development of such semantic resources. A pilot study in the exploitation of lexicographic material for the extraction of semantic relations has already been performed by Nikulásdóttir (2007) for Icelandic, building on classic studies for English (Alshawi 1987; Chodorow et al. 1985; Markowitz et al. 1986; Nakamura and Nagao 1988) and more recent work on Basque (Agirre et al. 2000). The pilot study gave promising results, with 94.77% of the analysed definitions being correctly or partly correctly analysed. A Grant of Excellence has just been awarded by the Icelandic Research Fund to the project “Viable Language Technology beyond English – Icelandic as a test case” (hereafter VLT), the first work package of which aims to extend Nikulásdóttir´s work in developing a semantic network for Icelandic. We hope that the resources developed and the experience acquired in this project will i) lay the foundation for the development of a WordNet-like (cf. Fellbaum 1998) resource for Icelandic, and ii) serve as guidelines for other less-resourced languages for automatically extracting semantic relations.

Sections 2 to 4 of this paper offer an overview of Nikulásdóttir (2007). Nikulásdóttir used definitions in the 2002 3rd Edition of *Íslensk orðabók* ‘Icelandic Dictionary’ (henceforth ÍO) for her pilot study. Section 2 describes the characteristic format of noun definitions in the dictionary (2 main formats) and the issues that relate to these definition formats. Section 3 reviews ten semantic relations which were automatically extracted from definitions of nouns in ÍO: hypernyms, synonyms, holonyms, meronyms, verbal endonyms, adjectival endonyms, attributes, biological family, equivalences, and references. In Section 4, the results of the automatic extraction process are evaluated. The 94.77% success rate of the automatic extraction provides an encouraging basis for further work. However, practical considerations for a less-resourced language like Icelandic require the ability to extract semantic relations from large corpora of free text. The newly-funded VLT Project aims to develop methodologies to address this issue by supplementing the pattern matching methodology of the pilot study with latent semantic and coordination techniques and established statistical methods. This is discussed in Section 5.

2 Definition Formats in the Icelandic Dictionary

Definitions of nouns in monolingual dictionaries, such as ÍO, use certain syntactic patterns repeatedly in the formulation of particular kinds of definition (Geeraerts 2003: 89). It is therefore possible to exploit the correlation between syntactic patterns and semantic relations for automatic extraction. The most common formats for noun definitions in ÍO are a) synonym definitions or synthetic definitions and b) a paraphrase, including *genus proximum* and *differentias specificas* (cf. Geeraerts 2003: 89). The use of the term *genus proximum* is, however, not unproblematic, since it should refer to the closest taxonomical hypernym. The head noun in a definitional paraphrase in a dictionary does not necessarily fit to that description, even if it represents a hypernym (Wiegand 1989: 548). We prefer to describe the paraphrasal definition as including a hypernym with features that distinguish the lemma from its co-hyponyms.

a) synonym definition:
fagnaður 1 ánægja, gleði
joy 1 *pleasure, gladness*

b) a paraphrase:
breiðband breitt tíðnisvið [*gen. prox.*] notað til fjarskipta, [...]
broadband *a broad frequency range used for telecommunication, [...]*

In the ÍO database, definitions are segmented with regard to meaning items. A meaning item is a subpart of a definition serving various lexicographic functions (N.B. a definition may comprise just one meaning item). The following definition for example includes three meaning items:

> **dílaburkni** [1] íslensk burknategund [2] *(Dryopteris assimilis)* [3] af þrílaufungsætt, með fjaðurskiptum blöðum, vex í gjám og kjarri
> **...** [1] *an Icelandic species of fern* [2] *(Dryopteris assimilis)* [3] *of the three-leaved variety, with feathered leaves, grows in crevices and thickets*

319 different functions are defined for meaning items in ÍO, 22 of which are exploited in the present study. For instance, meaning item [1] for *dílaburkni* provides the relation of **hypernym** to *burknategund* but meaning item [2] is discarded as it does not contain Icelandic lexemes.

All relevant meaning items were tagged for part of speech (POS) with the TnT statistical POS-tagger from Brants (2000), which had been trained on Icelandic data. We only used the word class information from the tagger, except for nouns, where case tags were included as well. This resulted in 106,977 POS-tagged meaning items. The method showed here was developed assuming that the first POS of a meaning item is an important indicator of what semantic relations are likely to be included in the meaning item and how these can be extracted. The vast majority of the meaning items start with a noun, 68.51%. Of these items 48% consist of only one noun. The second largest group of meaning items comprises those starting with an adjective, 13.73% of all analysed meaning items. All POS-tags were extracted from the items to build POS-patterns. These are named according to the first POS-tag, e.g. patterns extracted from items starting with a noun are called *N_Patterns*.

3 Extraction of semantic relations

After analysing the five most important groups of patterns, including those starting with a noun, an adjective, a pronoun, an adverb and a verb, algorithms for extracting semantic relations were developed. As indicated above, the algorithms begin with the first POS-tag of the definition, in order to narrow down the range of possible semantic relations. Every pattern-group is then analysed in a specific way, searching for POS-patterns or lexicosyntactic patterns indicating a semantic relation (cf. Hearst 1998). One aim of the analysis was to extract as many kinds of relations as possible, so that the success of using this methodology on different relation types could be evaluated. The study was therefore not limited to one or two relations, such as hypernymy and synonymy. All in all, ten relations were extracted, including equivalence and references, which are also marked in the ÍO database. These relations are reviewed next.

3.1 Hypernyms

As described in 2, the typical paraphrasal dictionary definition includes a hypernym of the lemma. Among the patterns indicating a hypernym are:

(1) `adj noun`
(2) `adj (,|conj) adj noun`
(3) `noun .+`

In all cases `noun` represents the hypernym:

spenna ótryggt (adj) ástand (noun)
(***tension*** *precarious situation*)
`hypernym(spenna, ástand)`

The pattern `noun .+` (i.e. noun plus anything) has several exceptions, where e.g. synonyms or more than one hypernym are extracted.

3.2 Synonyms

Normally, in dictionary definitions of nouns which consist of one noun or a list of nouns, each noun in the definition is a synonym of the lemma:

> **meiðing** barsmíð, líkamsárás
> ***beating-up*** *thrashing, physical assault*

The synonyms extracted in this way are rarely absolute synonyms and can have quite different connotations. This is characteristic of the general problem of synonymy. Cruse (1986:88) defines propositional synonymy as the relation between two syntactically identical words which, when interchanged in the same context, will not change the proposition of the corresponding sentence. In WordNet, synonymy is defined as the relation between words that are substitutable in *some* contexts (Miller 1998: 24). Synonymy could thus be seen as propositional synonymy where the substitutability only has to be valid in some contexts.

Despite this broad definition of synonymy, not all definitions that have the format of a synonym definition (cf. Section 2) can be seen as containing synonyms but rather represent a hypernym-hyponym relation:

> **garg** fuglahljóð
> ***screech*** *a bird-sound*

In this case an underlying “is-a-kind-of” is not explicitly expressed.

In the ÍO database, the meaning items of definitions in this style are mostly marked as equivalences. This is a misleading term, since equivalences normally hold between words in two different languages.

3.3 Equivalences and references

As stated in 3.2, there is a meaning item in ÍO labelled as “equivalence”. An equivalence should consist either of a noun or a listing of nouns, representing synonyms of the lemma. This prescription has not however been followed consistently in ÍO. Sometimes equivalences represent hypo- or hypernyms and even complex sentences are occasionally marked as equivalences.

ÍO also independently labels “references”, which are meaning items containing one noun or a list of nouns, somehow semantically related to the lemma. Randomly selected references represented hyponymy, hypernymy, antonymy and meronymy.

As equivalences and references are independently labelled in ÍO and are inconsistent in semantic type, they are simply extracted by label (and only if they contain only nouns, which are the target of this study).

3.4 Holonyms and meronyms

Holonyms are extracted where, in a first run, a hypernym *hluti* (‘part of’) has been recognized. In these definitions, the hypernym is rejected and the next noun is extracted as a holonym of the lemma:

> **fingurgómur** fremsti hluti fingurs [...]
> ***fingertip*** *the foremost part of a finger*
> `holonym(fingurgómur,fingurs)`

The lexico-syntactic pattern indicating meronymy is:

> `noun(, noun )*og noun.*`

where all nouns are meronyms of the lemma. The extracted meronyms are of different kinds: ‚X is part of Y’, ‚X and Z build Y’ (*bride* and *groom* build a *bridal couple*), to be a Y includes being X and Z’ (being *a troubadour* includes being *a poet* and *a musician*). Another kind of relation related to meronymy is the member-group relation. As with holonyms, an extracted hypernym is tested to look for a member-group indication. If it is the word *hópur* (‚group’) it will be rejected and the next noun extracted as the members of the group named by the lemma:

> **leshringur** hópur fólks sem [...]
> ***reading group*** *group of people that [...]*
> `member(leshringur, fólks)`

3.5 Verbal and adjectival endonyms

Sometimes nouns are paronyms of verbs or adjectives, which in turn constitute endonyms of the corresponding nouns (cf. Cruse 1986). These nouns are often defined in terms of the endonyms.

íhugun (n)	íhuga (v)
consideration	*consider*
björgun (n)	bjarga (v)
rescue	*rescue*

frægð (n)	frægur (adj)
fame	*famous*
heiðarleiki (n)	heiðarlegur (adj)
honesty	*honest*

In some cases the extracted endonym is not morphologically related to the noun, but it still has the analogous semantic relation:

eftirför (n)	elta (v)
chase	*chase*

The basic patterns for the extraction of endonyms are:

(1) *það að* verb
that to
(2) noun adv conj verb
(3) *það að vera* adj(, adj)*
that to be
(4) *e-ð* adj.*
sth.

3.6 Attributes

The extracted attributes do not correspond to attribute slots like SIZE or COLOR; they are in fact attribute values, like *big* or *red.* These provide valuable semantic information and can be used as a basis of differentiation between co-hyponyms. Another benefit of the attributes is the possibility of grouping co-hyponyms that have the same attribute, thus allowing extraction of synonymy or near synonymy that would otherwise have been hidden, as shown in table 1.

Lemma	Attribute	Hypernym
skella	*hávær (loud)*	*stúlka (girl)*
glumra	*hávær*	*stúlka*
bjalla	*hávær*	*stúlka*
heimasæta	*ógift (unmarried)*	*stúlka*
yngisstúlka	*ógift*	*stúlka*
ungfrú	*ógift*	*stúlka*

Table 1: lemmata with the same attribute and the same hypernym can be grouped to build a potential synset

The patterns used to extract attributes are the ones starting with an adjective:

(1) adj (adj)? noun.*
(2) adj (,|conj) adj noun.*

3.7 Biological family

Definitions of lemmata from the categories of flora and fauna often include encyclopaedic information additionally to a hypernym and an attribute. From these definitions the name of the biological family of the animate being denoted by the lemma can be extracted:

> **grænlilja** íslensk plöntutegund (*Orthilia secunda*) af vetrarliljuætt
> **...** *an Icelandic plant species (Orthilia secunda) of the wintery lily family*

```
family(grænlilja, vetrarliljuætt)
hypernym(grænlilja, plöntutegund)
attribute(grænlilja, íslensk)
```

4 Evaluation

The analysis tool is called “MerkOr”, from Icelandic *merking* (‘meaning’) and *orð* (‘word’). The results of MerkOr’s analysis of ÍO include 116,446 semantic relations between a lemma and a word included in its definition, with equivalence as the most frequent relation extracted. Table 2 shows the extracted relations, ordered by frequency:

Relation	**Number extracted**
Equivalence	51,390
Hypernym	43,066
Attribute	12,771
Biological family	2,817
Reference	2,140
Endonym - verb	1,286
Synonym	1,201
Meronym	731
Endonym - adj	662
Holonym	382
TOTAL	**116,446**

Table 2: Extracted relations by frequency

Note, however, that the equivalence and reference relations were extracted by the item number in the ÍO definition and not by pattern matching. If these two relations are excluded then there are eight relations extracted 62,916 times, with hypernyms being the largest group.

The first results of MerkOr are promising. First, from a high percentage of the definitions, at least one semantic relation was extracted. Table 3 shows this data.

	Total	Relation extracted
Definitions	77,348	96.45%
Meaning items	106,977	92.61%

Table 3: Analysed definitions and meaning items

A random selection of 1,034 definitions (about 1.34% of the total) was manually analysed as a gold standard against which the MerkOr results could be tested. The evaluation was run with respect to whole definitions rather than individual meaning items, as this information was thought to be more useful for dictionary makers. MerkOr extracted semantic relations from 957 of the definitions in the gold standard (92.55% extraction rate). The evaluation measures are defined as follows: *correct* indicates that all possible semantic relations are identified and no impossible relations are identified; *partly correct* indicates that some but not all possible relations are identified and also that no impossible relations are identified; *false* indicates that at least one impossible relation is indentified. Table 4 shows the test results.

correct	82.13%
partly correct	12.64%
false	5.22%
correct + partly correct	94.77%

Table 4: Accuracy of MerkOr for definitions

All in all 94.77% of the analysed definitions were correctly or partly correctly analysed. This is an encouraging result; the next question, however, is whether this methodology can be extended to free text.

5 Grant of Excellence – a database of semantic relations

Given the limited resources (people and money) in a small language community like Iceland, it is essential to develop LT modules in efficient ways. This is especially true for an extensive project like the development of a semantic database. Existing hand-built resources such as WordNet have been decades in the making; for Icelandic there is little alternative but to adopt and adapt more automated methodologies, such as those outlined above.

Building on these results, we aim to develop methods for extracting semantic relations from unstructured Icelandic texts, using lexico-syntactic patterns. As in the ÍO-project, we will strive for the extraction of both lexical and encyclopaedic relations. Such work requires vast amounts of tagged text and just such resources are being developed at the Árni Magnússon Institute for Icelandic Studies (Helgadóttir 2004).

The central pattern-based methodology will be extended with other techniques such as latent semantic analysis and coordination information (cf. Cederberg and Widdows 2003, Snow et al. 2005) and tested against established statistical methods for automatic thesaurus construction (cf. Grossman and Frieder 2004).

A tool with a graphical user interface will be developed that allows for manual corrections and extensions of the automatically extracted relations.

6 Conclusions and future work

Nikulásdóttir (2007) shows that automatic extraction of semantic relations from a monolingual dictionary works well for Icelandic. The challenge is to extend this work and test the feasibility of applying a similar approach to free text from a tagged corpus of Icelandic. This will be the task undertaken as part of VLT, the recently-awarded Grant of Excellence. We hope that this work will lay the foundation for the development of a WordNet-like resource for Icelandic.

References

Agirre, Eneko et al. (2000): Extraction of Semantic Relations from a Basque Monolingual Dictionary using Constraint Grammar. In: *CoRR* (cs.CL/0010025)

Alshawi, Hiyan (1987): Processing Dictionary Definitions with Phraseal Pattern Hierarchies. In: *Computational Linguistics* Vol. 13, Nr. 3-4, pp. 195-202.

Brants, Thorsten (2000): TnT – A Statistical Part-of-Speech Tagger. In: *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*, pp. 224-231, Seattle, WA.

Cederberg, Scott and Dominic Widdows (2003): Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction. In: *Proceedings of the International Conference on Natural Language Learning (CoNLL)*, pp. 111-118.

Chodorow, Martin; Roy J. Bird and George E. Heidorn (1985): Extracting Semantic Hierarchies from a Large-On-line Dictionary. In: *Proceedings of the 23rd Annual Meeting of the ACL*, pp. 299-304.

Cruse, Alan (1986): *Lexical Semantics*. Cambridge: Cambridge University Press.

Fellbaum, Christine (1998): *WordNet. An Electronic Lexical Database*. Cambridge Mass., London: MIT Press.

Geeraerts, Dirk (2003): Meaning and Definition. In: Piet van Sterkenburg (ed.): *A Practical Guide to Lexicography*. Amsterdam, Philadelphia: John Benjamins, pp. 83-93.

Grossman, David A. and Ophir Frieder (2004*): Information Retrieval: Algorithms and Heuristics*. Second Edition. Berlin: Springer.

Hearst, Marti A. (1998). Automated discovery of WordNet relations. In: Christiane Fellbaum (ed): *WordNet. An Electronic Lexical Database*. Cambridge Mass., London: MIT Press, pp.131-151.

Helgadóttir, Sigrún. (2004). Mörkuð íslensk málheild. [A Tagged Icelandic Corpus]. In Samspil tungu og tækni. Ministry of Education, Science and Culture, Reykjavík, Iceland, pp. 67-71.

Íslensk orðabók (2002). Mörður Árnason (ed.). Reykjavík: Edda.

Markowitz, Judith; Thomas Ahlswede and Martha Evens (1986): Semantically Significant Patterns in Dictionary Definitions. In: *Proceedings of the 24th Annual Meeting of the ACL*, pp. 112-119.

Nakamura, Junichi and Makoto Nagao (1988): Extraction of Semantic Information from an Ordinary English Dictionary and its Evaluation. *Proceedings of the Twelfth International Conference on Computational Linguistics*, pp. 459-464.

Nikulásdóttir, Anna Björk (2007): *Automatische Extrahierung von semantischen Relationen aus einem einsprachigen isländischen Wörterbuch*. M.A.-Thesis, University of Heidelberg, Germany.

Rögnvaldsson, Eiríkur (2008). Icelandic Language Technology Ten Years Later. In Collaboration: *Interoperability between People in the Creation of Language Resources for Less-resourced Languages*, pp. 1-5. SALTMIL workshop, LREC 2008. Marrakech.

Snow, Rion; Daniel Jurafsky and Andrew Y. Ng (2006): Semantic Taxonomy Induction from heterogenous Evidence. In: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pp. 801-808, Sydney.

Wiegand, Herbert Ernst (1989): Die lexicographische Definition im allgemeinen einsprachigen Wörterbuch. In: Franz Josef Hausmann et al. (eds.): *International Encyclopedia of Lexicography:* 003. Berlin, New York. (Handbooks of Linguistics and Communication Science 5.1) pp. 530-588.

All in the Family: A Comparison of SALDO and WordNet

Lars Borin and **Markus Forsberg**
Språkbanken, University of Gothenburg, Sweden
lars.borin@svenska.gu.se, markus.forsberg@gu.se

Abstract

SALDO is a free full-scale modern Swedish semantic and morphological lexical resource intended primarily for use in language technology applications. In this paper we present our work on SALDO, compare it with some other lexical-semantic resources – Wierzbicka's Natural Semantic Metalanguage, Princeton WordNet, and Roget-style thesauruses – and discuss some implications of the differences.

1 Introduction

SALDO, or SAL version 2, is a free modern Swedish semantic and morphological lexicon. The lexicon is available under Creative Commons Attribute-Share Alike license and LGPL 3.0.

SALDO started as *Svenskt associationslexikon* (Lönngren, 1992) – 'The Swedish Associative Thesaurus', a so far relatively unknown Swedish thesaurus with an unusual semantic organization. SAL has been published in paper form in two reports, from the Center for Computational Linguistics (Lönngren, 1998), and the Department of Linguistics (Lönngren, 1992), both at Uppsala University. Additionally, the headwords and their basic semantic characterizations have been available electronically, in the form of text files, from the very beginning.

The history of SAL has been documented by Lönngren (1989) and Borin (2005). Initially, text corpora were used as sources of the vocabulary which went into SAL, e.g., a Swedish textbook for foreigners and a corpus of popular-scientific articles. A small encyclopedia and some other sources provided the large number (over 3000) of proper nouns found in SAL. Eventually, a list of the headwords from *Svensk ordbok* (SO, 1986) was acquired from the NLP and Lexicology Unit at the University of Gothenburg, and the second paper edition of SAL (Lönngren, 1992) contained 71,750 entries. At the time of writing, SALDO contains 76,750 entries, the increased number being because many new words have been added, but also because a number of entries belong to more than one part of speech or more than one inflectional pattern (there are currently 73,909 distinct semantic units in SALDO).

The work described here first started in late 2003, when Lars Borin and Lennart Lönngren initiated a collaboration aiming at making SAL available for online browsing through Språkbanken (the Swedish Language Bank at the University of Gothenburg). In 2005, a computational linguistics student made a prototype graphical interface to SAL (SLV – Språkbanken Lexicon Visualization; Cabrera 2005). Using this interface, Lennart Lönngren was able to revise a considerable number of entries with respect to their semantic characterization, so that SALDO is in this respect no doubt a new edition of SAL, i.e., also as a semantic lexicon.

We soon realized, however, that in order to be really useful in language technology applications, SAL would have to be provided at least with inflectional morphological information. Thus the work on SALDO began.

2 SALDO: a Semantic Lexicon

As a semantic lexicon, SALDO is a kind of lexical-semantic network, superficially similar to WordNet (Fellbaum, 1998), but quite different from it in the principles by which it is structured.

The organizational principles of SALDO consist of two primitive semantic relations, one of which is obligatory and the other optional. Every entry in SALDO must have a *mother* (or *main descriptor*), a semantically closely related entry which is more central, i.e., semantically and/or morphologically less complex, probably more frequent, stylistically more unmarked and acquired earlier in first and second language acquisition, etc.[1] The mother will in practice often be either a hyperonym (superordinate concept) or synonym of the headword. However, it need not be either: Sometimes it is an antonym (opposite concept), and quite often it is a different part of speech from the headword, which takes us outside the realm of traditional lexical-semantic relations.

In order to make SALDO into a single hierarchy, an artificial most central entry, called PRIM, is used as the mother of 51 semantically unrelated entries at the top of the hierarchy, making all of SALDO into a single rooted tree. These 51 entries, which may be viewed as the semantic primitives of SALDO, are listed in figure 1, approximately translated.

The tree of SALDO roughly captures the notion of centrality by the 'depth' – the distance down from the PRIM root node – of an entry: the deeper an entry lies in the tree, the less central it is. The average depth of SALDO is 5.74 and the median depth is 6. The (single) deepest entry – *tjuvpojksglimt* 'rascal gleam' – is found at depth 15.

SALDO is a monolingual dictionary; it aspires to capture associative relations among the concepts of only one language, namely Swedish. Any claim to universality in SALDO must lie in the two basic relations, whereas the nodes connected by these relations are pre-existing, given by the lexical system of the particular language being described. Against this background, it is an instructive exercise to compare the topmost lexemes in SALDO – its 51 semantic primitives – with the semantic primitives of Wierzbicka and Goddard's *Natural Semantic Metalanguage* (NSM; Wierzbicka 1996; Goddard 2008), i.e., a semantic formalism with explicit claims to universality.

The NSM set of semantic primitives has undergone many revisions through the years. In figure 2 we reproduce the *Proposed semantic primes (2007)* from the NSM homepage <http://www.une.edu.au/bcss/linguistics/nsm/>. We find that the Swedish counterparts of the NSM primitives (Goddard and Karlsson, 2008) are generally found close to the top node in SALDO. Their depth in SALDO is indicated by numbers in parentheses in figure 2 (where a depth of one means a primitive concept in SALDO). It would be interesting to look closer into the differences between the two sets and their possible explanations, but considerations of space preclude any but the briefest remarks here. E.g., we note that in some cases, MSN treats as equally fundamental some concepts which in SALDO are related by the mother-child relation, and consequently one member in SALDO is seen as more central than the other(s): *bra* 'good' (depth 1) – *dålig* 'bad' (2); *mycken* 'much' (1) – *stor* 'big' (2) – *liten* 'small' (3).

all 'all'	*annan* 'other'	*använda* 'use'
att 'that'	*bara* 'only'	*bra* 'good'
genom 'through'	*den* 'it'	*fort* 'fast'
framme 'arrived'	*färg* 'color'	*för*^{2} 'for'
förbi 'gone/past'	*före* 'before'	*en*^{2} 'a/one'
göra 'do'	*ha* 'have'	*hur* 'how'
hända 'happen'	*i*^{2} 'in'	*ja* 'yes'
just 'just'	*kunna* 'be able'	*ljud* 'sound'
ljus 'light'	*med* 'with'	*men* 'but'
mycken 'much'	*måste* 'must'	*namn* 'name'
natur 'nature'	*när* 'when'	*och* 'and'
om 'if'	*om*^{2} 'about'	*på* 'on'
rak 'straight'	*röra* 'move'	*säga* 'say'
tal 'speech'	*till* 'to'	*tänka* 'think'
vad 'what'	*var* 'where'	*vara* 'be'
varm 'warm'	*vem* 'who'	*veta* 'know'
vid 'by'	*vilja* 'want'	*öppen* 'open'

Figure 1: SALDO's 51 semantic primitives

Some SALDO entries have in addition to the mother an optional *father* (or *determinative descriptor*), which is sometimes used to differentiate lexemes having the same mother.

SALDO is unusual in several respects:

- it contains a number of proper nouns and multi-word units, not normally found in conventional print or electronic dictionaries;

- it is strictly semantic in its organization; all entries are *lexemes*, i.e., semantic units; homonymous entries representing more than one part of speech are often treated as different, but always because of their semantics and never for inflectional reasons;

- the organizational principles of SALDO are different from those of lexical-semantic networks such as WordNet, in that the semantic relations are more loosely characterized in SALDO. They also differ from those of more conventional thesauruses, however, but in this case by having more, as well as more structured, sense relations among lexemes.

[1] Both the mother and the father (see below) relations are unique to SAL(DO); they were defined explicitly for this novel kind of lexical-semantically organized dictionary.

substantives: I (2), you (2), someone (2), people (3), something/thing (2), body (3);
relational substantives: kind (3), part (3);
determiners: this (3), the same (3), other/else (1);
quantifiers: one (2), two (2), some (2), all (1), much/many (2);
evaluators: good (1), bad (2);
descriptors: big (2), small (3);
mental predicates: think (1), know (1), want (1), feel (2), see (2), hear (2);
speech: say (1), words (4), true (3);
actions, events, movement, contact: do (1), happen (1), move (2), touch (2);
location, existence, possession, specification: be (somewhere) (1), there is (2), have (1), be (someone/something) (1);
life and death: live (2), die (3);
time: when/time (1), now (2), before (1), after (2), a long time (4), a short time (3), for some time (3), moment (4);
space: where/place (1), here (2), above (2), below (3), far (6), near (2), side (2), inside (2);
'logical' concepts: not (1), maybe (3), can (1), because (2), if (1);
intensifier, augmentor: very (2), more (2);
similarity: like (5).

Figure 2: NSM's 61 semantic primitives (depth in SALDO in parentheses)

Below, we give a few examples of entries with their mother and father lexemes, randomly selected under the letter "B":

balkong : hus ('balcony' : 'house')
bröd : mat + mjöl ('bread' : 'food' + 'flour')
brödföda : uppehälle ('daily bread' : 'subsistence')
bröllop : gifta sig ('wedding' : 'get married')
Bulgakov : författare + rysk ('Bulgakov' : 'author' + 'Russian')

It should be clear from these examples that the basic associative relations in SALDO are not intended as *definitions*, but as loose – but hopefully accurate and useful – semantic characterizations of lexical entries. On the other hand, they seek to characterize entries by (intrinsic) lexical-semantic associations, rather than by the (extrinsic) syntagmatic associations typically elicited in psychological and psycholinguistic word-association experiments (Lönngren, 1998). Like other forms of linguistic analysis, defining lexical entries using the SALDO relations is a skill which requires highly qualified linguistic training and a fair amount of practice for its mastery.

How SALDO is different from typical thesauruses becomes apparent when we consider that the two primitive lexical-semantic relations (*mother* and *father*) can form the basis of any number of derived relations, referred to below as *assets* (associative sets). Thus the m-sibling asset, lexemes having a common mother, is very interesting, as such sibling groups tend to correspond to natural semantic groupings. In this respect, SALDO's lexical families – made up by basic and derived relations – define a thesaurus-like structure, but one which is arrived at inductively, by the bottom-up process of assigning mothers to all lexical items, rather than deductively, by prespecifying by fiat a number of basic concepts under which all lexical items are then grouped, as in *Roget's thesaurus* (with 1000 pre-specified concepts) and its successors.

3 SALDO: a Morphological Lexicon

SAL did not contain any formal information about entries, not even an indication of part of speech. Thus, one important difference between SALDO and SAL is that SALDO now has full information about the part of speech and inflectional pattern of each entry.

The computational morphology of SALDO has been defined with the tool Functional Morphology (FM; Forsberg 2007), a tool that uses the typed functional programming language Haskell (Jones, 2003) as the description language and supports (compound) analysis, synthesis and compilation to a large number of other formats, including full form lists, paradigm tables, XML, XFST (Beesley and Karttunen, 2003), and GF (Ranta, 2004).

The starting point of SALDO's morphology was an FM implementation of modern Swedish developed by Markus Forsberg and Aarne Ranta at Chalmers University of Technology, which consists of an inflection engine covering the closed word classes and the most frequent paradigms in the open word classes. All in all, disregarding the noun compound forms that were not addressed properly, the existing implementation covered, roughly estimated, about 80% of the headwords of SALDO, but only less than a tenth of the inflectional patterns, or paradigms.

Many of the remaining paradigms are quite small. In essence, these are (1) the irregular words of traditional grammar and (2) paradigms with missing slots or more than one word form filling a particular slot.

Something which adds to the number of inflectional patterns is that we also encode some inherent features of words in the inflectional pattern designators, features which do not bear directly on the inflectional behavior of the word itself. However, they are potentially useful and comparatively easy to add simultaneously with the morphological information proper, but can be quite difficult to add later, e.g., the gender of nouns, agreement and anaphorical gender in proper names, etc.

In adding the morphological information to SALDO, we have used existing grammatical descriptions of Swedish inflectional morphology – above all *Svenska Akademiens grammatik* (Teleman et al., 1999), as well as the inflectional information provided in existing Swedish dictionaries, primarily *Nationalencyklopedins ordbok* (NEO, 1995), but also its predecessor *Svensk ordbok* (SO, 1986), and *Svenska Akademiens ordlista* (SAOL, 2006), plus empirically evidenced usage in corpora and on the internet.

4 SALDO in Comparison with WordNet

Princeton WordNet is built up from words in the open word classes, i.e., nouns, verbs, adjectives, and adverbs,[2] and a set of relations. The most important relation is the equivalence relation *synonymy* that defines the *synsets* (synonymy sets, sets of words that are interchangeable in some context). The other relations are over synsets: *antonymy*, *hyponymy*, *hyperonymy* (often called “hypernymy” in the WordNet literature), *meronymy*, *holonymy*, *troponymy*, and *entailment*. These relations are *typed* in the sense that they are only valid for a subset of the word classes.

SALDO, on the other hand, is concerned with all words, even the closed word classes such as prepositions and pronouns. The relations are more loosely defined through the untyped *mother* and *father* relations, but the resulting structure is strictly hierarchical and noncyclic.

The synsets of WordNet are the result of deliberate choices, and tend to be fairly small, whereas SALDO’s counterparts, the assets, are semantic groups that emerge gradually as the result of many individual decisions (although an examination of an asset may result in a change of the description), and which vary widely in size.

A concrete example is a comparison of the synsets of Princeton WordNet and the m-sibling asset of SALDO for an arbitrarily picked word: *sun* (and the Swedish counterpart *sol*).

Starting with Princeton WordNet <http://wordnetweb.princeton.edu/perl/webwn>, where we only consider the noun synsets, not the verbal ones, since the Swedish word *sol* has no verbal interpretation. Note that the synset memberships (the boldfaced items) are small, singleton sets in several cases.

sun, Sun (the star that is the source of light and heat for the planets in the solar system) “the sun contains 99.85% of the mass in the solar system”; “the Earth revolves around the Sun”;

sunlight, sunshine, sun (the rays of the sun) “the shingles were weathered by the sun and wind”;

sun (a person considered as a source of warmth or energy or glory etc);

sun (any star around which a planetary system revolves);

Sunday, Lord’s Day, Dominicus, Sun (first day of the week; observed as a day of rest and worship by most Christians)

If we now have a look at SALDO’s m-sibling asset for the lexeme *sol* ‘sun’ (there is one lexeme *sol* in SALDO), that is, the lexemes that share the same mother as *sol* (the verb *lysa* ‘shine’), we get the following asset. Here we have translated and grouped the lexemes into word classes for the sake of presentation, although, as mentioned already, no part-of-speech distinctions are made in SALDO.

verbs: *inform, sparkle, shine, twinkle, shimmer, lustre, flash, glitter, glimmer, glisten, gleam, flimmer, blink, illuminate*;

nouns: *light, star, moon, lantern, lamp, comet, flash, candle, light house*;

adjectives: *shining, fluorescent, light/bright*.

The lexeme *sol* is also related to a father, namely *himmel* ‘sky/heaven’. We continue by examining the full-sibling asset, i.e., those lexemes with *lysa* as mother and *himmel* as father, which is, of course, a subset of the m-sibling asset of *sol*.

nouns: *comet, moon, star*

Looking at the two examples it becomes clear that they are quite different. WordNet gives us its conception of a standard lexical semantic relation, synonymy, but SALDO gives us something else – associations rather than definitions. The sibling assets are clearly semantically related to the lexeme *sol*, but it reminds us about something we

[2] Numerals – cardinals and ordinals – are also included in Princeton WordNet, but labeled as nouns and adjectives (both cardinals and ordinals normally have both noun and adjective readings in WordNet).

might get if we asked a person to list words that they associate with the word *sun*. SALDO's assets are somewhat like Roget-style thesaurus entries, but smaller,[3] without the explicit separation usually made in thesuaruses of parts of speech, and of course including all parts of speech in the lexicon (there are currently 44 different parts of speech used in SALDO). SALDO occupies a position somewhere in between a Roget-style thesaurus and a Princeton-style wordnet in the family tree of lexical-semantically organized lexical resources.

5 Discussion

There is extensive empirical evidence in the literature for the usefulness of the Princeton WordNet,[4] but what about SALDO?

We have yet to perform any significant computational experiment, but we have a couple of ideas about in what kind of language technology applications SALDO may be useful.

SALDO could be useful component in computerized tools for *second language acquisition* of Swedish, since it is structured according to the *centrality principle*: going upwards in the semantic tree should give valid information for a second language learner. Also, the assets may provide semantic nuances that are not easily captured with a textbook definition.

We have also discussed whether SALDO could be used in a writing tool, where the associative links would help writers find appropriate ways of phrasing information content in varying ways in order to make the text livelier or to cater to different readerships.

Semantic information retrieval with different assets may provide interesting aspects on the data at hand. What these aspects could be are still open research questions. For example, what conclusions may we draw from the fact that a particular asset of a search word is populated or not?[5]

Finally, and a bit more far-fetched, but interesting idea, is *metaphor resolution*. A metaphor is a linguistic expression used to represent something else, and for a metaphor to be interpretable, it must be associatively related to what it represents. This is where SALDO comes into the picture: SALDO may potentially be able to generate a set of resolution candidates for a given metaphor.

6 Final Remarks

SALDO may be downloaded from its homepage <http://spraakbanken.gu.se/sal/eng>, where both the released versions and the development version may be accessed.

SALDO is also distributed through four web services: *an incremental fullform lookup service*, *an inflection engine service*, *a compound analysis service*, and an *experimental semantic visualizer*. The first three web services interface to the morphological component, and the last one generates static images of a lexeme's mother, its father, and its m-sibling asset. The web services are updated daily with the latest development version of SALDO.

A future plan is to augment and/or annotate SALDO with WordNet-like relations, such as hyperonymy, hyponymy, and antonymy. Furthermore, we intend to include the SynLex (Kann and Rosell, 2005), also referred to as "the people's synonym lexicon", an interesting free semantic resource, which has been created by asking voluntary users of an English-Swedish dictionary lookup service on the internet to judge the degree of synonymy between word pairs. With SynLex entries connected to SALDO senses (since SynLex provides only headwords), we could use the synonymy degree information at arbitrary cutoff points to create virtual "fuzzy wordnets" for Swedish. With the kind of degree-of-synonymy information present in SynLex – only about 5% of the word pairs in SynLex have the highest degree of synonymy, 5.0 – we could create a wordnet-like lexical resource where we can exactly quantify the 'near-synonymy' that is sometimes said to define WordNet synsets. This would partly address an oft-heard criticism of the WordNet concept, a criticism which hinges on a postulated universal linguistic principle of (full) *synonymy avoidance* (Carstairs-McCarthy, 1999). This being an intrinsic characteristic of human language – so the reasoning goes – a dictionary whose fundamental or-

[3] There is no main heading for *sun* in Roget 1911 <http://humanities.uchicago.edu/orgs/ARTFL/forms_unrest/ROGET.html>. Instead, the word is found under a number of headings, including *382. Heat, 420. Light* and *423. [Source of light, self-luminous body.] Luminary*, each containing a few tens of words or multi-word expressions.

[4] This is undoubtedly in no small part due to the Princeton WordNet being a completely free resource, as well as an English resource; cf. the contrasting case of the EuroWordNet.

[5] In fact, the original SAL project was initiated with information retrieval and automatic text indexing applications in mind (Lönngren, 1998).

ganization is based on the notion of (even near-) synonymy almost by definition cannot present a faithful reflection of our lexical knowledge, at least not from a linguistic point of view.

Acknowledgments

We would like to express our gratitude to Lennart Lönngren for creating SAL version 1 and for continuously being available as a sounding board for our ideas.

Thanks to two anonymous reviewers for spotting some unclarities in our text.

SALDO has been developed with Swedish public funding. After 2003, the University of Gothenburg through Språkbanken has financed the main part of the work on SALDO.

During 2006–2008, SALDO has been partly supported by the VR project *Library-Based Grammar Engineering* (2005-4211; PI Aarne Ranta, Chalmers University of Technology).

Since 2008, part the funding comes from the VR/DISC project *Safeguarding the future of Språkbanken* (2007-7430; PI Lars Borin, Språkbanken, University of Gothenburg).

References

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Publications, Stanford.

Lars Borin. 2005. Mannen är faderns mormor: *Svenskt associationslexikon* reinkarnerat. *LexicoNordica*, 12:39–54.

Isabelle Cabrera. 2005. Språkbanken lexicon visualization. Rapport de stage. Projet réalisé au Département de Langue Suédoise, Université de Göteborg, Suède.

Andrew Carstairs-McCarthy. 1999. *The Origins of Complex Language*. Oxford University Press, Oxford.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Mass.

Markus Forsberg. 2007. *Three Tools for Language Processing: BNF Converter, Functional Morphology, and Extract*. Ph.D. thesis, Göteborg University and Chalmers University of Technology.

Cliff Goddard and Susanna Karlsson. 2008. Rethinking *think* in contrastive perspective: Swedish vs. English. In Cliff Goddard, editor, *Cross-Linguistic Semantics*, pages 225–240. John Benjamins, Amsterdam.

Cliff Goddard, editor. 2008. *Cross-Linguistic Semantics*. John Benjamins, Amsterdam.

Simon P. Jones. 2003. *Haskell 98 Language and Libraries: The Revised Report*. Cambridge University Press, Cambridge, May.

Viggo Kann and Magnus Rosell. 2005. Free construction of a swedish dictionary of synonyms. In *NoDaLiDa 2005*, Joensuu.

Lennart Lönngren. 1989. *Svenskt associationslexikon: Rapport från ett projekt inom datorstödd lexikografi*. Centrum för datorlingvistik. Uppsala universitet. Rapport UCDL-R-89-1.

Lennart Lönngren. 1992. *Svenskt associationslexikon. Del I-IV*. Institutionen för lingvistik. Uppsala universitet.

Lennart Lönngren. 1998. A Swedish associative thesaurus. In *Euralex '98 proceedings, Vol. 2*, pages 467–474.

NEO. 1995. *Nationalencyklopedins ordbok*. Bra Böcker, Höganäs.

A. Ranta. 2004. Grammatical Framework: A type-theoretical grammar formalism. *The Journal of Functional Programming*, 14(2):145–189.

SAOL. 2006. *Svenska Akademiens ordlista över svenska språket*. Norstedts Akademiska Förlag, Stockholm.

SO. 1986. *Svensk ordbok*. Esselte Studium, Stockholm.

Ulf Teleman, Staffan Hellberg, and Erik Andersson. 1999. *Svenska Akademiens grammatik, 1–4*. NorstedtsOrdbok, Stockholm.

Anna Wierzbicka. 1996. *Semantics: Primes and Universals*. Oxford University Press, USA.

NorNet - a monolingual wordnet of modern Norwegian

Ruth Vatvedt Fjeld
Universitetet i Oslo
Norway
r.e.v.fjeld@iln.uio.no

Lars Nygaard
Kaldera Language Technology
Oslo, Norway
ln@kaldera.no

Abstract

NorNet is an attempt to derive a wordnet automatically from a traditional dictionary for Norwegian Bokmål by means of some simple rules for extracting information from its definitions. Only synonymy and hyponymy are investigated, and in this first version of NorNet approximately 80 000 lexical relations are described and all nouns in the dictionary are thereby ordered in sets. The method chosen seems to work well and will be used in further refining the wordnet and also include verbs and adjectives.

1 Introduction

A wordnet is an onomasiological dictionary where the main goal is to link words together in semantic fields based on semantic relations. Thesauruses, of which the best known is Roget (1852), are the traditional precursors to wordnets. The lexicographer Ivar Aasens made the first attempt of an Norwegian thesaurus with Norsk maalbunad, printed post mortem in 1925. Aasen thought of this thesaurus as his main work.

The first modern semantic database was the Princeton Wordnet[1]. The EuroWordNet project[2] implemented similar databases for several European languages. In the Nordic countries, DanNet[3] and SwordNet[4] the Swedish part of EuroWordNet, are the most elaborated data Apart from a preliminary version of the SIMPLE-lexicon (Lenci et.al., 2000), there has not been any attempts so far to build wordnets manually for Norwegian, but there has been made some attempts to generate wordnets automatically.

Dyvik (2002) generated a thesaurus from an English-Norwegian parallel-corpus by means of the so-called mirror method. The method uses translational correspondences from a parallel corpus to distinguish word senses and infer semantic relations.

Nygaard (2006) compiles sets of partially disambiguated lexical relations based on an automatic analysis of Bokmålsordboka, a traditional standard monolingual dictionary (Wangensteen, 2005).

2 NorNet

The aim of the NorNet project was to create a wordnet for Norweigan. The method chosen was to start with the lexical relations produced by the system described in Nygaard (2006), map out the hyperonyms and the synonyms of the lemmas, manually review the results and resolve remaining ambiguity; thus creating a full wordnet. The material has a very good coverage of the lexicon, since it is based on a traditional dictionary. In addidion, the error rate is fairly low (about 3 per cent). This method made it possible to create an extensive wordnet with a fairly small budget.

An advantage of using a monolingual dictionary as the basis for NorNet, is that the output is a model of the internal, semantic structure of the

[1] http://wordnet.princeton.edu/
[2] http://www.illc.uva.nl/EuroWordNet
[3] http://wordnet.dk
[4] http://www.ling.lu.se/projects/Swordnet

dictionary. This provides lexicographers with a tool for identifying inconsistencies and omissions in the dictionary. In particular, a large number of circular definitions have been identified.

NorNet now consists of a large set of lexical relations, approximately 80 000. For the time being, NorNet only contains nouns. The addition of adjectives and verbs is currently being investigated.

3 Method

The study of lexical relations have been given much attention in modern lexicology. Following Vossen (1998) who states that our general knowledge of semantic relations are too complex to be adequately described jet, we have chosen the relations most used in traditional dictionaries: synonymy and hyponymy.

These lexical relations are used as a basis for NorNet, and they were produced through a rather simple procedure, using the quite predictable structure of dictionary entries.

3.1 Analysis of definitions

The definitions in the dictionary were part-of-speech-tagged, and relations were extracted using a simple rule:

> if the definition consists of a single noun, or a comma-separated list of single nouns, then those nouns are synonyms to the defined word. If the definition consists of a modified noun, then the first noun in the definition is the hyperonym of the defined word.

The following are the definitions of "ananas" (pineapple) in *Bokmålsordboka:*
1. plante av slekten Ananas i ananasfamilien (plant of the genus Ananas, in the Ananas family)
2. frukt av ananas (fruit of ananas)
From these definitions, the program infers that

- sense 1 of "ananas" is a hyponym of "plante" (plant)
- sense 2 of "ananas" is a hyponym of "frukt" (fruit)

The definition of "anakoret" *(anchorite*) is "eneboer, eremitt" (*recluse, hermit*). The program infers that

- "anakoret" is a synonym to "eneboer" (*recluse*)
- "anakoret" is a synonym to "eremitt" (*hermit*)

There are some exceptions to this, due to non-standard definitions, e.g. negative definitions, meronymic or collective definitions and use of meta-language. For example "abessinier" is defined as "eldre betegnelse for etiopier" (*older designation for Ethiopian*). The program would wrongly infer that the hyperonym for "abessinier" is "betegnelse" (*designation),* thus "betegnelse" is added to a stoplist of words that may not be considered as hyperonyms.

3.2 Analysis of compound words

The dictionary also contains fairly extensive information about compounding. This formed the basis of a second set of relations. The word "rødvin" (*red wine)* is segmented as "rød~vin", allowing the program to infer that "vin" (*wine*) is the hypernym of "rødvin".

Of course, there are a number of compounds in Norwegian that are idiosyncratic, i.e. where the head is not the hypernym of the compound. This is typically in metaphorical use of one part of the compounds, as in "tankekors" (*puzzle*, lit. *thought cross*), which obviously is not a kind of cross. However, most of these words are given definitions in the dictionary, and the program allows relations from the definitions override relations from compounding information.
Additionally, the fact that most idiosyncratic compounds and most non-compound words are listed in the dictionary, makes automatic compound analysis feasible as a method for enriching the wordnet with a large number of compounds.

3.3 Remaining ambiguity

All the relations in NorNet based on the dictionary definitions are *partially disambiguated*: The sense of the lower entry in the hierarchy is known (jf. “ananas” in sect. 3.1), but there is no rule to which sense of the higher part that is to be chosen, e.g. if "ananas" is a hyponym of

- omdannet fruktemne (transformed ovary)
- godt resultat (good result)
- resultat (result)
- avkastning (earnings)
- produkt (product)
- følge (consequence)
- utbytte (yield)

Before this remaining ambiguity is resolved, the relations cannot be used to build a full wordnet. Consider the definition of “kommunist” (communist): “tilhenger av kommunisme” (supporter of communism). The word “tilhenger” is polysemous in Norwegian; it can either mean “supporter” or “trailer” (e.g. of a car or truck). Even if we at this stage correctly infer that a communist is a kind of “tilhenger”, we do not know if it is in the sense of “supporter” or “trailer”.

Because of the low precision of current efforts in automatic word sense disambiguation, and since a manual review of the material was judged to be necessary anyway, this ambiguity resolution was done manually in the NorNet project.

3.4 Manual review

In addition to disambiguation, the review process uncovered a wide variety of errors in the material.

The most frequent type of errors were caused by the analysis program itself, either by mistakes in the part-of-speech tagging, omissions in the exception lists or technical errors. A typical example is when listing all the hyperonyms under "person", the noun "pose" (flaunting person) occurs. But this word also means "bag, sack" in Norwegian, and consequently a large amount of hyponyms for "bag" are included under "person". These were to be sorted out by hand.

In addition, through the review, a number of mistakes in the dictionary itself were discovered, such as missing senses and missing words, inconsistent definitions, unsystematic co-hyponymy. For example, “apologet” (*apologist*) is defined as “forsvarer, særlig av kristendommen” (lit. *defender, in particular of Christianity*). However, the entry for “forsvarer” (*defender*) lacks this sense of the word (only containing the legal and sports-related senses).

4 Conclusion

NorNet reflects both the strengths and weaknesses of the traditional human-oriented dictionary. Dictionaries have traditionally been edited using an alphabetically structured word list. This ordering is, of course, completely arbitrary, and consequently there is a risk of inconsistency and incomplete description of the lexicon.

On the other hand, a traditional lexicon is just the result of a long tradition, often developed through several years with many editors. In new editions, mistakes have been corrected, lakunaes filled and new word senses has been added. As years go by, the traditionally made dictionaries are quite good, in spite of lacking methodology.

Using the method described in this paper, new lexical resources can make the best possible use of this knowledge and this tradition, while creating a tool for correcting the inconsistencies and omissions that occur.

References

Helge Dyvik. 2002. Translations as semantic mirrors: from parallel corpus to wordnet. In Karin Aijmer and Bengt Altenberg (ed.): *Papers from the 23rd International Conference on English Language Research on Computerized Corpora*, Göteborg.

Allesandro Lenci, Nuria Bel, Fredrica Busa, Nicoletta Calzolari, Elisabetta Gola, Monica Monachini, Antoine Ogonowski, Ivonne Peters, Wim Peters, Nilda Ruimy, Marta Villegas and Antonio Zampolli. 2000. Simple: A General Framework for the Development of Multilingual lexicons. In *International Journal of Lexicography* 13(4):249-263.

Lars Nygaard.2006. *Frå ordbok til ordnett.* Cand. Philol.-thesis, University of Oslo.

Piek Vossen. 1998. *EuroWordNet. A Multilingual Database with Lexical Semantic Networks.* Amsterdam.

Boye Wangensteen (ed.). 2005. *Bokmålsordboka: definisjons- og rettskrivningsordbok*. Kunnskapsforlaget, Oslo.

Automatic Relation Extraction – Can Synonym Extraction Benefit from Antonym Knowledge?

Anna Lobanova, Jennifer Spenader, Tim van de Cruys,
Tom van der Kleij, Erik Tjong Kim Sang
University of Groningen
{a.lobanova@| j.spenader@| t.van.de.cruys@| a.a.j.van.der.kleij@ai.| e.f.tjong.kim.sang@}rug.nl

Abstract

We use automatically extracted word pairs from one lexical relation to filter out incorrect pairs of another relation. Initial results for improving Dutch synonyms by filtering out antonyms show a small precision improvement.

1 Introduction

Automatic extraction of lexical relations is useful for improving the coverage of existing computational lexical resources. For example, representation of lexical knowledge in WordNet (Fellbaum, 1998) is based on the *synsets*, or sets of synonyms like (*rich, affluent, flush, loaded, moneyed, wealthy*). Being able to extract synonyms automatically would lead to a consistent way of improving and extending the representation of synonyms in wordnets across different languages. However, an important problem of current distribution-based methods of synonym extraction is that they produce noise. As Lin et al. (2003) point out, an automatically obtained list of the top-20 distributionally similar words of *adversary* includes not only synonyms like *opponent* and *antagonist* but also contrasted words like *supporter* and even antonyms like *ally*:

> ***adversary**: enemy, foe, ally, antagonist, opponent, rival, detractor, neighbor, supporter, competitor, partner, trading partner, accuser, terrorist, critic, Republican, advocate, skeptic, challenger* (Lin et al., 2003)

This is due to a similar distribution of antonyms and synonyms in text (Lucerto et al., 2004). Lin et al. (2003) suggested to perform a two-step relation extraction approach in which synonym extraction is followed by a step in which semantically incompatible word pairs are filtered out. A pair of words was considered semantically incompatible if it occurred in the two surface patterns *from X to Y* and *either X or Y*. The results of the combined approach were good but the authors did not evaluate the impact of the second step.

Our assumption is that the extra filtering step is useful for improving the quality of automatic relation extraction, in particular synonyms. The goal of this paper is to validate this assumption. We present an experiment with Dutch synonym extraction in which erroneously extracted antonyms are filtered out in a post-process. We show that the filtering step does indeed improve the quality of the first extraction step.

In the next section, we describe methods we used to extract antonyms and synonyms automatically. In section 3, we show how the results for the two relations can be combined and present the results of this approach, our conclusions are summarized in section 4.

2 Automatic extraction of antonyms and synonyms

In this section we describe our work on automatic extraction of antonyms and synonyms. We used two pattern-based approaches to extract antonyms. The first uses two manually selected text patterns (section 2.1). In the second approach, text patterns indicating an antonym relation were *learned* from a collection of texts using a small set of antonym pairs as seeds (section 2.2). Hearst (1992) was the first to use preselected lexico-syntactic patterns to automatically extract hypernym-hyponym pairs from text. Since then, text patterns have been used to extract different lexico-semantic relations, in most cases hyponyms and meronyms (Berland and Charniak, 1999; Pantel and Pennacchiotti, 2006; Tjong Kim Sang and Hofmann, 2007). Synonym extraction has instead focused on using distributional methods. We present our work on automatic extraction of synonyms in section 2.3.

2.1 Antonyms derived from chosen patterns

For the first experiment we chose two text patterns which we expected to contain antonym pairs frequently:

- *zowel X als Y* (*X as well as Y*), for example *zowel mannen als vrouwen* (*men as well as women*)
- *tussen X en Y* (*between X and Y*), for example *tussen goed en kwaad* (*between good and evil*)

We searched in the Twente News Corpus (300 million words) for these text patterns and selected all lower case nouns X and Y which appeared inside both patterns at least twice. The result was a list of 270 antonym candidates. These candidates have been assessed by five native speakers of Dutch. They had the choice of labeling a word pair either as antonym (e.g. *rich/poor*), synonym (e.g. *rich/wealthy*), co-hyponym (e.g. *cat* and *dog* are co-hyponyms of *animal*) or unrelated (none of the above relations). The results of this assessment can be found in Table 1. Percentages of word pairs which received the same label from all participants can be found in the row Unanimous. Percentages for word pairs that received the same label from three or more participants are displayed in the row Majority. 14 pairs (5%) did not receive a majority label.

	Antonyms	Synon.	Co-hyp.	Unrel.
Majority	34%	0%	54%	7%
Unanimous	22%	0%	16%	0%

Table 1: Human evaluation of the 270 pairs extracted by means of chosen patterns. Word pairs could be classified as antonyms, synonyms, co-hyponyms or unrelated. 14 pairs (5%) did not receive a majority label.

The precision of the two patterns was not very high. 34% of the pairs were labeled as antonyms but 54% were assigned the label co-hyponym. The other two categories occurred rarely.

This approach shows that a small number of text patterns is already useful for extracting candidate antonym pairs. Incorporating more text patterns could lead to finding more good pairs. Manual selection of patterns is time-consuming. It is hard to think of all possible productive patterns. Indeed, some infrequent patterns might still provide a valuable contribution. Learning patterns from text and antonym pair examples is a fast and more objective alternative. We discuss this approach next.

2.2 Antonyms derived from learned patterns

To extract patterns automatically we used two sets of seeds consisting of 6 and 18 well established antonyms. The algorithm we used was based on the approach of Ravichandran and Hovy (2002). All sentences containing one of the seed pairs were extracted from Dutch CLEF corpus (Jijkoun et al., 2003), the antonyms were replaced by a wildcard token, 50 most frequently occurring patterns that contained seed pairs at least twice were used to find all word pairs that co-occurred in the positions of the wildcard tokens in the corpus. Depending on the number of times a pattern contained an already known antonym pair and the total number of times that pattern was found in the corpus, each pattern was given a score. Patterns with a score above the threshold were used to calculate the antonymy score (A_i) for each word pair that occurred in them. This score is the probability that the i-th pair is an actual antonym pair, given how often it occurred with each pattern (C_{i_j}) and the scores of these patterns (S_j):

$$A_i = 1 - \Pi_j (1 - S_j)^{C_{ij}} \qquad (1)$$

Pairs with a score $\geq$0.9 were used as new seeds in the following iteration. The entire process of identifying patterns and using those to extract new antonyms was repeated iteratively six times.

After six cycles, the seed sets of 6 and 18 elements had resulted in lists of respectively 1189 and 1355 antonym pairs. Pairs with a score $\geq$0.6 were checked by the human assessors. In the set of 6 seeds, 9 out of 197 checked pairs were antonyms according to EuroWordNet (5%). In the result set obtained with 18 seeds, 10 out of 172 checked pairs were antonyms according to EuroWordNet (6%). Pairs were then evaluated as antonyms, synonyms, co-hyponyms, or none of the above by five participants. The results are presented in Table 2.

The assessment results are comparable to those for the chosen text patterns in Table 1. The precision scores were around 30% but the number of extracted pairs was smaller (an average of 185 in comparison with the 270 of the chosen patterns). Note that the percentages of antonyms found by the assessors are a lot higher than the percentages in EuroWordNet. The antonym relation in EuroWordNet is incomplete.

	Antonyms	Synon.	Co-hyp.	Unrel.
6 seeds				
Majority	27%	1%	39%	31%
Unanimous	16%	0%	9%	15%
18 seeds				
Majority	33%	0%	35%	28%
Unanimous	20%	0%	9%	15%

Table 2: Human evaluation of the word pairs extracted by means of learned patterns: 197 with 6 seeds and 172 with 18 seeds. Word pairs could be classified as antonyms, synonyms, co-hyponyms or unrelated.

2.3 Automatic Extraction of Synonyms

The automatic extraction of synonyms has been carried out with standard dependency-based distributional similarity measures (Lin, 1998; Van de Cruys, 2006; Padó and Lapata, 2007). For each noun, a vector has been constructed, containing the frequencies of the dependency relations in which the noun appears. For example, a noun like *apple*, has features like red_{adj} and eat_{obj}. Dependency triples have been extracted from the CLEF corpus (Jijkoun et al., 2003). The 10,000 most frequent nouns have been used, together with the 60,000 most frequent dependency features, yielding a frequency matrix of 10K nouns by 60K dependency features. This matrix has been adapted with pointwise mutual information (Church and Hanks, 1990) for weighting purposes. Next, the noun by noun similarity matrix has been calculated using the cosine similarity measure. Finally, for each noun, all nouns that exceed a certain cosine similarity threshold are selected as the noun's candidate synonyms.

3 Using Antonyms in Synonym Extraction

We derived noun synonym candidates with distribution-based methods, cosine similarity and pointwise mutual information, as described in section 2.3. Next, we removed all synonym candidates which did not contain a word that was present in one of the two sets with antonyms derived in the previous experiments (see section 2; for the learned patterns, we used the set derived from 18 seeds). This resulted in two sets with unfiltered synonym candidates (114 and 80 word pairs, respectively) which will be used as baselines.

	cut-off (cosine)					
	.40	.30	.20	.18	.15	.10
Baseline (unfiltered)						
Precision	.008	.025	.053	.045	.036	.014
Recall	.003	.005	.035	.038	.048	.099
$F_{\beta=1}$	.004	.008	.042	.041	.041	.025
Filtered						
Precision	.008	.025	.055	.047	.039	.015
Recall	.003	.005	.035	.038	.048	.099
$F_{\beta=1}$	.004	.008	.042	.042	.043	.026

Table 3: Effects of filtering out antonyms derived with chosen patterns from a set of 114 candidate synonyms: a small positive effect on the low-cut-off sets.

Next, we removed from the synonym lists the candidate pairs that also occurred in the antonym lists. This produced two sets of filtered synonym pairs. We computed precision and recall scores for the filtered and the unfiltered synonym lists by comparing them with the synonyms in the Dutch part of EuroWordNet while using six different threshold values determined by the cosine similarity value of the word pairs. The results can be found in Tables 3 and 4.

When using antonyms derived with learned patterns, filtering out antonyms from a set of candidate synonyms had a large negative effect on the $F_{\beta=1}$ rates of high-cut-off sets (Table 4). The approach worked better with antonyms which had been extracted with chosen text patterns. Here we observed a small positive effect on the $F_{\beta=1}$ rates of low-cut-off sets (Table 3). The difference between the two approaches is surprising, given that the quality of the two sets of antonyms was similar according to human assessors (Tables 1 and 2).

Inspection of the results showed that the performance drop associated with the second set of antonyms was caused by a single synonym being present in the antonym list (see Table 5). If the synonym pair had not been classified as an antonym pair then the results of the second filter would have been similar to the first. This reveals a weakness of using learned patterns for identifying relations. The learner might use low-precision patterns which could be harmful for the quality of the results of the extraction process.

However, even without the incorrect pair in the antonym data, the positive effect would be small. In order to obtain a larger positive effect, we need more antonyms. This means that we either should use more data or use more extraction patterns.

	cut-off (cosine)					
	.40	.30	.20	.18	.15	.10
Baseline (unfiltered)						
Precision	.025	.035	.077	.097	.071	.024
Recall	.017	.025	.053	.091	.120	.174
$F_{\beta=1}$	.020	.029	.063	.094	.090	.042
Filtered						
Precision	.013	.023	.070	.090	.069	.024
Recall	.004	.013	.041	.078	.107	.161
$F_{\beta=1}$	.006	.017	.051	.084	.084	.042

Table 4: Effects of filtering out antonyms derived with learned patterns from a set of 80 candidate synonyms: a large negative effect on the high-cut-off sets.

Antonym extraction was based on a text collection of 300 million words and it is unlikely that we will be able to collect a substantial number of extra text soon. Using more extraction patterns has the risk of generating additional false positives with a negative effect on the quality of antonyms.

4 Concluding remarks

We have described an experiment in which automatically extracted antonyms were used to filter out suspected errors from an automatically derived list of synonyms. We used two different methods for producing the antonyms and found that the ones produced by chosen high-quality text patterns were best suited for this approach. However, we only measured a small increase in the quality of the filtered synonym list in comparison with the unfiltered list.

In order to enlarge the observed positive effect, we need a larger set of antonyms. We have argued that both using more data and finding more extraction patterns will be difficult to achieve. One way to work around this problem is by replacing antonymy with a relation which is more frequent, for example, hypernymy. Future research will have to show if this will lead to improved relation extraction results.

References

Matthew Berland and Eugene Charniak. 1999. Finding parts in very large corpora. In *Proceedings of ACL 99*. Maryland, MD, USA.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information & lexicography. *Computational Linguistics*, 16(1):22–29.

CH	LE	
x		bewondering-afkeer (admiration-resent)
x		export-import (export-import)
x		jongen-meisje (boy-girl)
x	x	man-vrouw (man-woman)
	x	uitvoer-invoer (export-import)
	x	waarde-norm (value-norm)
	x	werkelijkheid-realiteit (reality-reality)
	x	werknemer-ambtenaar (employee-civil servant)
	x	werknemer-werkgever (employee-employer)

Table 5: The antonym pairs which were used for filtering: found by chosen patterns (CH) or by learned patterns (LE). Only one is a synonym: *werkelijkheid* - *realiteit* (both mean *reality*).

Christiane Fellbaum. 1998. *WordNet – An Electronic Lexical Database*. The MIT Press.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of ACL-92*. Newark, Delaware, USA.

Valentin Jijkoun, Gilad Mishne, and Maarten de Rijke. 2003. Preprocessing documents to answer dutch questions. In *Proceedings of BNAIC'03*. Nijmegen, The Netherlands.

Dekang Lin, Shaojun Zhao, Lijuan Qin, and Ming Zhou. 2003. Identifying synonyms among distributionally similar words. In *Proceedings of IJCAI 2003*. Acapulco, Mexico.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of* COLING/ACL *98*, Montreal, Canada.

Cupertino Lucerto, David Pinto, and Héctor Jiménez-Salazar. 2004. An automatic method to identify antonymy relations. In *Workshop on Lexical Resources and the Web for Word Sense Disambiguation*. IBERAMIA 2004, Puebla, Mexico.

Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Patrick Pantel and Marco Pennacchiotti. 2006. Expresso: Leveraging generic patterns for utomatically harvesting semantic relations. In *Proceedings of ACL 2006*. Sydney, Australia.

Deepak Ravichandran and Eduard Hovy. 2002. Learning surface text patterns for a question answering system. In *Proceedings of ACL 2002*. Pittsburg, PA, USA.

Erik Tjong Kim Sang and Katja Hofmann. 2007. Automatic extraction of dutch hypernym-hyponym pairs. In *Proceedings of CLIN-2006*. Leuven, Belgium.

Tim Van de Cruys. 2006. Semantic clustering in dutch. In *Proceedings of CLIN-2005*, pages 17–32. University of Amsterdam, Amsterdam, The Netherlands.

The Semantic Relations of Artifacts in DanNet

Sanni Nimb
Det Danske Sprog- og Litteraturselskab (Society for Danish Language and Literature)
Christians Brygge 1, DK-1219, Copenhagen, Denmark
sn@dsl.dk

Abstract

This paper presents the newly released first version of the Danish wordnet, DanNet, focusing on the lexicon model and on the semantic description of artifacts. Apart from being the necessary resource for computational processing of Danish text material such as automatic indexing and information retrieval, the first version of DanNet also makes it possible to carry out linguistic investigations on parts of the Danish lexicon, due to the large number of well structured and consistent lexical data. One example is an investigation on the hyponymy relation at different levels of conceptual domains in Danish, showing a tendency of far more non-taxonomical sister concepts at the general language level than at the basic and specific levels of the language. Another example is an investigation of the distribution of the manually assigned relations in the synsets having an artifact sense in DanNet, showing that the telic relation 'used_for' describing the purpose of the artifact is by far the most frequently applied relation for this group of words. The paper also discusses the differences between the information found in dictionaries and the information to be included in a wordnet.

1 Introduction

The first version of the Danish wordnet, DanNet, was released in March 2009 as an open-source resource (see http://wordnet.dk). DanNet is the product of a joint project between two institutions, The University of Copenhagen, Center for Language Technology (CST), previously having compiled a pilot version of a computational semantic lexicon for Danish, SIMPLE-DK (Pedersen and Paggio, 2004), and the Society for Danish Language and Literature (DSL) that compiled the Danish dictionary which was used as the basis for the wordnet ((Den Danske Ordbog (henceforth DDO (DDO, 2003-2005)).

The 4 years (2005-2009), resulting in the first version, were funded by the Danish Research Council (3,000,000 DKK). In 2008 an additional 3-year funding of 1,000,000 DKK within the DK-Clarin project ensures that the wordnet will be extended by 25,000 synsets.

The first version of DanNet contains approx. 41,000 synsets (34,000 noun synsets, 6,000 verb synsets and 1,000 adjectival synsets). A synset is a set of synonymous lemmas referring to the same concept. e.g. {lys; stearinlys} (candle), {raritet; sjældenhed}(rarity); {humorist; humørbombe; humørspreder} (humorist) and {hoppe} (to jump). Often a synset contains just one single lemma. 26,458 noun lemmas, 3,094 verb lemmas and 809 adjective lemmas are described in the first version. Many of them are polysemous and we have focused on describing at least the main senses of the lemmas.

All synsets in the first version of DanNet are described with hyponymy relations as well as ontological type such as [Living+Object], [Artifact +Object+Part], [Human+Occupation], [Property] etc. 27,000 of the 41,000 synsets in the first version describe nouns having a concrete sense. Of these, approx 12,000 synsets, those referring to objects or human beings, are fully described with information on meronymy, near synonymy, connotation etc., in the case of humans the typical role of the person (e.g. humorist: entertain) and in the case of artifacts also information on origin (how it was made), purpose (what it is used for) as well as agents and instruments involved in the use of the artifact.

A small subset of the synsets in DanNet is linked to Princeton WordNet, and the aim is that 8,000 have been linked by the end of 2010.

The wordnet was established on purely monolingual grounds, and not, as is the case for many other wordnets, by translating synonym sets from i.e. Princeton WordNet to the language in question, in this case Danish. This method – the so-called merge approach – was chosen due to the fact that a corpus-based dictionary of Danish was completed in 2005 and accessible in a machine-readable version with hyperonymy information explicitly specified for each of the approx. 100,000 sense definitions. First of all, this made it possible to build a Danish wordnet using semi-automatic methods, and we estimate that approx. 50% of the data in DanNet has been semi-automatically produced without further adding of data than what is found in DDO. But not less important, it guaranteed that the senses included in the wordnet were actually frequent in general language texts, as the aim of DanNet was to establish a linguistic resource for computational processing of Danish text material, for example automatic indexing, information retrieval, and automatic sense annotation.

Apart from offering linguistic data to developers within the language technology community, DanNet also makes it possible to carry out a wide range of lexical investigations on the Danish lexicon which have not been possible before, due to the systematic organization of the semantics we find in the definitions in DDO as well as completely new data on certain semantic relations not deducible from DDO.

2 The hyponymy hierarchy in DanNet

The wordnet was semi-automatically built by extracting all the senses in DDO having the same specified hypernym (genus proximum). The compiler of the wordnet then organized the proposed hyponymy hierarchy by either simply accepting the hypernym from DDO (which also involved a disambiguation in the many cases of polysemous genus expressions) or by manually selecting a new, and from a structural point of view more precise, hypernym, e.g. in the cases where the genus proximum in DDO was chosen arbitrarily among several synonymous possibilities, or in the cases where genus expressions referred to concepts on a higher level in the hierarchy than the nearest one from a structural point of view. One example of the latter case is 'budcykel' (carrier cycle used to bring out goods to customers) which has the genus proximum 'cykel' (bicycle) in DDO although the structurally seen nearest hypernym is 'ladcykel' (carrier cycle) – in DanNet it is therefore inserted as hyponym to 'ladcykel' instead (see Figure 1).

More challenging was the task of choosing between often more than one suitable hypernym. In some of these cases the synset has been linked to two hypernyms in DanNet: an offroader is both a kind of car and a kind of motorcycle, and a 'havestol' (outdoor chair) is both a chair and a piece of garden furniture. But in general only one hypernym was selected, i.e. the one with the highest number of relevant semantic relations to be inherited: 'slips' (a tie) is for this reason in DanNet described as a 'beklædningsgenstand' (a piece of garment) although defined as a piece of fabric in DDO.

In order to facilitate the practical use of the wordnet as a resource in formal ontologies, the so-called taxonomical hyponyms defined by the test: X is a kind of Y (Cruse, 2002) have been separated from the hyponyms for which the test does not hold (Pedersen and Sørensen, 2006, Pedersen et al., forthcoming). E.g. for the concept 'bicycle' the different kinds of bicycle (a mountain bike, a racer bike, a carrier cycle) are taxonomical in contrast to those hyponyms not being kinds of bicycles but instead describing a property transversely to the taxonomical group. Some examples are 'herrecykel' (gentleman's bicycle), and 'jernhest' (old bike). While members of the last group, which in DanNet are considered to be 'orthogonal' and assigned a special feature, are compatible with any hyponym of bicycle (a gentleman's bicycle as well as an old bicycle can at the same time be a racer bike or a mountain bike), members of the taxonomical group are only compatible with the members of the orthogonal group (a racer bike cannot be a mountain bike). In Figure 1, the orthogonal synset 'herrecykel' (gentleman's bicycle) is illustrated by a rhombus, in contrast to the taxonomical hyponyms 'ladcykel' (carrier cycle) and 'klubcykel' (standard bicycle).

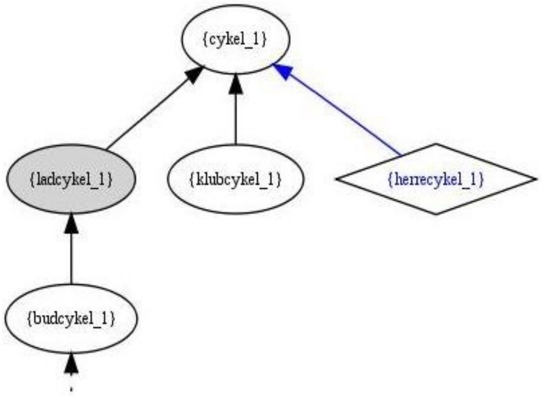


Figure 1. Some hyponyms of 'cykel' (bicycle). The rhombus figure indicates orthogonal hyponymy.

The encoded data on orthogonal versus taxonomical hyponymy in DanNet represents a new description of Danish concepts, the information on different categories of hyponyms not being deducible from the data in a traditional semasiological dictionary like DDO. See (Pedersen and Sørensen, 2006), (Pedersen and Nimb, 2008) and (Pedersen et al., forthcoming), for further discussion of the hyponymy relation in DanNet, also in the case of verbs.

The orthogonal feature makes it possible to carry out linguistic investigations on the nature of the hyponymy relation between Danish words. Concepts can be classified as belonging to three different levels according to Dirven and Verspoor (1998, p. 38): the general level (plant, animal, garment), the basic level (tree, dog, trousers) and the specific level (oak, labrador, jeans). If we consider the hyponymy hierarchy for the approx. 6,800 concrete objects in DanNet, we find a very even distribution between the number of taxonomical and orthogonal co-hyponyms at the general language level. In other words, the direct hyponyms of ‘genstand’ (object), that is concepts like garment, toy, tool, and vehicle, have many orthogonal sister concepts which, in principle, are compatible with any taxonomical hyponym of ‘genstand’, such as ‘ejendom’ (property), ‘blikfang’ (eye catcher), ‘eksemplar’ (specimen), ‘helligdom’ (shrine), ‘kopi’ (copy), 'nyhed' (novelty), 'opfindelse' (invention), 'original' (original) and ‘værdigenstand’ (article of value). In Danish we have many words denoting any kind of object which is owned, copied, invented, new, valuable etc. We find a much smaller percentage of orthogonal hyponyms the further down we move in the DanNet hyponymy hierarchy, also when it comes to the generally quite large sets of hyponyms of the basic level concepts (e.g. book: 28 taxonomical and 14 orthogonal hyponyms; shoe: 28 taxonomical and 5 orthogonal hyponyms, trousers: 16 taxonomical and 0 orthogonal hyponyms). The concepts at the specific language level seem to have very few orthogonal sister concepts.

3 The set of semantic relations in DanNet

The set of semantic relations in DanNet is based on the wordnet relations from EuroWordNet (Vossen, 1998), extended by three relations from the SIMPLE lexicon. In the SIMPLE model (Lenci et al., 2000), semantic relations are organized according to the four qualia roles (Pustejovsky, 1995), relating to inheritance structure, origin, composition and purpose. None of the EuroWordNet relations cover the origin dimension and the purpose dimension of a concept. During the compiling of the Danish SIMPLE lexicon (Pedersen and Paggio, 2004), it turned out that the four-dimensional qualia structure in general ensured most semantic aspects of a word sense to be described in the lexicon. Therefore, the two SIMPLE relations 'made_by' and 'used_for' were included in DanNet. Also the relation 'concerns' from the SIMPLE model was added. Furthermore some relations on synonymy are part of the wordnet set of relations. See Table 1.

Formal Role (INHERITANCE)	has_hyperonym has_hyponym is_a way_of
Agentive Role (ORIGIN)	made_by (from SIMPLE)
Constitutive Role (COMPOSITION)	has_holo_made_of has_holo_part has_holo_member has_holo_location has_mero_made_of has_mero_part has_mero_member concerns (from SIMPLE) involved_agent involved_patient involved_instrument
Telic Role (PURPOSE)	used_for (from SIMPLE) used_for_object role_agent role_patient
Synonymy	near_synonym near_antonym xpos_near_synonym

Table 1 Semantic relations in DanNet

4 A concept and its relations in DanNet

The relations assigned to a concept, e.g. the basic-level concept 'bog' (book), see Table 2, is in DanNet mainly based on the DDO sense definitions. In addition to this, an examination of the hyponyms of the concept also proved necessary as the set of hyponyms often reveals a number of central semantic aspects of the hypernym in question which are not mentioned in the DDO definition. Consider for example the many hyponyms of 'bog' (book) which describe the topic of the book thus making it clear that the topic is in fact a central semantic aspect of a book, even though this is not mentioned in the definition of 'bog' itself in DDO. We find 'fuglebog' (bird book, concerns: bird), 'kogebog' (cookery book: concerns: cooking), 'kriminalroman' (detective novel, crime novel: concerns: crime). The semantic relation 'concerns: topic' has therefore been assigned at the top level of the 'bog'-hierarchy in DanNet and is subsequently restricted to a more precise synset for those hyponyms having a specific topic sense.

Ontological type	[LanguageRepresentation+ Artifact+Object]
Formal role/ INHERITANCE	has_hyperonym: 'genstand' (object)
Agentive role/ ORIGIN	made_by: skrive (write); trykke (print)
Constitutive role/ COMPOSITION	has_mero_made_of: papir (paper) has_mero_part: tekst (text), side (page), ryg (back), titel (title) concerns: emne (topic) involved_agent: forfatter (writer) involved_agent: læser (reader)
Telic role/ PURPOSE	used_for: læse (to read)
Synonymy	near_synonym: hæfte (booklet; pamphlet)

Table 2. The semantic relations of 'bog' (book) in DanNet

In DanNet, the aim has been to describe explicitly as much semantics as possible by giving precise relations to other concepts in order to compensate the likely deficit of world knowledge in NLP software using lexical data like DanNet. Veale and Hao (2008) claim that even the kind of knowledge we normally find in dictionaries does not cover what it takes to make a computer understand everyday language, and that wordnets should be enriched with information on stereotypes and culturally-inherited associations. This is outside the scope of DanNet at its current stage the aim being instead to define the native speaker’s lexical knowledge about a concept and focus on the prototypical semantic aspects. From an ideal point of view, DDO would contain exactly this level of information so that the information found here just needed to be translated into semantic relations in DanNet, but due to the fact that dictionary definitions lean on the language-user's ability of making assumptions (Svensén, 1993) this is often far from beeing the case. Also for syntactic reasons DDO does not always bring all the information needed in DanNet. The definition in DDO had to be a well-formed, not too complicated or long phrase, and this is probably the reason why nothing is said about the topic in the case of books, nor about books typically having a title, being written by an author, read by a reader etc. Furthermore, the entries in DDO are meant to be read as a whole, implying that some semantic aspects might emerge from the examples, the list of connotations etc. Finally and maybe most importantly, the DDO definitions were created in a bottom-up way, without schematic specifications for a given group of words in order to ensure all relevant semantic aspects to be covered systematically. Therefore it is not surprising that we often find a discrepancy between the sometimes quite large number of relations which from a systematic point of view should be described for a given sense in order to reflect the native speaker's lexical knowledge, and the ones which are explicitly described in the definition of the word in DDO.

Comparing DDO and DanNet, we can conclude that in the case of artifacts DanNet in general contains more information on the meronymy relations than DDO does, especially in the cases of the basic-level concepts. In DanNet we find information on books having pages, a back and a title, and on shops having display windows, information not found in DDO. DanNet also contains far more information than DDO does on the typical user of an artifact (e.g. easy reader / pupil, hymn book / church goer). In Table 3 we present a range of examples of cases where information on the typical user has been added in DanNet, compared to what is mentioned in DDO. What is interesting in these cases is that the artifact lemma is often morphologically closely related to the user or vice versa, as in the examples of shop, shopkeeper, shopper; pharmacy, pharmacist; bakery, baker; pilot licence, pilot.

Synset	**Added information in DanNet compared to DDO**
flyvecertificat (pilot licence):	involved_agent: pilot (pilot)
briller (glasses)	involved_agent: person (person)
forskningsbibliotek (research library)	involved_agent: forsker (researcher)
læbestift (lipstick)	involved_agent: kvinde (woman)
barberkost (shaving brush)	involved_agent: mand (man)

Synset	Added information in DanNet compared to DDO
ægteskab (marriage)	involved_agent: ægtepar (married couple)
apotek (pharmacy)	involved_agent: apoteker (pharmacist)
bageri (bakery):	involved_agent: bager (baker)
registreringsattest (vehicle registration certificate)	involved_agent 'motorkontor' (motoring office).

Table 3. Examples of added information in DanNet compared to what is described in DDO

A statistical investigation of the manually added relations (i.e., those not automatically inherited from the hypernym of the synset) in the synsets of 6,800 object artifacts gives an idea of the most important relations when describing an artifact by semantic relations. See Table 4.

Semantic relation	Percentage of 6,800 artifact objects described with the relation
used_for	28% (book/to read)
has_mero_part	14% (book/page)
concerns	9% (christmas decorations/christmas)
made_by	6% (clothes/to sew)
involved_agent	6% (guitar/guitarist)
has_holo_part	5% (page/book)
has_mero_madeof	5% (clothes/fabric)
has_holo_location	3% (carpet/floor)
near_synonym	3% (book/pamphlet)
Others relations	1% or less

Table 4. The distribution of the percentage of manually assigned relation types in 6,800 synsets with an artifact object sense (inherited relations not included).

The frequent use of the DanNet relations taken from the SIMPLE lexicon model (used_for, made_by, and concerns) supports the decision of extending the set of standard WordNet relations. It should be remarked that this type of information is often deducible from the DDO definition, in opposition to the information in DanNet on the involved user.

The number of manual assignments of relations also indicates how often we find lexical restrictions between the relations in artifact synsets. In DanNet a general 'used_for' relation is always assigned at the top hypernym of a certain group of artifacts (e.g. tool: used_for: to use; garment: used_for: to dress). Also the involved_agent relation is assigned here with the value 'person' (person), e.g. tool: involved_agent: person. Whenever the inherited relation value is too imprecise and a manual assignment of the two relations is applied for a hyponym, it reflects a lexical relation between the artifact synset itself, the synset describing the kind of use, and the synset describing the kind of user. We find these cases relatively often, since one out of four cases of a manual assignment of the used_for relation, e.g. for shaving brush to shave, and for pilot licence to fly, also has resulted in a restriction on the type of user, e.g. shaving brush: man; and pilot licence: pilot.

5 Conclusion

DanNet contains a high number of well-structured and consistent semantic data on the Danish word senses, and in several cases also more information than what can be found in the definitions in the dictionary on which the wordnet is based, e.g. on different groups of hyponyms and on the involved user of artifacts. The investigations in the DanNet data of 1) the distribution of taxonomical and orthogonal hyponyms at different conceptual levels and 2) the distribution of the different relations used to describe artifact synsets, which have been presented here, shed new light on the semantic relations between a group of concepts in the Danish lexicon and is just a minor example of the types of lexical-semantic studies that can be carried out on a wordnet like DanNet.

References

D.A. Cruse. 2002. Hyponymy and Its Varieties. Green, R., Bean, C.A., Myaeng, S.H. (eds.) *The Semantics of Relationships: An Interdisciplinary Perspective, Information Science and Knowledge Management*. Springer Verlag.

DDO = Hjorth, E., Kristensen, K. et al. (eds.). 2003-2005. *Den Danske Ordbog 1-6* ('The Danish Dictionary 1-6'). Gyldendal and Society for Danish Language and Literature, Denmark.

R. Dirven and M. Verspoor (eds.). 1998. *Cognititive Exploration of Language and Linguistics*, John Benjamin Publishing Company Amsterdam/ Philadelphia.

A. Lenci., N. Bel, F. Busa, N. Calzolari, E. Gola, M. Monachini, A. Ogonowski,, I. Peters, W. Peters, N. Ruimy, M. Villegas. and A. Zampolli. 2000. SIMPLE – A General Framework for the Development of Multilingual Lexicons. T. Fontenelle (ed.) *International Journal of Lexicography* Vol 13, pp. 249-263. Oxford University Press.

B.S. Pedersen and P. Paggio. 2004. The Danish SIMPLE Lexicon and its Application in Content-based Querying. *Nordic Journal of Linguistics* Vol 27(1), pp. 97-127. Cambridge University Press.

B.S. Pedersen and N. Sørensen. 2006. Towards Sounder Taxonomies. A. Oltramari, Chu-Ren Huang, A. Lenci, P. Buuitelaar, C. Fellbaum (eds) *Wordnets*. Ontolex 2006 at 5th International Conference on Language Resources and Evaluation, pp. 9-16. Genova, Italy.

B.S. Pedersen and S. Nimb. 2008. Event Hierarchies in DanNet. A. Tanács, D. Csendes, V. Vincze, C. Fellbaum, and P. Vossen (eds) *Proceedings of Global WordNet Conference,* University of Szeged, Hungary, pp. 339-349. University of Szeged, Juhász Press Ltd., Hungary.

B.S. Pedersen, S. Nimb, J. Asmussen, N. H. Sørensen, L. Trap-Jensen and H. Lorentzen (forthcoming). *DanNet - the challenge of compiling a WordNet for Danish by reusing a monolingual dictionary.* Language Resources and Evaluation. Springer Netherlands.

J. Pustejovsky. 1995. *The Generative Lexicon*. The MIT Press, Cambridge, Massachusetts.

B. Svensén. 1993. *Practical Lexicography. Principles and Methods of Dictionary-making* [translated from the Swedish Handbok i lexikografi (1987) by J. Sykes and K. Schofield] Oxford: Oxford University Press.

T. Veale and Y. Hao. 2008. Enriching WordNet with Folk Knowledge and Stereotypes. *Proceedings of the Fourth Global Wordnet Conference*, University of Szeged, Hungary, pp. 453-461. University of Szeged, Juhász Press Ltd., Hungary.

P. Vossen (ed.). 1999. *EuroWordNet, A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, The Netherlands.

Ontologies vs. classification systems

Bodil Nistrup Madsen
Copenhagen Business School
Copenhagen, Denmark
bnm.isv@cbs.dk

Hanne Erdman Thomsen
Copenhagen Business School
Copenhagen, Denmark
het.isv@cbs.dk

Abstract

What is an ontology compared to a classification system? Is a taxonomy a kind of classification system or a kind of ontology? These are questions that we meet when working with people from industry and public authorities, who need methods and tools for concept clarification, for developing meta data sets or for obtaining advanced search facilities. In this paper we will present an attempt at answering these questions. We will give a presentation of various types of ontologies and briefly introduce terminological ontologies. Furthermore we will argue that classification systems, e.g. product classification systems and meta data taxonomies, should be based on ontologies.

1 Introduction

In recent years many authors have discussed the nature of ontologies and proposed various definitions and subtypes of ontologies for various purposes, among them Gruber (2007), Guarino (1998), Gómez-Pérez et al. (2004). According to CEN (2004) ontologies and taxonomies are types of knowledge structuring, as shown in Figure 1.

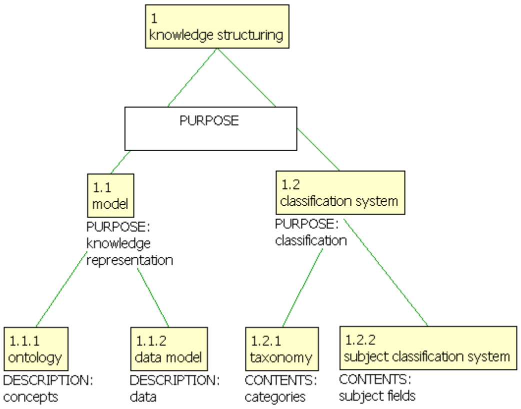


Figure 1: Some concepts related to knowledge structuring.

The ontology in Figure 1 comprises concepts (boxes with systematic notations) and subdivision criteria (boxes with text in capital letters). The concepts are related by means of type relations (lines between the concept boxes) and further described by means of feature specifications each consisting of an attribute-value pair (e.g. PURPOSE: knowledge representation).

According to the ontology in Figure 1 one may distinguish *models* and *classification systems* as follows: The purpose of a model is to *give a simplified representation of knowledge about phenomena*, whereas the purpose of a classification system is *the subdivision of phenomena into classes that form the basis for ordering 'things'*.

Very often a conceptual data model, represented by means of an ER diagram or an UML diagram, is referred to as ontology. Our recommendation is to use the term ontology only as defined here.

2 Various types of ontologies

In 2007, ISO Technical Committee 37, *Terminology and Other Language Resources* (ISO TC 37), set up an Ontology Task Force with the aim of proposing a strategy for the work on ontologies within TC 37. As a basis for this strategy, the Task Force will develop an overview of related ongoing projects, existing standards and proposals for future projects within TC 37 as well as an overview of examples of ontologies and projects 'outside' TC37. The first step in the work of the Ontology Task Force is to describe different types of knowledge representation resources, and to clarify the differences between these. One of the results is a systematic overview in the form of an ontology of ontologies which comprises proposals for definitions of the different types of ontology.

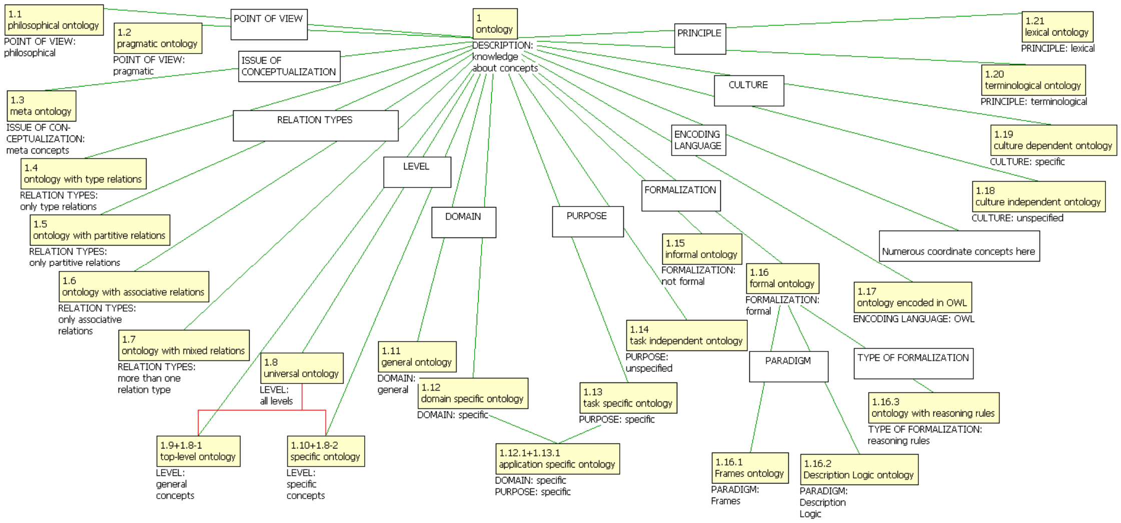


Figure 2: Ontology of ontologies.

Figure 2 presents this ontology of ontologies. The description of the concepts is to a great extent based on Guarino (1998). In this ontology characteristics and subdivision criteria are introduced that clearly distinguish the types of ontologies, e.g. LEVEL, DOMAIN and PURPOSE. The broken lines between concepts represent part-whole relations.

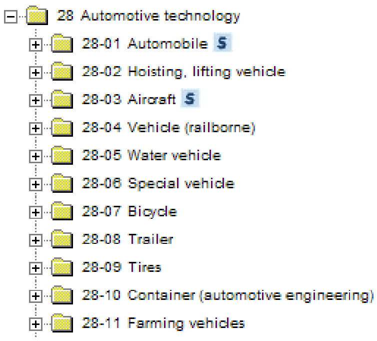
The ontology in Figure 2 may be characterized as a terminological ontology, i.e. an ontology that is based on the terminological method, making use of characteristics and subdivision criteria, cf. ISO 704 (2000).

A terminological ontology is a domain specific ontology. We use the term *terminological ontology* as a synonym of the term *concept system*, which is normally used in terminology work, cf. for example ISO 704 (2000). Gruber (2007) describes an ontology in the following way: *An ontology specifies a vocabulary with which to make assertions, which may be inputs or outputs of knowledge agents (such as a software program). … an ontology must be formulated in some representation language* … In our view, the demand for a representation language narrows the concept, i.e. Gruber's definition describes the concept *formal ontology* in Figure 2.

3 Ontologies as the basis for classification systems

As already mentioned, we distinguish *ontology* and *classification system* with respect to purpose. However, we strongly recommend that a classification system is built on the basis of a terminological ontology or by using the principles of terminological ontologies.

In the extract of the product classification system eCl@ss in Figure 3, it is evident that by using principles of terminological ontologies, this system could be structured in a more logical way, and thus could be intuitively easier to use: *automobile*, *aircraft*, *railborne vehicle* and *water vehicle* are distinguished with respect to "channel of transportation". For example *automobiles* are meant for traveling on streets or roads while *aircrafts* are designed to travel through the air. *Farming vehicles* and *hoisting, lifting vehicles* are characterized with respect to purpose. The order of the classes does not make this clear.



28 Automotive technology
28-01 Automobile S
28-02 Hoisting, lifting vehicle
28-03 Aircraft S
28-04 Vehicle (railborne)
28-05 Water vehicle
28-06 Special vehicle
28-07 Bicycle
28-08 Trailer
28-09 Tires
28-10 Container (automotive engineering)
28-11 Farming vehicles

Figure 3: Extract of a product classification system.

Figure 4 presents an ontology with concepts corresponding to the classes in Figure 3. Since some of the classes in Figure 3 do not refer to automobiles, the top concept chosen is *vehicle*.

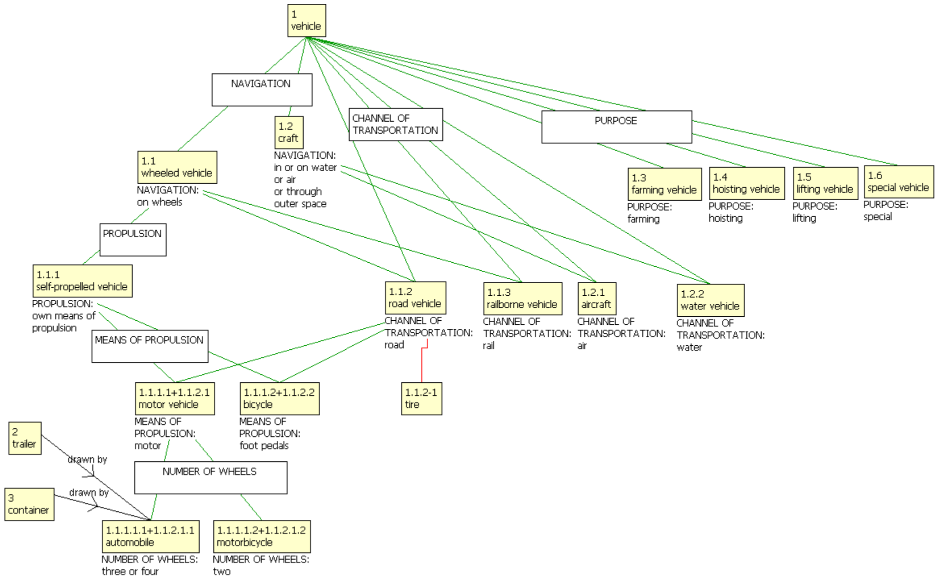
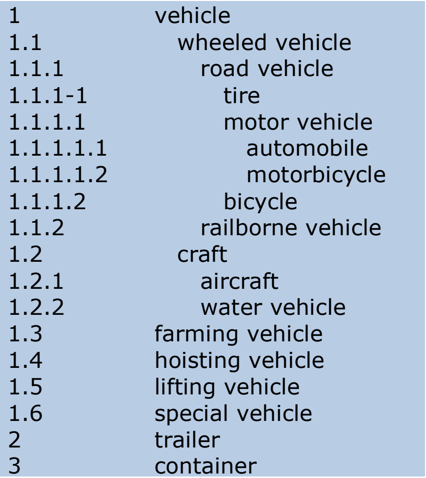


Figure 4: Ontology of vehicles.

In the ontology in Figure 4 the concepts are clearly delimited from each other by means of subdivision criteria: NAVIGATION, CHANNEL OF TRANSPORTATION, etc. It may be useful to introduce subdivision criteria also in a classification system in order to make this clear.



1	vehicle
1.1	wheeled vehicle
1.1.1	road vehicle
1.1.1-1	tire
1.1.1.1	motor vehicle
1.1.1.1.1	automobile
1.1.1.1.2	motorbicycle
1.1.1.2	bicycle
1.1.2	railborne vehicle
1.2	craft
1.2.1	aircraft
1.2.2	water vehicle
1.3	farming vehicle
1.4	hoisting vehicle
1.5	lifting vehicle
1.6	special vehicle
2	trailer
3	container

Figure 5: Extract of a classification system.

It is not intuitively understandable why the class *Bicycle* belongs to *Automotive technology* in Figure 3, but it may be because this class comprises *motor driven bicycles*. However, a closer look into the class *Bicycle*, reveals that the class also comprises the class *Bike*.

During the concept clarification process it turned out that there was a need for introducing the two concepts *wheeled vehicle* and *craft* which were not in the classification in Figure 3. Based on the ontology in Figure 4, a classification list like the one in Figure 5 can be developed.

When building a classification system on the basis of an ontology, some simplifications will typically be made. In Figure 5 the concept *self-propelled vehicle,* which is a superordinate concept to *motor vehicle* and *bicycle,* is not found as a class. One may also consider to leave out the class *bicycle* for the above mentioned reasons.

As already mentioned, it may be useful to introduce subdivision criteria in order to make explicit the differences between the classes.

4 Classification systems compared to concept systems

A characteristic of a classification system is that the nodes are not always concepts, but often groups of concepts. This is true in the Semantic Types of UMLS (Unified Medical Language System), cf. Figure 6.

The Semantic Network consists of (1) a set of broad subject categories, or Semantic Types, that provide a consistent categorization of all concepts represented in the UMLS Metathesaurus®, and (2) a set of useful and important relationships, or Semantic Relations, that exist between Semantic Types, cf. (Bodenreider, 2005) and the Semantic Network Fact Sheet (http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html).

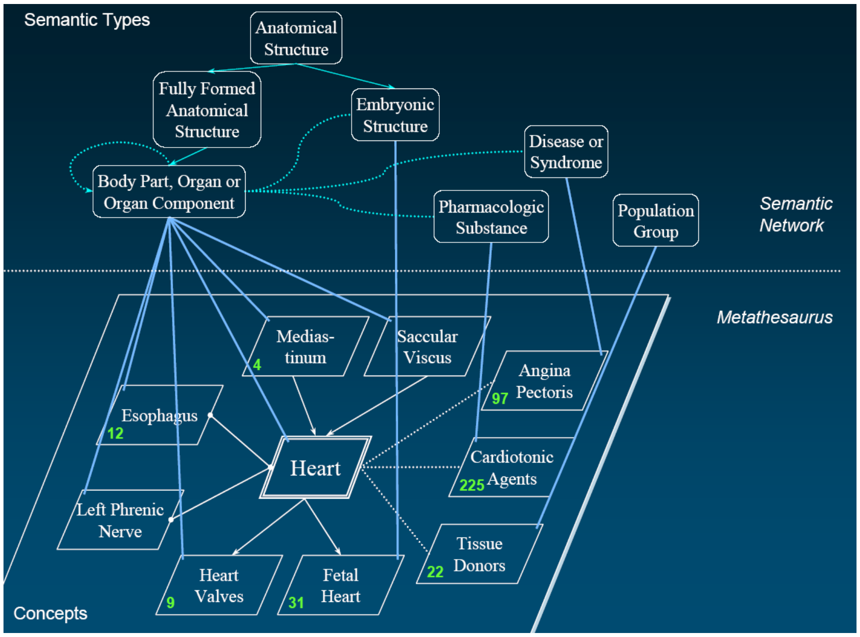


Figure 6: Example from UMLS.

An example of a semantic type is '*Body Part, Organ or Organ Component*', which conflates three concepts: *body part*, *organ* and *organ component*. In an ontology these three would be separate concepts (nodes).

5 Ontologies as the basis for meta data taxonomies

In order to facilitate data exchange and interoperability, it is important to be able to describe elements of data collections systematically and unambiguously. This is the reason why metadata registries comprising sets of metadata categories with negotiated definitions and examples, exist in many fields.

When defining a set of metadata categories it is very useful to base it on a kind of systematization, e.g. a taxonomy, specifying main categories, categories and subcategories. Otherwise one may end up with an incomplete and inconsistent set of categories that is very difficult to use and to extend.

In order to obtain a well structured taxonomy we will argue that it should be based on the elaboration of a terminological ontology. In this way the concepts of the domain and their interrelations are clarified. In some cases it is even possible to generate a taxonomy on the basis of an ontology, i.e. some concepts of the ontology may more or less automatically be transformed into categories of the taxonomy. In other cases, the ontology renders the knowledge which forms the basis for the construction of the taxonomy.

6 Data categories for linguistic resources

ISO 12620:1999, *Computer assisted terminology management — Data Categories* specifies data categories used in terminological resources. These data categories are classified in three major groups and ten sub-groups:

Term and term-related data categories:
A.1 term
A.2 term-related information
A.3 equivalence
Descriptive data categories:
A.4 subject field
A.5 concept-related description
A.6 concept relation
A.7 conceptual structures
A.8 note
Administrative data categories:
A.9 documentary language
A.10 administrative information

This structure is not homogenous, i.e. it reflects various subdividing criteria (dimensions), and it does not give a very clear overview of the data categories.

One dimension is for example term-related information vs. concept-related description. Here it is not clear why e.g. *subject field* and *concept relation* do not fall within the group: *concept-related description*.

In 2003, it was proposed to set up a Data Category Registry (DCR) in TC 37 for all kinds of lexical data. Since this DCR also includes data categories of dictionaries, the above structure was not very appropriate. Consequently it was decided to give up a classification of the categories. In our opinion it will, however, be difficult to ensure completeness, consistency, user-friendliness and extensibility of the above mentioned DCR, if there is no structure at all of the data categories.

7 Ontologies as the basis for meta data taxonomies

Figure 7 presents an extract of a terminological ontology for concepts pertaining to semantic information that may be registered in lexical data collections, such as e.g. termbases and electronic dictionaries. The three main types of semantic information are *subject classification*, *content specification* and *semantic relation*.

This ontology uses type relations, part whole relations and associative relations (lines with the designation of the relation type and an arrow indicating the direction of the relation).

The group of concepts on the right hand side, which are related by means of associative and part-whole relations, contribute to a better understanding of the concepts that are central for semantic information. For example, it is illustrated that a *content specification* describes the *intension* of a concept, and that the *intension* consists of *characteristic features*.

8 The Danish standard of lexical resources

The Danish Standard DS 2394-1:1998 comprises a taxonomy for the classification of lexical data, the STANLEX taxonomy. In STANLEX the main groups of information types are structured according to the linguistic disciplines: etymological information, grammatical information, graphical information, phonetic information, semantic information and usage. Examples of categories and sub categories are shown in Table 1.

9 From ontology to taxonomy

The ‘backbone’ of the ontology in Figure 7 consists of the top concept *semantic information* and the subordinate concepts which are related to this concept by means of type relations: *lexical paraphrase*, *analytic definition* etc. These concepts will typically form the background for categories to be included in a taxonomy. As already mentioned, the concepts that are related by means of part-whole relations or associative relations typically give a better understanding of the central concepts, but it will often not be relevant to introduce corresponding categories in a taxonomy.

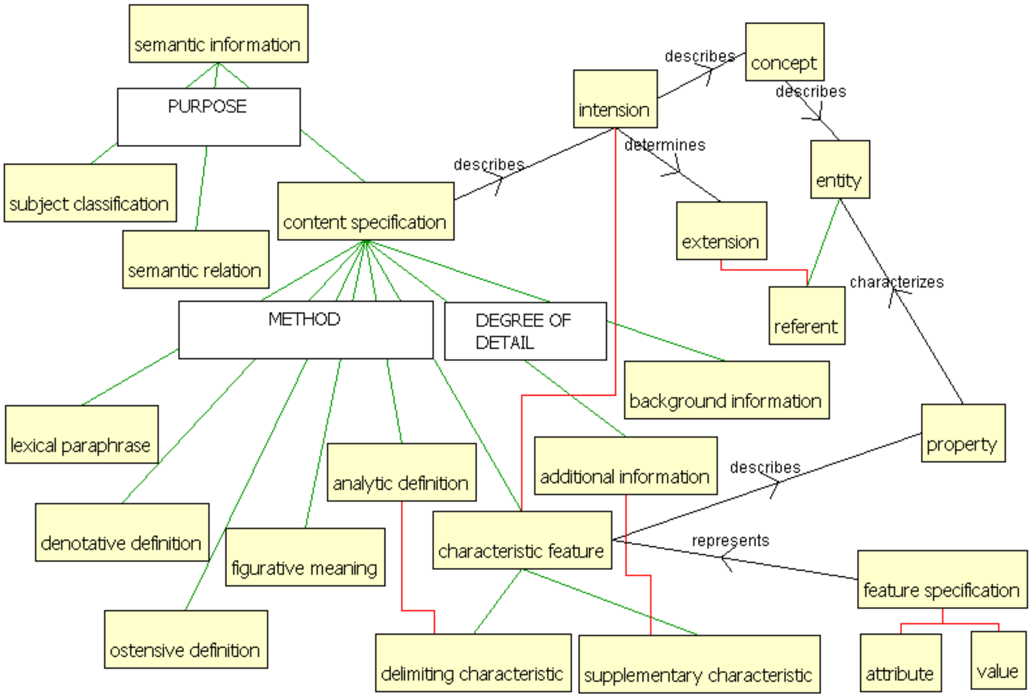


Figure 7: Ontology of semantic information.

Main group	Category	Subcategory
Semantic information	Subject classification	• Classification system • Normative subject classification • Nonnormative subject classification
	Semantic relations	• Concept system • Position of concept in concept system • Generic relation • Partitive relation • Successive relation • Causal relation • Associative relation • Antonymy • Metonymy • Equivalence within one language • Equivalence between two or more languages • Equivalence constraint
	Content specification	• Lexical paraphrase • Analytic definition • Denotative definition • Ostensive definition • Additional information • Background information • Characteristic feature • Figurative meaning

Table 1: Categories and subcategories of Semantic Information.

The nodes in a taxonomy represent categories, not concepts, and a taxonomy category may sometimes correspond to more concepts. This may be more user friendly, since the user of the taxonomy will then not have to worry about subtle distinctions. For example, in Figure 7, the concept *additional information* refers to information in the form of supplementary characteristics, while *background information* gives further information about historical, technical, legal or other aspects of the semantics of the lexical entry. In a taxonomy, one might decide to 'merge' the two concepts *additional information* and *background information* into one category, since it may be difficult for the user of the taxonomy to choose between them.

Sometimes the taxonomy will not comprise the 'lowest' levels of a hierarchy in the corresponding ontology. For example there may not be a need for distinguishing between *delimiting characteristics* and *supplementary characteristics* in the taxonomy. This is the case in the Danish Standard of lexical data categories.

In some cases it may be relevant to convert concepts of an ontology participating solely in associative or part-whole relations into categories in a taxonomy. For example it may be relevant to include the categories *feature specification*, *attribute* and *value* from Figure 5 as taxonomy categories.

10 Conclusion

In this paper we have argued that by applying principles of terminological ontologies when constructing a product classification system or a metadata taxonomy, it is possible to obtain a clear and intuitively understandable structure and in this way to obtain completeness, consistency, user-friendliness and extensibility. In some cases an ontology may be mapped directly into a classification system, but in other cases it will be necessary and useful to introduce adjustments into the classification system compared to the ontology. The principles that we introduce here are relevant for the development of all kinds of classification systems.

References

Bodenreider, Olivier. 2005. Consistency between Metathesaurus and Semantic Network Workshop on *The Future of the UMLS Semantic Network.* NLM.

DS 2394-1. 1998. *Lexical data collections – Description of data categories and data structure – Part 1: Taxonomy for the classification of information types*, Danish Standards.

CWA 15045. 2005. *CEN Workshop Agreement: Multilingual Catalogue Strategies for eCommerce and eBusiness.*

eCl@ss: http://www.eclass-online.com/

Gruber, Tom. 2008. *Ontology. Encyclopedia of Database Systems*, Ling Liu and M. Tamer Özsu (Eds.), Springer-Verlag. Looked up on February 23 2009: http://tomgruber.org/writing/ontology-definition-2007.htm

Gómez-Pérez, Asunción; Mariano Fernández-López & Oscar Corcho. 2004. *Ontological Engineering – with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. London: Springer Verlag.

Guarino, Nicola. 1998. Formal Ontology and Information Systems. *Formal Ontology in Information Systems, Proceedings of the First International Conference (FOIS'98)*, Amsterdam: IOS Press.

ISO 704. 2000. *Terminology work — Principles and methods.* Genève: ISO.

Extending the adverbial coverage of a French wordnet

Benoît Sagot
INRIA Paris-Rocquencourt / Paris 7
Paris, France
benoit.sagot@inria.fr

Karën Fort
INIST
Nancy, France
karen.fort@inist.fr

Fabienne Venant
INRIA Nancy Grand-Est
Nancy, France
venantfa@loria.fr

Abstract

This paper presents a work on extending the adverbial entries of WOLF, a semantic lexical resource for French. This work is based on the exploitation of the derivation and synonymy relations; the latter are extracted from the DicoSyn synonyms database. The resulting semantic resource, which is freely available, is manually evaluated and validated in an exhaustive manner.

1 Introduction

Nowadays, the availability of resources for Natural Language Processing (NLP) remains a hot topic, in particular for French. The situation is slightly improving as compared to English as far as morphological and syntactic resources are concerned (Sagot et al., 2006). However, this is not yet the case for semantic resources, despite efforts made to provide a freely-available wordnet for French, WOLF (see Section 2.2).

In this paper, we describe a first step in this direction. Restricting our area of investigation to adverbs, our goal is to complete WOLF, thanks to the morphological and syntactic lexicon Le*fff* (Sagot et al., 2006) and the synonyms database DicoSyn (Ploux and Victorri, 1998).

This paper is organized as follows. In Section 2, we introduce the three resources used in our work. In Section 3 we describe how we extended WOLF thanks to two complementary techniques. Finally, in Section 4 we detail the results of the exhaustive manual evaluation of the resulting entries.

2 Ressources

2.1 Le*fff* and the Lexique-Grammaire tables

Le*fff* (Lexique des Formes Fléchies du Français, *Lexicon of French Inflected Forms*) (Sagot et al., 2006), is a large-coverage morphological and syntactic lexicon for French which is freely available.[1] Le*fff* aims at conciliating linguistic relevance and usability in NLP applications. In particular, it is used in several parsers that rely on various formalisms (LFG, TAG). Le*fff*, currently in version 3, covers all categories and is progressively enriched with syntactic and semantic information, notably by comparing it to other syntactic resources (Danlos and Sagot, 2007). Thus, adverbial entries in Le*fff* were enhanced (Sagot and Fort, 2007) thanks to the Lexique-Grammaire tables of adverbs in *-ment*, the so-called Molinier tables (Molinier and Levrier, 2000).

In French, adverbs ending in *-ment* form a large class of adverbs. Moreover, as opposed to other adverbs, it is an open class. Those adverbs form a morphologically homogeneous class, since most of them are built according to the pattern adjective + *ment*. Numerous other adverbs exist, and in particular a large amount of adverbial phrases, but they lie beyond the scope of this work.

2.2 WOLF

WOLF (WOrdnet Libre du Français, *Free French Wordnet*) is a semantic lexical resource for French, freely available (Sagot and Fišer, 2008).[2] It is a *wordnet*, based on the model of the Princeton WordNet (PWN), the first wordnet ever developed, which deals with English (Fellbaum, 1998). Like any wordnet, WOLF is a lexical database in which words (lexemes, literals) are divided by parts-of-speech and organized into a hierarchy of nodes. Each node has a unique id, and represents a *concept* or *synset* (set of synonyms). It groups a certain amount of synonymous lexemes that denote this concept. For example, in the PWN (version 2.0), the synset ENG20-02853224-n contains the

[1] http://gforge.inria.fr/projects/alexina/

[2] http://wolf.gforge.inria.fr/

lexemes {*car, auto, automobile, machine, motorcar*}. Lexemes can be single words as well as multi-word expressions, taking also into account metaphoric and idiomatic usage. Synsets also contain a short gloss, and are related to other synsets. For example, the above-mentioned synset is related to the synset {*motor vehicle, automotive vehicle*} by a hypernymy relation, and to the synset {*cab, hack, taxi, taxicab*} by a hyponymy relation.

WOLF was built using the PWN 2.0 and various multilingual resources, thanks to two complementary approaches. Polysemous lexemes were dealt with using an approach that relies on parallel corpora in five languages, including French, that were word-aligned. Several multilingual lexicons were extracted from those aligned corpora, taking into account three to five of the available languages (precision and recall of these lexicons vary w.r.t. the number of languages taken into account). Multilingual lexicons were semantically disambiguated thanks to wordnets for the corresponding languages. On the other hand, monosemous PWN lexemes only required bilingual lexicons that were extracted from wiki resources (Wikipedia, Wiktionary) and thesauri. Nominal and verbal sub-wordnets of WOLF were evaluated against the French wordnet built during the EuroWordNet project.[3]

WOLF contains all PWN 2.0 synsets, including those for which no French lexeme is known. The latest version of WOLF before this work, version 0.1.4, includes French adverbial lexemes for only 676 of the 3,664 adverbial synsets, i.e., only 18.4%, and only 983 lexeme-synset pairs corresponding to only 665 unique adverbial lemmas. For this reason, we applied two complementary techniques to improve WOLF's coverage. One of those techniques relies on the morphological and semantic derivation relation that often exists between an adverbial synset and its corresponding adjectival synset, both in English and French. The other technique relies on the expoitation of the synonyms database DicoSyn.

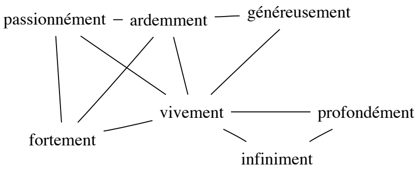


Figure 1: Extract from the adverbial synonymy graph

2.3 DicoSyn and the cliques of synonyms

DicoSyn is an electronic dictionary of synonyms, whose latest versions are available for online usage.[4] The initial base (Ploux and Victorri, 1998) was created merging seven French classic dictionaries (Bailly, Benac, Du Chazaud, Guizot, Lafaye, Larousse and Robert) from which the synonymic relations were extracted. The major advantage of this dictionary is that it explicitly shows the graph of the synonymy relation.[5] Ploux and Victorri designed Visusyn, that allows to explore the graph. It is then possible to automatically visualize and characterize the semantic properties of a unit, using the sub-graph it constitutes with its synonyms (François et al., 2002; Venant, 2004), or to study in a more global way the semantic characteristics of a whole lexical paradigm (Venant, 2007). We were thus able to exploit a graph of adverbial synonyms. As DicoSyn does not contain any indication regarding categories, this graph was built mapping DicoSyn with the adverbs in *-ment* from Le*fff*. The graph comprises 1,597 nodes (adverbs) and 4,344 (synonymy) connections. Among those nodes some are not adverbs ending in *-ment*, but synonyms of such adverbs (for example, *bien* is a node of the graph due to the fact that DicoSyn indicates that it is a synonym of *amplement* or *copieusement*). Figure 1 presents an extract from this graph.

We exploited this graph using the notion of clique. A clique is a the largest possible set of nodes connected as pairs. Thus, the graph in figure 1 contains 3 cliques: {*ardemment, fortement, passionnément, vivement*} (we cannot add *généreusement* which is neither a synonym

[3] The wordnet developed during the EuroWordNet project (Vossen, P., 1999) is the only other French wordnet. It contains only nominal and verbal synset, but no adjectival or adverbial synsets. Moreover, important license problems explain why it is rarely used in the research community. Finally, and partly for the same reason, it has not been improved since its creation. Those are the three main motivations for the development of WOLF.

[4] http://dico.isc.cnrs.fr/ and http://elsap1.unicaen.fr/dicosyn.html

[5] It is of course a partial synonymy relation, that Ploux and Victorri define in the following way: "Two lexical units are in a synonymy relation if any of the occurrences of one of them can be replaced by any occurrence of the other in a certain number of environments, without profoundly modifying the meaning of the text it belongs to." It is a symmetric relation.

of *fortement*, or *passionnément*), {*ardemment, généreusement, vivement*} and {*infiniment, profondément, vivement*}. The obtained adverbial graph comprises 2,247 cliques. The idea behind this is that a clique corresponds to a possible usage of the adverb. A clique being a set of synonyms, it more or less corresponds to a WordNet synset. Thus, cliques constitute the structural unit of the graph semantic analysis.

3 Extending WOLF

As previously stated, we first extended WOLF in order to increase the number of non-empty adverbial synsets (for which at least one French lexeme exists) as well as the number of lexemes in each non-empty synset. To do so, we used two types of relations between lexemes: the derivation relation, between an adverb ending in *-ment* and its corresponding adjective, and the synonymy relation between adverbs, as defined by the cliques in DicoSyn .

3.1 Using the derivation relation

The method based on the derivation relation arose from the two following observations:

- The PWN includes a derivation relation (*derived*) that links some adverbial synsets to one or more adjectival synsets. This link indicates that some adjectival lexemes in the adjectival synset allow the construction, using morphological derivation (*-ly* suffix), of some adverbial lexemes of the adverbial synset. Naturally, this link also indicates a semantic connection between the two synsets.

- The mechanism of morphological and semantic derivation between adjectives and adverbs is often parallel in English (adjective + *ly*) and French (adjective$_{\text{fem,sing}}$ + *ment*).[6]

We therefore collected, for each adverbial synset, the (French) adjectives in the adjectival synset connected through the *derived* relation. We then applied the morphological derivation algorithm to those adjectives.[7] The obtained adverbs which appear in Le*fff* were kept and allocated to the adverbial synset (with a note specifying that the lexeme–synset links were built using morphological derivation).

Let us consider, for example, the ENG20-00115661-b synset. In WOLF 0.1.4, it only contains the (correct) lexemes *toujours* and *invariablement*. Yet, this synset is connected to the adjectival synset ENG20-02417249-a through a *derived* relation and the latter comprises the lexemes *permanent*, *invariable* and *perpétuel*. Therefore, the potential adverbs *permanentement*, *invariablement* and *perpétuellement* are built. The first one is removed, as it does not appear in Le*fff*, the second one confirms a lexeme that already belonged to the adverbial synset, and the last one allows the creation of a new lexeme–synset connection. In the end, the ENG20-00115661-b synset is transformed into {*toujours*, *invariablement*, *perpétuellement*}.

Using this method, the number of adverbial lexeme–synset relations in WOLF raised from 983 to 1,536 (+56%). The number of non-empty adverbial synsets raised from 676 to 969 (+43%). The number of adverbial lexemes in WOLF raised from 665 to 889 (+23%).

3.2 Using the synonymy relation

Once the adverbial synsets of WOLF completed using the derivation relation between adverbs ending in *-ment* and adjectives, we applied a method based on the synonymy relation, as defined by the DicoSyn cliques. Three steps were necessary.

1. We first associated to each lexeme–synset connection a weighting rate according to their origin (see section 2.2). If a connection was built (among other sources) from bilingual resources (wiki resources), it receives a rate of 5. If the connection was built using aligned multilingual corpora, the rate is 4, if one of the corpus contained at least 4 languages, 3 if they all contained only 3 languages. In all other cases, including for connections built using the derivation relation, a rate of 2 is associated to the connection.

2. Each adverbial synset is then associated to the DicoSyn clique which corresponds the most, i.e. not simply containing the highest number of lexemes in common, but rather maximizing the sum of the rates of the lexemes shared by the clique and the synset.

[6] This is of course not always true (see courante/couramment and many others), but it is still a reasonable heuristics.

[7] The feminine singular form of the adjective being taken from Le*fff*.

3. Each synset is then completed with all the lexemes (adverbs) belonging to the associated clique.

For example, let us consider the ENG20-00115661-b synset, the very same synset we previously detailled. Once extended using the derivation relation, it contained the adverbs *toujours*, *invariablement* and *perpétuellement*. As the first two were built using the French Wiktionary, they receive a rate of 5. The adverb *perpétuellement*, built by derivation, receives a rate of 2. Therefore, the clique maximizing the sum of the rates of the common lexemes is {*éternellement*, *invariablement*, *perpétuellement*, *sans cesse*, *toujours*}. Two adverbs were thus added to the ENG20-00115661-b synset, the multi-word adverb *sans cesse* and the *-ment* adverb *perpétuellement*.

Using those methods, we increased the number of lexeme–adverbial synset relations from 1,536 to 2,149, which represents a 28.5% increase.

4 Evaluation of the extended WOLF

4.1 Methodology

We conducted a manual evaluation of all the adverbial synsets we obtained, i.e. of the 2,149 lexeme–synset pairs, comprising 1,025 adverbial lexemes. Each author manually validated the couples comprising one fourth of the lexemes; the remaining fourth being evaluated by the three authors, thus allowing for inter-validator agreement calculus.

Validating a lexeme–synset pair consists in assigning it one of the following codes:

- OK: correct association;
- SC (Semantically close): one of the meaning of the lexeme is semantically close to that of the synset (hyponym, hypernym, pseudo-synonym);
- SR (Semantically related): one of the meaning of the lexeme is semantically related (but less close) to that of the synset;
- NR (Non Related): no meaning of the lexeme is related to that of the synset;
- CC (Composed Component): false association, but the lexeme is one of the component of a multi-word lexeme which would fit in the synset;

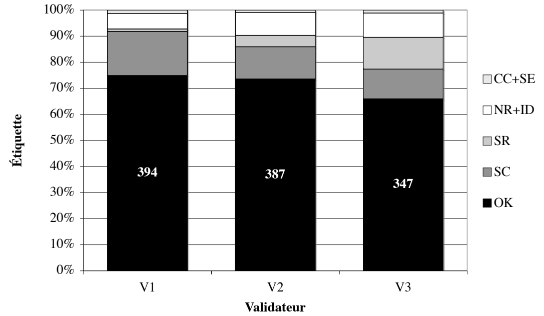


Figure 2: Comparison between the 3 validators of the distribution of evaluation codes for the same 530 lexeme–synset pairs

- ID (Incorrect Derivation): false association, due to a derivation issue such as an ambiguity of the intermediary adjective or the lack of parallel between morphological and semantic derivation (see, for example, *absolument* in the synset defined by *in a royal manner*)
- SE (Spelling Error): spelling error in the lexeme, the association is to be rejected;
- WC (Wrong Category): false association, due to an erroneous part-of-speech tagging of the lexeme (see, for example, *bougonnerie*)

4.2 Inter-validator agreement

For one fourth of the lexemes, the three authors carried out the evaluation independently. If we replace all the codes other than OK by a unique NONOK code, the three validators agree on 366 of the 530 lexeme-synset pairs, i.e., 69% of such pairs are validated three times NONOK or three times OK. The latter case (all validators agree the pair is correct) covers 292 lexeme-synset pairs (55%). Examining the distribution of the codes for each validator, we noticed differences in terms of tolerance level (see figure 2). As the boundary between codes like SC, SR and NR is difficult to define objectively, the variety of decisions about them is not surprising. On the opposite, over the 456 pairs judged OK by at least one of the validators, only 292 were validated (OK) by the three validators (64%) and 94 by two validators (20,6%). The agreement rate is therefore quite low. This can be explained by the difficulty of the task (some synsets cannot be easily differentiated) and by the scarcity of some adverbs.

The analysis of those results led us to associate a unique code to the lexeme–synset pairs evaluated by the three validators, in the following way:

- OK if the three evaluations are OK-OK-OK, OK-OK-SC, OK-OK-SR, or OK-SC-SC ;
- SC if they are OK-SC-SR or SC-SC-SR ;
- SR in the other cases where there is one or two OK amongst the three, as well as in the SC-SR-SR and SR-SR-SR cases;
- SE (ID, CC, WC) in the other cases, if a validator gave the SE code (ID, CC, WC);
- NR in the remaining cases.

Needless to say that the lexeme–synset pairs evaluated by only one validator keeps the code s/he gave them.

4.3 Evaluation results and obtained resource

The results are quite promising (see table 1), as we obtain more than 68% of correct lexeme-synset associations (OK). We kept 1,461 of the 2,149 lexeme-synset relations that we built automatically (as compared to 983 before this work, which were not manually validated). WOLF now contains 871 adverbial lexemes (as compared to 665 when we started) belonging to 871 non empty synsets (as compared to the initial 676). Therefore, the improvements in WOLF cover not only its quality, due to the manual validation, but also the number of synsets.

Total	OK	SC	SR	NR
2 145	**1 461**	296	147	162
100%	**68,1%**	13,8%	6,9%	7,6%

ID	CC	WC	SE
41	26	13	3
1,9%	1,2%	0,6%	0,1%

Table 1: Results of the manual validation

5 Conclusion and prospects

At a time when the lack of large scale lexical resources for French weights on NLP research, we showed the interest of using several existing resources to enrich or diversify their content. The Le*fff*–WOLF interaction, through DicoSyn, allowed us to enrich WOLF both in terms of quality and quantity. This work led to an increase of nearly 55% of the adverbial lexeme–synset relations in WOLF.

Those encouraging results also show that it is worthwhile exploiting a lexicon as a graph, at least as far as the automatic access to semantic information is concerned. The synonymy and the adverbs ending in *-ment* were ideal for this experiment and encourage us to explore other paradigmatic (hypernymy, antonymy) or syntagmatic (through corpus analysis) relations, as well as other parts-of-speech, like, for example, the nouns ending in *-ité* or the verbs in *-ifier* and *-iser*.

References

Laurence Danlos and Benoît Sagot. 2007. Comparaison du Lexique-Grammaire et de Dicovalence: vers une intégration dans le Le*fff*. In *Actes de TALN 07*, Toulouse, France.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Jacques François, Bernard Victorri, and Jean-Luc Manguin. 2002. Polysémie adjectivale et synonymie : l'éventail des sens de curieux. *La polysémie*.

C. Molinier and F. Levrier. 2000. *Grammaire des adverbes. Description des formes en* -ment. Droz, Geneva, Switzerland.

Sabine Ploux and Bernard Victorri. 1998. Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes. *Traitement Automatique des Langues (T.A.L.)*, 39(1):161–182.

Benoît Sagot and Darja Fišer. 2008. Building a free french wordnet from multilingual resources. In *Actes de Ontolex 2008*, Marrakech, Morocco. (à paraître).

Benoît Sagot and Karën Fort. 2007. Améliorer un lexique syntaxique à l'aide des tables du lexique-grammaire – adverbes en -ment. In *Actes du Colloque Lexique et Grammaire*, Bonifacio, France.

Benoît Sagot, Lionel Clément, Éric Villemonte de La Clergerie, and Pierre Boullier. 2006. The Le*fff* 2 syntactic lexicon for French: architecture, acquisition, use. In *Proc. of LREC'06*.

Fabienne Venant. 2004. Polysémie et calcul du sens. In *Actes de JADT 2004*, Leuven, Belgium.

Fabienne Venant. 2007. Une exploration géométrique de la structure sémantique du lexique adjectival franais. *Traitement Automatique des Langues (T.A.L.)*, 47(2).

Vossen, P. 1999. *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*. Kluwer, Dordrecht.

Building a Resource of Ontological and Formal Lexical-Semantic Knowledge

Dennis Spohr
Institut für Linguistik/Romanistik, Universität Stuttgart
Stuttgart, Germany
{dennis.spohr}@ling.uni-stuttgart.de

Abstract

In this paper, we give details on our ongoing efforts to building a lexical resource that provides fine-grained lexical-semantic analyses of French verbs, in addition to a formal organisation of the ontological concepts that are used to describe them. For implementing this information, we make use of technology developed in the context of the Semantic Web, such as the Web Ontology Language OWL, Description Logic reasoners, and the Semantic Web Rule Language SWRL.

We motivate our efforts by comparing our verb analyses to those found in the French EuroWordNet (Vossen, 1998). We further show the necessity of detailed lexical-semantic knowledge – including information about presuppositions and inferences – as well as of ontological type information e.g. on permitted fillers for argument slots, for the successful completion of computational linguistic tasks. Since our resource is primarily intended for computational use, we will outline possible applications of the modelled information.

1 Introduction and Motivation

A number of large-scale lexical resources containing lexical-semantic information have been created and mapped to resources of ontological knowledge, such as WordNet and FrameNet (Fellbaum, 1998; Baker et al., 1998). Although impressive in quantitative terms, what these resources lack to a large extent is an in-depth formal lexical-semantic analysis, e.g. one that provides presuppositional and inferential information. However, this knowledge is required in order to be able to successfully perform automatic reasoning tasks such as the recognition of textual entailment.

While these resources might still serve as a solid basis for starting in-depth lexical-semantic analysis of English lexical items, there is no such resource of comparable quality for French. Although there is a French EuroWordNet (Vossen, 1998), its usability is questionable particularly because it contains a lot of inaccuracies in the description of the verbal domain. As for existing ontological resources, they tend to describe continuants (i.e. entities that are persistent through time, such as objects or organisations) with far more accuracy and detail than occurrents (i.e. entities that have temporal parts, such as events or processes).

In this paper, we will show our approach to building a resource that provides in-depth analyses of the lexical semantics of French verbs and that takes into account also presuppositional and inferential information. The lexical-semantic analyses are tightly linked to concepts in an *ontology of occurrents*. However, despite the references to large-scale lexical-semantic resources, the purpose of this paper is not to present a finished large-scale resource that is capable of directly competing with existing ones, but rather to illustrate ongoing work on principles for modelling a formal combination of syntactic, lexical-semantic, and ontological information in a single resource.

In the following section, we will introduce the necessary background and look at the extent to which existing resources could be used in the creation process. The process itself is described in detail in Section 3.

2 Background and Related Work

2.1 Formalisms

The formalisms that are used for building the resource have been developed in the field of the *Semantic Web*, a research area devoted among others to providing tools and formalisms for assigning meaning to web content (Berners-Lee et al., 2001).

In particular, we make use of the Web Ontology Language OWL (Bechhofer et al., 2004) and the Semantic Web Rule Language SWRL (Horrocks et al., 2004). While these formalisms have been described at length in the relevant literature, we will quickly summarise the main characteristics that are necessary for the comprehension of the paper.

OWL. The Web Ontology Language (Bechhofer et al., 2004) is a formalism based on the Resource Description Framework RDF[1] and can be expressed in XML syntax. Its main building blocks are classes (corresponding to one-place predicates in first-order logic), properties (two-place predicates) and individuals (instances of classes). OWL comes in three sublanguages, which differ wrt. their expressivity: OWL Lite is the least expressive sublanguage and allows for simple class definitions; OWL DL is based on description logic, a decidable fragment of first-order logic, which allows for all OWL constructs but restricts the use of some of them in order to maintain decidability of reasoning; OWL Full is the most expressive sublanguage and imposes no restrictions on the language constructs, however at the cost of decidability. For example, in OWL Full it is possible to express that a class is an instance of another class, which is disallowed in OWL DL.

SWRL. The Semantic Web Rule Language (Horrocks et al., 2004) adds expressivity to OWL in that it allows for the expression of Horn-like rules, i.e. disjunctive rules with at most one positive literal, for example

$$\neg hasFather(x,y) \vee \neg hasBrother(y,z) \vee hasUncle(x,z)$$

which is equivalent to the following rule:

$$hasFather(x,y) \wedge hasBrother(y,z) \rightarrow hasUncle(x,z)$$

SWRL can be expressed directly in OWL syntax – so the resulting documents are still OWL compliant – and the rules can be interpreted and executed by tools such as the Jess® rule engine[2].

2.2 Lexical-semantic resources and ontologies

EuroWordNet. The EuroWordNet project (Vossen, 1998) aimed at providing resources similar to Princeton WordNet (Fellbaum, 1998) for seven European languages, all of which are connected through an interlingual index (ILI) that contains a set of language-independent concepts. The ILI is linked to the so-called EuroWordNet Top Ontology, an upper-ontology-like collection of features that have been designed to describe the lexical-semantic relations in the wordnet. The French version of EuroWordNet contains roughly 8,300 verb senses and 24,500 noun senses, which are organised into 22,745 synonym sets and linked using lexical-semantic relations like hyponymy and meronymy.

In contrast to the scale of the resource in terms of covered senses, the detail of description is generally limited to taxonomic relations between synonym sets and does not include information on argument structure. However, the probably biggest drawback of the French EuroWordNet lies in its inaccuracy and even partial incorrectness, mainly wrt. to the verbal descriptions, both of which probably stem from semi-automatically translating English synsets into French (Dutoit et al., 1998). Therefore, only the noun hierarchy can be considered as a useful starting point for building other lexical resources, whereas the verb hierarchy can only provide a rough sketch as to the interpretation and organisation of the senses.

Other resources. Apart from EuroWordNet, there is no large-scale lexical resource of French that provides qualitatively adequate lexical-semantic analyses. While resources such as FrameNet and VerbNet (Baker et al., 1998; Kipper-Schuler, 2006) exist for English, none of these have been extended to French in a comparable way yet.

2.3 Ontologies

SUMO. Together with DOLCE (see below), the Suggested Upper Merged Ontology (Niles and Pease, 2001) is one of the most widely used ones in the NLP community, among others due to the fact that mappings have been created to Princeton WordNet (Niles and Pease, 2003) and the EuroWordNet ILI (Spohr, 2008a). SUMO comes with MILO, a mid-level ontology, as well as domain ontology extensions, which in total contain 20,000 terms and 70,000 axioms. While originally implemented in SUO-KIF – a formalism intended as first-order language – SUMO has also been translated to OWL Full, with the attempt to preserve as much as possible of the original axiomatisation.

[1] http://www.w3.org/RDF/

[2] http://www.jessrules.com/

Despite its quantitative size and degree of formalisation, SUMO has been criticised primarily wrt. the usability of its axiomatisations, since they are questionable from a modelling perspective (e.g. instances being concepts at the same time and relations being modelled as concepts). Moreover, SUMO seems to lack a clear theoretical basis, as it adopts ideas from different ontological theories (Sonntag et al., 2007).

DOLCE. The Descriptive Ontology for Linguistic and Cognitive Engineering (Gangemi et al., 2003a) is an upper-level ontology that has been designed with a strongly cognitive bias. Its classes and the relations among them have been implemented with the OntoClean methodology (Guarino and Welty, 2002), which gives the resource a formally and theoretically more solid basis than e.g. SUMO. As was mentioned above, DOLCE has also been mapped to Princeton WordNet (Gangemi et al., 2003b).

DOLCE is the first reference module of the WonderWeb library of foundational ontologies, and it has a number of extensions (e.g. an ontology of information objects). In total, DOLCE and its extensions comprise roughly 200 classes and 300 properties, and they are available as OWL versions.

Next to this version of DOLCE, which is called *DOLCE-Lite-Plus*, there exists a version called *DOLCE-Ultralite* (DUL), which uses friendly names for classes and properties and simple class restrictions.[3] For these reasons, and since DUL is – as DOLCE-Lite-Plus – expressed in OWL DL, it provides a solid formal basis for the definition of a lexical-semantic and ontological resource. In total, DUL contains roughly 200 classes and 130 properties.

3 Creation and Computational Use of the Resource

In the following, we will discuss the different steps in the process of building the resource. The manual analysis that precedes the other ones will be omitted here since it has been discussed at length in (Martin et al., to appear). However, it is important to notice that at the end of this analysis step, we have obtained a formal lexical-semantic representation of different senses of a verb that contains information about presuppositions and inferences, in addition to information about sense-specific restrictions on the ontological type of argument slot fillers (e.g. "the subject has to be human" or "a directional prepositional phrase has to be present").

3.1 Interfaces between syntactic, ontological and lexical-semantic knowledge

In this section, we will explain how we model the knowledge obtained from the manual analysis, on the one hand in the form of a kind of "lexical entry" for the different senses, on the other hand in the form of ontological concepts and inference rules.

Ontological argument restrictions in the lexicon. On the basis of the above analysis, we create a small subhierarchy of classes in our lexicon, corresponding to the senses of a verb. The classes are organised hierarchically (as shown in Figure 1) in order to be able to express generalisations that hold for more than one sense, and in order to be able to complete reasoning tasks such as "is the occurrence of *pousser* in this sentence a physical sense of *pousser*?".

- *pousser*
  - *pousser_figurative*
    - *pousser_conceptual*
      - *pousser_conceptual*$_1$
      - *pousser_conceptual*$_2$
      - . . .
    - *pousser_psychological*
      - *pousser_psychological*$_1$
      - . . .
  - *pousser_physical*
    - . . .

Figure 1: Hierarchy of senses of *pousser*

Each of the "leaf classes" (e.g. *pousser_conceptual*$_1$) represents a specific configuration of syntactic and ontological parameters, which are modelled as necessary and sufficient conditions on the definition of the respective class. These axioms are the result of expressing the findings and intuitions wrt. the ontological type of the arguments in the manual analysis step in terms of DOLCE-Ultralite concepts. Figure 2 below shows such a configuration for one of the conceptual senses of *pousser*.

The formalisation is to be interpreted as follows: in order to be classified as an instance of

[3] http://wiki.loa-cnr.it/index.php/LoaWiki:DOLCE-UltraLite

$$
\begin{aligned}
\textit{pousser\_conceptual}_1 \equiv\ & \textit{pousser} \\
& \exists\ \textit{subj}\ (\exists\ \textit{canDenote dul:Organism}) \\
& \forall\ \textit{subj}\ (\exists\ \textit{canDenote dul:Organism}) \\
& \exists\ \textit{obj}\ (\exists\ \textit{canDenote dul:Abstract}) \\
& \forall\ \textit{obj}\ (\exists\ \textit{canDenote dul:Abstract}) \\
& \geq\ 3\ \textit{arg}\ \ \textit{owl:Thing}
\end{aligned}
$$

Figure 2: Axiomatisation of $\textit{pousser\_conceptual}_1$

$\textit{pousser\_conceptual}_1$, it is both necessary and sufficient to be an instance of *pousser*, with a subject that can denote an organism, with a direct object that can denote something abstract, and with at least one more argument (i.e. the number of arguments is at least 3; *owl:Thing* just refers to "any kind of entity"). The predicate *canDenote* used in the formalisation captures the polysemy of the nominal argument, since the classes that represent nouns contain as axioms the ontological concepts they can denote, such as e.g. the class $\textit{faim}_1$ with the axiom $\exists$ *canDenote dul:SocialObjectAttribute*. So in other words, the object part of the example above states that the value of the *obj* property of *pousser* has to be an instance of a class that can denote something abstract (i.e. *dul:Abstract* or any of its subclasses). An example of an instance of this sense of *pousser* is given in sentence 1 below.

(1) *Pierre a poussé ma faim*
Pierre has pushed my hunger
jusqu'à la rage.
to the point of fury.

As can be seen in the figure, we have implemented a very tight link between ontological and syntactic information. In addition to this, we have a further link from the syntax to the ontological and formal lexical-semantic analysis, which will be illustrated in the following.

Inference rules. In order to model the inferences triggered by the syntactic configurations shown above, a formalism that goes beyond the expressivity of OWL is needed, e.g. to be able to make assertions about the entities involved. For this we make use of SWRL rules that contain a specific syntactic configuration in the rule body (e.g. $pousser(?e) \wedge subj(?e, ?x) \wedge obj(?e, ?y)$) and a resulting lexical-semantic output configuration in the rule head (e.g. $PUSHING(?e) \wedge agent(?e, ?x) \wedge VECTOR(?v) \wedge source(?e, ?v) \ldots$). Such a rule is interpreted for example as "if we have an instance *e* of *pousser* with subject *x* and object *y*, then *e* is also an instance of a *PUSHING*-event, with *x* as agent and a vector *v* as source . . . ". Thus, rules implement a crucial link between the syntax on the one hand, and lexical-semantic and ontological knowledge on the other.

Ontology of occurrents. As can be seen in the rule excerpt above, we make use of other ontological concepts in addition to the ones defined in DUL, such as *PUSHING* and *VECTOR*. Taxonomically, they are located below the DUL concepts in the hierarchy, as they represent more specific cases of the ones defined there, e.g. *PUSHING* as a more specific kind of *Action*. The aim of this *ontology of occurrents* is to also assign axiomatic definitions and inference rules to the concepts therein, in order to generalise conceptual properties over specific lexical realisations, i.e. verb senses. This ontology is still work in progress, and since we intend to design it according to the OntoClean principles, we have used DUL to sort of "prestructure" our concept hierarchy. However, defining essential and rigid properties or identity criteria of occurrents is an entire topic of its own, and will be part of future research.

3.2 Computational use

In the following paragraphs, we will briefly explain how the resource can be used for automatic word-sense disambiguation and calculation of inferences.

Disambiguation of verbs in context. The primary factors that can be used for the disambiguation of verb senses is the ontological type of the syntactic arguments. As was shown in Figure 2 above, these are modelled as necessary and sufficient conditions in the respective class definition.

For disambiguating a sentence like the one in (1), we would first assume syntactic input that provides at least information about the predicate (*pousser*), its arguments (e.g. *subject = Pierre*, *object = faim* etc.) as well as the tense used (in this case *passé composé*). The fillers of the argument slots are then looked up in selectional preference lists of the respective predicate (Spohr, 2008b), which contain information about the most probable ontological types per argument slot, and the sense of the noun whose ontological type scores highest is selected and asserted in the resource.

For example, after having selected a sense of *faim*, we assert an individual *x* as an instance of the class $faim_1$ and link it to the predicate by means of the *subj* relation, i.e. *subj(x)*. On the basis of (i) the syntactic configuration, (ii) the necessary and sufficient conditions in the classes for *pousser*, and (iii) the sense selection for the nominal arguments, a description logic reasoner (e.g. Pellet; (Sirin et al., 2007)) is run and infers a sense of the predicate *pousser* that has been used in this particular sentence.

Calculation of inferences. Once a sense has been selected by the reasoner, the system can execute the SWRL rules that have been defined for the respective senses in order to calculate the inferences that are licensed on the basis of the previous sense selection. As was mentioned in Section 3.1, the appropriateness of a rule is further determined by the syntactic context in which the verbal predicate has been used, and which has to match with the one stated in the rule body. The new statements that result from the rule execution are then asserted in the resource. They represent the logical form of the input sentence, based on the ontologically enriched manual lexical-semantic analysis. This information, which is directly encoded in OWL, can then further be made available to other applications.

3.3 Current state of and future plans

As was already mentioned in the introduction, the resource is not in a state of being applied to real-life tasks. The lexical-semantic analysis of verbs as well as the definition of the ontology are still work in progress, and the current size in terms of senses covered is very small. Nonetheless, sample tests on selected corpus sentences have been able to serve as a proof of concept for the rich formalisation of verbs as being done in our project. Therefore, with the formal principles of modelling lexical-semantic and ontological information defined, we intend to tackle the quantitative size of the resource in the future.

4 Conclusion

In this paper, we have provided details on the process of building a lexical resource of French that contains a high level of detail wrt. the lexical-semantic and ontological analysis of the verbal domain, with focus on the interplay between syntactic, lexical-semantic and ontological information. In addition to motivating the necessity of a high level of detail in the modelling of this knowledge, we have presented ongoing efforts in designing an ontology of occurrents and, finally, outlined the potential of the resulting resource for use in computational scenarios.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the joint COLING/ACL 1998*, Montreal, Canada.

Sean Bechhofer, Frank van Harmelen, Jim Hendler, Ian Horrocks, Deborah L. McGuinness, Peter F. Patel-Schneider, and Lynn Andrea Stein. 2004. OWL Web Ontology Language Reference. W3C Recommendation. `http://www.w3.org/TR/owl-ref/`.

Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The Semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, 284(5):34–43.

Dominique Dutoit, Laurent Catherin, and Andreas Wagner. 1998. Specification of German & French WNs. EuroWordNet (LE-8328) Deliverable: 2D002.

Christiane Fellbaum, editor. 1998. *WordNet – An Electronic Lexical Database*. MIT Press, Cambridge, MA, USA.

Aldo Gangemi, Nicola Guarino, Claudio Masolo, and Alessandro Oltramari. 2003a. Sweetening WordNet with DOLCE. *AI Magazine*, 24(3):13–24.

Aldo Gangemi, Roberto Navigli, and Paola Velardi. 2003b. The OntoWordNet Project: extension and axiomatization of conceptual relations in WordNet. In *Proceedings of ODBASE*, Catania, Italy. Springer.

Nicola Guarino and Christopher Welty. 2002. Evaluating Ontological Decisions with OntoClean. *Communications of the ACM*, 45(2):61–65.

Ian Horrocks, Peter F. Patel-Schneider, Harold Boley, Said Tabet, Benjamin Grosof, and Mike Dean. 2004. SWRL: A Semantic Web Rule Language Combining OWL and RuleML. W3C Member Submission. `http://www.w3.org/Submission/SWRL/`.

Karin Kipper-Schuler. 2006. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, Computer and Information Science Dept., University of Pennsylvania, Philadelphia.

Fabienne Martin, Dennis Spohr, and Achim Stein. (to appear). Representing a Resource of Formal

Lexical-Semantic Descriptions in the Web Ontology Language. *GSCL Forum – Special Issue on Lexical-Semantic and Ontological Resources*.

Ian Niles and Adam Pease. 2001. Towards a Standard Upper Ontology. In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS 2001)*, Ogunquit, ME.

Ian Niles and Adam Pease. 2003. Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In *Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE '03)*, Las Vegas, NV.

Evren Sirin, Bijan Parsia, Bernardo Cuenca Grau, Aditya Kalyanpur, and Yarden Katz. 2007. Pellet: A practical OWL-DL reasoner. *Journal of Web Semantics*, 5(2).

Daniel Sonntag, Ralf Engel, Gerd Herzog, Alexander Pfalzgraf, Norbert Pfleger, Massimo Romanelli, and Norbert Reithinger. 2007. SmartWeb Handheld – Multimodal Interaction with Ontological Knowledge Bases and Semantic Web Services. In Thomas S. Huang, Anton Nijholt, Maja Pantic, and Alex Pentland, editors, *Artificial Intelligence for Human Computing*, volume 4451 of *Lecture Notes in Artificial Intelligence*, pages 272–295. Springer, Heidelberg.

Dennis Spohr. 2008a. A General Methodology for Mapping EuroWordNets to the Suggested Upper Merged Ontology. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)*, Marrakech, Morocco.

Dennis Spohr. 2008b. Extraction of Selectional Preferences for French using a Mapping from EuroWordNet to the Suggested Upper Merged Ontology. In *Proceedings of the 4th Global WordNet Conference*, Szeged, Hungary.

Piek Vossen, editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers.