Matthew Whelpton

From human-oriented dictionaries to computer-oriented lexical resources – trying to pin down words

1 Introduction

Dictionaries are one of the most familiar linguistic resources to which an ordinary native speaker of a (standardised) language is likely to have access; indeed the process of dictionary creation has served a crucial role historically in the standardisation of European languages and represents an important activity in the creation and maintainance of standards today. In the age of information technology, the successors of the paper dictionary have continued to exert great influence, playing a central role in the field of natural language processing (NLP), supplying text and speech processors with essential information on word form, category, and meaning in activities as diverse as grammatical parsing, information extraction, and machine translation.

Although some of these NLP resources are essentially electronic versions of paper dictionaries, the demands of computer applications are in many ways much greater than those of human users: computers require that information be presented in a manageable format for algorithmic manipulation (e.g. in a relational database where each piece of information is classified and linked explicitly to others) and that the information itself be systematic and absolutely explicit. The

Orð og tunga 14 (2012), 1–18. © Stofnun Árna Magnússonar í íslenskum fræðum,
Reykjavík.

human user of a dictionary brings to the dictionary a host of implicit knowledge and cognitive skills that aid in dictionary use, in particular a range of assumed world and cultural knowledge and the common sense ability to deploy that knowledge appropriately in interpreting the dictionary entry. A computer on the other hand comes to the electronic "dictionary" knowing nothing at all in advance and has only the "sense" that is represented by the processing algorithms available to it. To be effective, a computational lexical resource must therefore represent the relevant information in a fully explicit and systematic way, encoding information that to human users would seem obvious and unnecessary.

This paper focuses on the meaning of words (lexical semantics) and some important computational resources that have been developed to make lexical semantic information available to computers for a variety of NLP tasks. It is intended as a survey article, describing the properties of three major lexical semantic resources (WordNet, DanNet and SALDO), which provide a frame of reference for current work on two Icelandic projects, reported in this volume. Anna Björk Nikulásdóttir reports on a project developing semi-automatic means for extracting information on lexical semantic relations from text corpora (*Íslenskur merkingarbrunnur*, cf Nikulásdóttir & Whelpton 2009, 2010a, 2010b); Jón Hilmar Jónsson reports on a project which is manually developing a network of lexical sense relations (*Íslenskt orðanet*, cf Jónsson 2008, 2009a, 2009b, 2009c; Úlfarsdóttir 2006).

Section 2 introduces one of the oldest and most influential lexical semantic resources, the Princeton WordNet, and reviews some of the central lexical semantic relations around which the resource is structured: synonymy, hyponymy, meronymy, antonymy, and troponymy. Section 3 introduces DanNet, a lexical semantic resource for Danish, conforming to the international standards of wordnet development; a number of challenges faced by DanNet are reviewed, in particular the challenge of converting traditional dictionary information into a computer-tractable form and the challenge of addressing deficiencies in the relation set of the original Princeton WordNet. Section 4 introduces SALDO, a morphological and lexical semantic database for Swedish, organised on radically different lines to the wordnets, as it attempts to model the degree of centrality of lexicalised concepts in Swedish rather than encoding specific lexical semantic relations between them. Section 5 concludes this survey and points on to the papers introducing the two Icelandic resources.

2 Wordnet

2.1 Background

The Princeton WordNet[1] (Miller 1995, Fellbaum 1998) is a lexical database of English constructed to represent word sense relations. It was developed under the direction of the psychologist, George Miller, and its original aims were explicitly psycholinguistic in nature. As Miller (1998a: xv) explains, the original WordNet project included two psycholinguistic hypotheses: (i) the separability hypothesis "that the lexical component of language can be isolated and studied in its own right", i.e. that the mental lexicon has a distinct organisation and identity from the combinatorial systems of grammar and the expressive system of phonology; (ii) the patterning hypothesis "that people could not master and have readily available all the lexical knowledge needed to use a natural language unless they could take advantage of systematic patterns and relations among the meanings that words can be used to express". The WordNet project was always, however, a project in computational psycholinguistics and another important hypothesis is related to the issue of computational tractability and scalability: the comprehensiveness hypothesis "that computational linguistics, if it were ever to process natural languages as people do, would need to have available a store of lexical knowledge as extensive as people have".

The challenge was to decide how a comprehensive lexical semantic database for computation might be structured. One of the earliest and most influential forms of lexical semantic analysis was componential analysis, i.e. the analysis of the meaning of a word like *man* as HUMAN + MALE + ADULT. However, by 1985 it was becoming clear that there was no easily identifiable list of "conceptual atoms" and following contemporary developments in the field, Miller adopted the idea that word meaning could be characterised in terms of systematic relationships to other words (Miller 1998a: xvi): for instance, *table* could be related to *furniture* by an IS-A-KIND-OF relation: this would not make the claim that *furniture* was a component of the meaning of *table*, merely that there was a systematic relationship of a particular kind between the meaning of *table* (whatever that was) and the meaning of *furniture* (whatever that was).

[1] http://wordnet.princeton.edu/

The WordNet database is therefore structured in terms of a number of sense relations which appear to be psychologically relevant in the characterisation of word meaning. Further the database is organised around part of speech, on the basis of evidence that word storage in the mental lexicon is sensitive to part of speech. The current discussion relates to WordNet 3.0, which contains around 155,000 word forms (unique strings), of which just over 115,000 are nouns; the rest are verbs, adjectives and adverbs. In the following sections, we will review some of the main lexical sense relations that determine the organisation of WordNet.

2.2 Synonyms and synsets

The basic building block of WordNet is the synset or "set of synonyms" (Icelandic: *samheiti*; Greek: *syn* 'same' + *onyma* 'name'). In WordNet, synonymy is defined as having the same sense in a particular context.

(1) the nurse gave him a flu shot/injection/*pellet

- synset: = {shot, injection}

(2) the shot/pellet/*injection buzzed past his ear

- synset: = {shot, pellet}

Sentence (1) identifies a particular "sense", glossed in WordNet 3.0 as "the act of putting a liquid into the body by means of a syringe". This sense can be expressed by *shot* and by *injection* but not by *pellet*; *shot* and *injection* are therefore synonyms and form a synset. Sentence (2) identifies another "sense", glossed in WordNet 3.0 as "a solid missile discharged from a firearm". This second sense can be expressed by *shot* and by *pellet* but not by *injection*. This illustrates two important points about the organisation of WordNet.

First, the basic building block of the network is in fact a particular sense or concept; that sense can be expressed by one or more different word forms. This is very different from a traditional dictionary, whose basic building block is the word itself: the forms *shot* and *injection* would be listed separately in a traditional dictionary and each would be listed with the relevant sense as part of its entry. In WordNet, the sense itself represents a unique entry and the forms associated with it are grouped in a synset. WordNet is sense-oriented; a traditional dictionary is word-oriented.

Second, WordNet does not distinguish between polysemy and homonymy. Polysemy is when a single word (lexeme) is associated with more than one sense. The word *shot* would be a good example, as it can express the sense associated with either Sentence (1) or Sentence (2). Homonymy is when two different words (lexemes) happen to have the same form: the classic example of this in English is the word-form *bank*, which can refer either to the side of a river or to a particular kind of financial institution. The intuition here is that the two senses are completely unrelated and that it is no more than a historical coincidence that they are expressed by the same word-form. WordNet remains completely agnostic on this distinction between polysemy and homonymy because its basic building block is the sense, each sense having one entry and being associated with a set of one or more word-forms which can express that sense in a certain context, i.e. the synset. It is the synset in WordNet which stands in sense-relations to other synsets and we will now review some of the main relations around which the database is structured.

2.3 Hyponymy~Hypernymy

Hyponymy is also known as the IS_A relation, typically the subkind relation. For instance, *mare* is a hyponym (Icelandic *undirheiti*; Greek: *hypo* 'under' + *onyma* 'name') of *horse*; and conversely, *horse* is a hypernym (Icelandic *yfirheiti*; Greek: *hyper* 'over' + *onyma* 'name') of *mare*, because *a mare is a (kind of) horse*. Hyponymy naturally creates hierarchies:

(3) a mare IS_A horse IS_A mammal IS_A animal

This is especially true of natural kinds, for which the hyponymy hierarchy can become quite articulated.

According to the hierarchy in Figure 1, both *mare* and *stallion* are hyponyms of *horse*, i.e. they are co-hyponyms; *animal* is the root of this hierarchy. In fact, WordNet has a considerably more articulated hierarchy than is shown here, with much greater depth. For instance, *stallion* is in fact a co-hyponym with *gelding*: both are *male horses* but the latter is castrated and the former not: this means that there is a lexical gap in the hierarchy because there is no specialised lexeme in

English for a male horse which covers both castrated and uncastrated varieties. WordNet sometimes fills this gap with multiword expressions: in this case, the hypernym for *stallion* and *gelding* is given as *male horse*, and it is *male horse* which is the co-hyponym of *mare*. At the top of the hierarchy, are a number of abstract terms which root the tree: so the top of the hierarchy for *mare* is not in fact *animal* but *entity* (*entity* is in fact at the root of all noun hyponymy hierarchies).

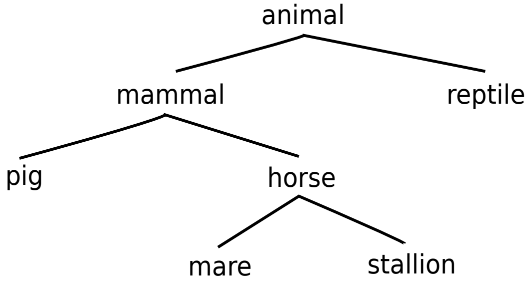

Figure 1. *A (partial) hyponymy hierarchy*

Such hierarchies are therefore lexical ontologies, i.e. classifications of the kinds of things that can be referred to in the language. Lexical ontologies must therefore confront the tension between scientific ontologies and folk ontologies, i.e. between the classification established as objective by the natural sciences and the classification established by popular usage and belief. A wordnet is often a compromise between these two and not always a consistent one. So, for instance, WordNet conforms to the scientific ontology for *whale*: it is given as a hyponym of *mammal* and is glossed as "any of the larger cetacean mammals having a streamlined body and breathing through a blowhole on the head". However, *tomato* is given as a hyponym of *vegetable* despite biologically being a *fruit*; nevertheless the gloss acknowledges the scientific classification and hints at the reason for the *vegetable*-classification: "mildly acid red or yellow pulpy fruit eaten as a vegetable", i.e. the hyponym relation is assigned on the basis of the use that is made of the entity, rather than its biological status – this is a functional hyponym not a nature-kind hyponym. It is important to stress here the difference between WordNet and a traditional dictionary: the main semantic information is the lexical semantic relation (hyponymy) and not the gloss; a computer using WordNet to build a semantic representation will treat *tomato* as a *vegetable*.

2.4 Meronymy~holonymy

Meronymy is the part-relation. For instance, *nose* is a meronym (Icelandic: *hlutheiti*; Greek *meros* 'part' + *onyma* 'name') of *face*; conversely, *face* is the holonym (Icelandic: *heildheiti*; Greek *holos* 'whole' + *onyma* 'name') of *nose*. The meronymy relation raises the important issue of modality: whether the relation actually must hold or merely can hold. With natural-kind hyponymy, the relation is necessary: every mare is a horse and no mare is not a horse. With meronymy, the relation is often one of possibility rather than necessity. So, for instance, meronyms of *face* include *beard*, which is only possible on some faces and never necessary. This shows that meronymy in WordNet is not even associated with typicality, as *beard* is not a typical part of *face* in general: women´s faces don´t typically have beards and even for men´s faces beards would only be typical in some cultures.

2.5 Antonymy

Antonymy is the relation of oppositeness and is important for the classification of adjectives.

(4) If the water is hot, then the water is not cold, and vice versa.

Hot is the antonym (Icelandic *andheiti*; Greek *anta* 'opposite' + *onyma* 'name') of *cold* and vice versa. It turns out, however, that not every adjective has an antonym, even when it is a synonym for an adjective that does. For instance, *torrid* is a synonym of *hot* (*hot/torrid weather*); *hot* is an antonym of *cold*; yet *torrid* is not an antonym of *cold*. Adjective networks in WordNet therefore often have a "bicycle" structure (cf. Figure 2).

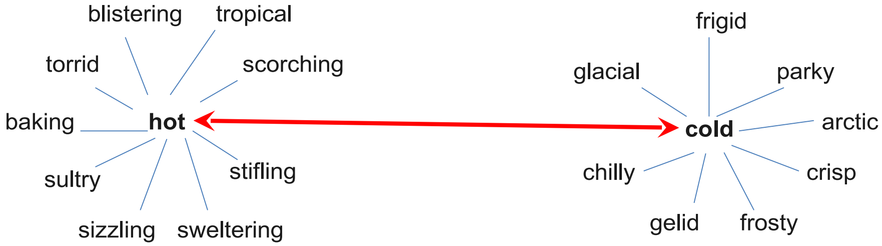

Figure 2. *The "bicycle" structure of antonymy in WordNet*

2.6 Troponymy

Verbs in WordNet 3.0 are largely organised in terms of the hyponymy relation, just like nouns. For instance, *to amble* is listed as a hyponym of *to walk*. Fellbaum (1998: 79) points out that verb hyponymy is not straightforwardly equivalent to noun hyponymy, however. She cites Lyons (1977: 294) for the observation that verbs do not fit naturally into the hyponymy paraphrase for nouns, without nominalisation:

(5) x is a (kind of) y

(6) ?To amble is (*a kind of) to walk

(7) ?Ambling is (a kind of) walking

Even Sentence (7) would not necessarily be accepted by all native speakers. However, if the manner aspect of the relation is emphasised, then the paraphrase works much more effectively:

(8) To amble is to walk in an ambling manner.

(9) To V_1 is to V_2 in a particular manner/way.

Fellbaum & Miller (1990) dub such manner hyponyms, **troponyms**. As Fellbaum goes on to observe, however, the complexity of the relation between a verb and its hypernym is much richer and more complex than this paraphrase suggests and I will leave this issue here.

We will now turn to another major wordnet resource which explicitly addresses some of the problems with the original Princeton WordNet – DanNet, a wordnet for Danish.

3 DanNet

DanNet[2] is a lexical semantic database for Danish, conforming to international wordnet standards; it was developed from a monolingual Danish dictionary (*Den Danske Ordbog*, DDO) as a collaborative project between SIMPLE-DK (Centre for Language Technology at the University of Copenhagen) and the publisher of DDO (the Society for Danish Language and Literature, Danish Ministry of Culture). As of 2009, it con-

[2] See www.wordnet.dk for download and general information; see http://andreord.dk/ord/ to browse DanNet

tained 50,000 synsets. Pedersen et al. (2009) report on the problems of converting a human-use dictionary to a lexical semantic database and the limitations of using the classical lexical semantic relations of the Princeton WordNet. A good example of both problems comes in the discussion of the word *butik* ‚shop‘:

> For example all hyponyms of *butik* (shop) inherit the involved agent *handlende* (shopkeeper). Thus, the DanNet editor is prompted to identify the involved agent of the more restricted hyponym: that the shopkeeper of a pharmacy is a pharmacist, the shopkeeper of a bakery is a baker and so on. Such information is only rarely specified in DDO definitions (although sometimes provided implicitly as examples of word formation), but this information is seen as highly relevant in a wordnet.
>
> (Pedersen et al. 2009: 273)

The DanNet developers address two issues in this passage.

The first issue is the gap between a traditional dictionary and an NLP resource: the traditional dictionary entries of DDO leave implicit the relation between a pharmacy and a pharmacist, a bakery and a baker, and so on, because the inference of such a relation can be left to the world knowledge and common sense of a human user; a computer however will not automatically make such inferences. The DanNet developers therefore state the shopkeeper for every subkind of shop, where such an entity is lexicalised. The manual process of adding such information is streamlined by exploiting the inheritance relation: an algorithm is set to prompt the developer for "missing information" that can be inferred on the basis of established relations; so, if a relation is explicitly stated for a hypernym, then it is likely that all hyponyms will have a specific equivalent of this relation: if all shops have a shopkeeper (involved agent) and a bakery is shop (hyponymy), then a bakery will have a shopkeeper, for whom the language may well have a specialised lexical item.

The second issue implied in the quoted passage concerns the limits of the original WordNet relation set. Notice that the information being added to DanNet here (involved agent) is not a relation in the original WordNet: speakers of a language implicitly understand a systematic relation between an entity and its typical owner or user; this information is therefore added systematically to DanNet. In fact, the developers of DanNet extend the classical relations of WordNet, following the work of the computational semanticist, Pustejovsky, whose *Generative Lexicon* (Pustejovsky 1995) proposes four “qualia roles” that form part of the representation of noun meaning:

(10) Formal: contrastive relation to other objects (affecting inheritance relations)

(11) Agentive: origin

(12) Constitutive: how an object is made up (its parts and organisation)

(13) Telic: purpose or function

The range of relations used in DanNet are shown in Figure 3 (based on Pedersen et al. 2009, Figure 12).

FORMAL	**AGENTIVE**	**CONSTITUTIVE**	**TELIC**
has_hypernym	made_by	has_holo_made_of	used_for
has_hyponym		has_mero_made_of	used_for_object
is_a_way_of (troponym)		has_holo_part	role_agent
		has_mero_part	role_patient
		has_holo_member	
		has_mero_member	
		has_holo_location	
		has_mero_location	
		concerns	
		involved_agent	
		involved_patient	
		involved_instrument	

Figure 3. *Relations in DanNet*

We have already seen how involved_agent can be used to link senses. Notice also the more fine-grained range of distinctions that have been added to the meronymy~holonymy relation. As we saw earlier, the traditional meronymy relation relates to typical parts, so a cabin might have a roof as a part. However, a log-cabin will have logs as a part, in the sense that it is made of logs, a slightly different kind of part relation. Similarly, a congregation may have a minister as a “part” in the sense that the minister is a member of the congregation; and England has London as a “part” in the sense that it is a location within the larger location. These distinctions have important implications for inferencing: a human user will reflexively accommodate them but a computer must be provided with the information explicitly and systematically.

These changes to the original WordNet relation set are essentially extensions and elaborations. However, DanNet also addresses a fundamental problem with the original relation which provides the backbone to Princeton WordNet: hyponymy. To be useful for inferencing, the hyponymy relation should: (i) hold of subkinds (a dog is a kind of animal; a cat is a kind of animal), where (ii) co-hyponyms are mutually exclusive (if something is a dog then it is not a cat and vice versa). Pedersen et al. (2009: 277) point out that the traditional use of hyponymy covers a broader range of relations than this, in a way which is problematic for NLP applications: so-called, **hyponymy overload**. Consider the following examples:

(14) *oliemaleri* 'oil painting', *blomstermaleri* 'flower painting', *fidusmaleri* 'pseudo-art', *akvarel* 'water colour', *marinebillede* 'seascape', *klatmaleri* 'daub'

Each of these terms is a hyponym of *maleri* 'painting'. However, if one thinks in terms of mutually-exclusive subkinds, two candidate pairs emerge:

(15) *oliemaleri* 'oil painting' vs *akvarel* ´water colour'

- subkinds of painting distinguished by the paint used

(16) *blomstermaleri* 'flower painting' vs *marinebillede* 'seascape'

- subkinds of painting distinguished by the subject depicted

It is perfectly possible to have an oil painting which is also a flower painting – and it is perfectly possible that that item is also a "daub". Notice that the two mutually exclusive pairs are exclusive along a particular dimension: paintings classified by the kind of paint used; paintings classified by the subject depicted. Pedersen et al. therefore adopt a proposal by Huang et al. (2008) which allows hyponyms defined by a particular dimension of description to be grouped together: *oliemaleri* 'oil painting' and *akvarel* ´water colour' are said to be **paranyms**, terms associated with the same dimension of description. The paranym relation allows co-hyponyms to be clustered into mutually-exclusive subsets.

The terms *fidusmaleri* 'pseudo-art' and *klatmaleri* 'daub' remain problems, however, because any painting can in fact be a daub: this term does not really describe a subkind as such but rather a subjective evaluation of an item; a daub may indeed be a kind of painting but any painting can be termed a daub if the speaker assesses its quality to be at a certain level. Terms like *klatmaleri* 'daub' therefore cut across the hyponyms of *maleri* 'painting': they are orthogonal to the classification. DanNet therefore allows the hyponym relation to be tagged with a feature ORTHO which indicates that the term represents an evaluation that can apply to any "co-hyponym" of the term.

Even this relatively brief discussion illustrates well the challenges that face the construction of a wordnet which is to be sufficiently richly and systematically elaborated to be used in advanced NLP applications. We will now turn to a third resource, quite unlike the two that we have reviewed so far, which is developed around a very different aspect of sense organisation.

4 SALDO

SALDO[3] is "a free full-scale modern Swedish semantic and morphological lexical resource intended primarily for use in language technology applications" (Borin & Forsberg 2009: 7). It is based on a much looser associative relation that we typically find in wordnets – especially as the relation is not sensitive to part of speech. In fact there is only one obligatory relation in SALDO (mother) and one optional relation (father). The mother will be a more central concept, i.e. semantically and/or morphologically less complex, probably more frequent, stylistically more unmarked, and acquired earlier in first and second language acquisition. The father will be a differentiating term (often a domain-specifier) (Borin & Forsberg 2009: 7f). For example, the noun *sol* 'sun' has as a mother the verb *lysa* 'shine'; the father of *sol* 'sun' is *himmel* 'sky', i.e. the domain or context in which the shining takes place (Borin & Forsberg 2009: 10). The simplest way to grasp the essential intuition upon which SALDO´s semantic classification is built is to imagine how you would define a word if you were dealing with someone with very limited vocabulary. You might attempt to indicate what the sun was by saying that it was the thing in the sky which shines. Shining is the most salient characteristic of the sun and the sky is the place that one needs to look to find it.

[3] http://spraakbanken.gu.se/eng/saldo

Where a wordnet has a hyponymy hierarchy, SALDO has a centrality hierarchy based on motherhood. So, *sol* 'sun' has a number of siblings that share the same mother, *lysa* 'shine':

(17) 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*

Some of these are full siblings that also share the same father, *himmel* 'sky':

(18) *comet, moon, star*

At the core of SALDO are the roots of these hierarchies: 51 lexical primitives on which all other items depend (Borin & Forsberg 2009: 9, their Figure 1).

(19) *all* 'all', *annan* 'other', *använda* 'use', *att* 'that', *bara* 'only', *bra* 'good', *genom* 'through', *den* 'it', *fort* 'fast', *framme* 'arrived', *färg* 'color', *för2* 'for', *förbi* 'gone/past', *före* 'before', *en2* 'a/one', *göra* 'do', *ha* 'have', *hur* 'how', *hända* 'happen', *i2* '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', *om2* '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'

This way of looking at the semantic relations between words is obviously very different from the wordnets. One striking difference, when considering the roots of the hierarchies, is that in WordNet we find abstract terms like "entity" which are added to draw together the forest of more lexically articulated and conceptually substantive trees beneath, whereas in SALDO we find highly frequent and often substantive terms such as "light" and "warm" and "say". This is because SALDO is driven to a large extent by conceptual saliency and centrality and to that extent it is reminiscent of the core vocabulary in Wierzbicka and Goddard's Natural Semantic Metalanguage (NSM)[4] (Wierzbicka 1996; Goddard 2008), with which Borin & Forsberg (2009: 8f) compare their work.

[4] http://www.une.edu.au/bcss/linguistics/nsm/semantics-in-brief.php

nsm was developed in support of a program of “reductive paraphrase“, in which the meaning of complex expressions is given using simple terms. The simple terms express irreducible fundamental concepts which have exponents in all languages. nsm is therefore intended as a kind of universal conceptual interlingua. Like saldo, the primitive terms of nsm are descriptively substantive and relatively high frequency; of the 51 lexical primitives of saldo and 61 semantic primitives of nsm, there are 17 shared terms, including: good, do, think, want, when, where, not, if. It proves to be significant, however, that nsm aims at a set of universal paraphrase terms which can be used for building sense definitions of lexical items in all languages, whereas saldo (saldo *Instruktion*, p. 10) aims at ”så homogena och intuitivt tilltalande horisontella lexemklasser som möjligt”[5] for Swedish, in which the small lexical groupings emerge organically from the internal properties of the Swedish vocabulary system, rather than being imposed externally from a preconceived typology (”Större strukturer i lexikonet växer fram organiskt, utan kontroll ’uppifrån’ eller ’utifrån’.”[6]) One nice example of this is the relative centrality of comparative *like*. It is a central term in nsm because the relation of comparison is understood as a primitive conceptual relation. It is, however, four steps from the core of saldo. Another example discussed by Borin & Forsberg (2009: 9) concerns antonymy. In saldo, antonyms can be related by a mother-child relation: in saldo, the mother of *dålig* ‘bad’ is *bra* ‘good’; the father of *dålig* ‘bad’ is *motsats* ‘opposite’. So in saldo, *bra* ‘good’ is treated as a primitive concept and *dålig* ‘bad’ derived with respect to it by opposition or contrast; in nsm, *good* and *bad* are treated as primitive evaluative terms which can be used to paraphrase classes of more complex expressions.

Although saldo and nsm differ radically from the wordnets in the kinds of terms that we find at the roots of their hierarchies, they nevertheless show significant differences related to their contrasting attitudes to universal conceptual structure versus language-particular lexical organisation.

[5] "a horizontal grouping of lexemes which is as homogeneous and intuitively appealing as possible" (my translation).

[6] "Larger structures in the lexicon develop organically, without imposition ‘from above’ or ‘from outside’.” (my translation).

5 Conclusion

This paper began by setting up a contrast between the demands placed on the traditional dictionary for human use and the lexical resource for NLP use. As the human user brings a vast amount of world and cultural knowledge to the task of dictionary use, supplemented by robust common sense reasoning skills, the dictionary creator can assume all sorts of semantic information as understood; as a computer brings nothing to the lexical semantic resource, independent of the algorithms it has been programmed with, the creator of an NLP resource must include a rich set of information in a systematic and explicit manner and in a format which is suitable for algorithmic manipulation. It is not surprising then to find the creators of each of these resources treading the delicate line between the modelling of linguistic organisation and of conceptual organisation.

As the final discussion concerning the differences between SALDO and NSM show, there is also a tension between potentially universal properties of linguistic organisation and the idiosyncratic properties of particular languages. NSM aims at a universal paraphrase language for the conceptual primitives underlying lexical organisation in human languages; SALDO is emphatically monolingual in its approach. The tension between universal and particular is built into WordNet: at the root of the WordNet hierarchies are abstract terms such as "entity" which serve to root the forest of hyponymy hierarchies beneath them and which are likely to be shared by wordnets for other languages; but the bulk of the relational information represented is potentially idiosyncratic and reflected in the distribution of lexical gaps and the elaboration of hyponymy distinctions further down the tree. Nevertheless, the Princeton WordNet was developed as an analysis of English lexical semantic organisation and as such is a monolingual resource. Similarly, DanNet was explicitly monolingual in its methodology, basing its structure on a monolingual corpus-based dictionary, rather than translation from the Princeton WordNet. This monolingual emphasis is shared by both Icelandic resources presented in this volume, which seek to characterise the lexical semantic organisation of Icelandic in its own terms, without importing a structure from resources developed for other languages (e.g. by translation of WordNet or DanNet).

Another important characteristic shared by all three of the resources surveyed here is that they are manually constructed. This places an enormous burden on project resources in terms of time, money and manpower. For a small community such as Iceland, this is a critical issue (Rögnvaldsson et al. 2009). In this respect, the two Icelandic projects differ in approach but both provide reason for cautious optimism. Jón Hilmar Jónsson´s *Íslenskt orðanet* adopts a manual methodology and yet despite the practical constraints that this imposes has achieved impressive progress in developing a monolingual sense-oriented resource for Icelandic; Anna Björk Nikulásdóttir´s *Íslenskur merkingarbrunnur* is developing a variety of semi-automatic methods to extract lexical semantic relations from text corpora, which is currently showing promising results. It is to be hoped that the contrasting methodologies (semi-automatic and manual) will prove to be complementary and allow the two projects to collaborate effectively in the development of robust lexical semantic resources for Icelandic.

References

Baroni, M., Murphy, B., Barbu, E., & Poesio, M. 2010. Strudel: A Corpus-Based Semantic Model Based on Properties and Types. *Cognitive Science* 34: 222–254.

Borin, L., & Forsberg, M. 2009. All in the Family: A Comparison of SALDO and WordNet. In: B.S. Pedersen, A. Braasch, S. Nimb, and R. Vatvedt Fjeld, (Eds.). *Proceedings of the Workshop "Wordnets and other Lexical Semantic Resources - between Lexical Semantics, Lexicography, Terminology and Formal Ontologies" at 17th Nordic Conference on Computational Linguistics (NODALIDA) 14-16th May 2009*, NEALT Proceedings Series Volume 7, pp. 7–12. Odense, Denmark.

Cruse, D. A. 1991. *Lexical semantics*. Cambridge: Cambridge University Press.

Cruse, D. A. 2002. Hyponymy and its varieties. In: R. Green, C. A. Bean, & S. H. Myaeng (Eds.). *The semantics of relationships: An interdisciplinary perspective, information science and knowledge management*, pp. 2–21. Springer.

DDO = *Den Danske Ordbog* (´The Danish dictionary´) 1–6. 2003–5. Eds.: E. Hjorth, K. Kristensen, et al. Copenhagen: Gyldendal and Society for Danish Language and Literature.

Fellbaum, C. (Ed.) 1998a. *WordNet. An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fellbaum, C. 1998b. Introduction. In: C. Fellbaum (Ed.). *WordNet. An Electronic Lexical Database*, pp. 1–19. MIT Press, Cambridge, MA.

Fellbaum, C., & Miller, G.A. 1990. Folk psychology or semantic entailment? A reply to Rips and Conrad. *The Psychological Review* 97: 565–570.

Huang, C., Hsiao, P., Su, I., & Ke, X. 2008. Paranymy: Enriching ontological knowledge in WordNets. In*: Proceedings of the fourth global WordNet conference*, pp. 221–228. Szeged, Hungary.

Goddard, C. (Ed.) 2008. *Cross-Linguistic Semantics*. Amsterdam: John Benjamins.

Jónsson, J.H. 2008. Í áttina að samfelldri orðabók – nokkrir megindrættir í Íslensku orðaneti. *Orð og tunga* 10: 29–45.

Jónsson, J.H. 2009a. Ordforbindelser: Grunnelementer i ordboken? *LexicoNordica* 16: 161–179.

Jónsson, J.H. 2009b. Lemmatisation of Multi-word Lexical Units: Motivation and Benefits. In: H. Bergenholtz, S. Nielsen & S. Tarp, (Eds.). *Lexicography at a Crossroads. Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow*, pp. 165–194. Bern: Peter Lang.

Jónsson, J.H. 2009c. Lexicographic description: An onomasiological approach on the basis of phraseology. In: S. Nielsen & S. Tarp, (Eds.). *Lexicography in the 21st Century. In honour of Henning Bergenholtz,* pp. 257–280. Amsterdam: John Benjamins Publishing Company.

Miller, G.A. 1998a. Foreword. In: C. Fellbaum (Ed.). *WordNet. An Electronic Lexical Database*, pp. xv–xxii. Cambridge, MA: MIT Press.

Miller, G.A. 1998b. Nouns in Wordnet. In: C. Fellbaum (Ed.). *WordNet. An Electronic Lexical Database*, pp. 23–46. Cambridge, MA: MIT Press.

Nikulásdóttir, A. & Whelpton, M. 2010a. Lexicon Acquisition through Noun Clustering. *LexicoNordica* 17: 141–161.

Nikulásdóttir, A. & Whelpton, M. 2010b. Extraction of Semantic Relations as a Basis for a Future Semantic Database for Icelandic. In: *Proceedings of 7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages (Workshop 22 of 7th Language Resources and Evaluation Conference)*, pp. 33–39. Valletta, Malta.

Nikulásdóttir, A. & Whelpton, M. 2009. Automatic extraction of semantic relations for less-resourced languages. In: B.S. Pedersen, A. Braasch, S. Nimb, and R. Vatvedt Fjeld, (Eds.). *Proceedings of the Workshop "Wordnets and other Lexical Semantic Resources - between Lexical Semantics, Lexicography, Terminology and Formal Ontologies" at 17th Nordic Conference on Computational Linguistics (NODALIDA) 14–16th May 2009*, NEALT Proceedings Series Volume 7, pp. 1–6. Odense, Denmark.

Pedersen, B.S., Nimb, S., Asmussen, J., Sørensen, N.H., Trap-Jensen, L. & Lorentzen, H. 2009. DanNet: the Challenge of Compiling a Wordnet for Danish by Reusing a Monolingual Dictionary. *Language Resources and Evaluation* 43: 269–299.

NSM. http://www.une.edu.au/bcss/linguistics/nsm/semantics-in-brief.php. (9th June 2011).

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

Rögnvaldsson, R., Loftsson, H., Bjarnadóttir, K., Helgadóttir, S., Nikulásdóttir, A., Whelpton, M., & Ingason, A.K. 2009. Icelandic Language Resources and Technology: Status and Prospects. In: R. Domeij, K. Koskenniemi, S. Krauwer, B. Maegaard, E. Rögnvaldsson, and K. de Smedt, (Eds.). *Proceedings of the NODALIDA 2009 Workshop Nordic Perspectives on the CLARIN Infrastructure of Language Resources*, NEALT Proceedings Series Volume 5, pp. 27–32. Odense, Denmark.

SALDO *Instruktion*. Web resource: https://svn.spraakdata.gu.se/repos/sblex/pub/saldo_instruktion.pdf (8.11.2011).

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

Úlfarsdóttir, Þ. 2006. Málfræðileg mörkun orðasambanda. *Orð og tunga* 8: 117–144.

Útdráttur

Orðabækur eru gerðar fyrir fólk en hins vegar eru mörg rafræn orðfræðileg málsöfn sett saman með tölvur í huga. Þeim er ætlað að geyma upplýsingar um form, notkun og merkingu orða á þann hátt að tölvur geti greint mannlegt mál á markvissan hátt, til þess að draga fram upplýsingar úr textum eða tali og til þess að draga ályktanir af þeim upplýsingum sem þannig er aflað. Ganga má út frá því að lifandi notendur viti ýmislegt fyrirfram, annaðhvort af skynsemi sinni eða almennri þekkingu, en aftur á móti hefur tölva enga fyrirfram gefna vitneskju. Í greininni er fjallað um ýmiss konar dæmigerðar merkingarfræðilegar upplýsingar í svonefndum orðanetum eins og WordNet (fyrir ensku) og DanNet (fyrir dönsku), en einnig í orðfræðilegum gagnasöfnum sem eru í grundvallaratriðum annarrar gerðar eins og SALDO (fyrir sænsku).

Lykilorð

merkingarfræði, merking orða, málgreining, orðanet, merkingarvensl

Keywords

lexical semantics, natural language processing, wordnets, semantic relations

Dr. Matthew Whelpton
Faculty of Foreign Languages, Literature and Linguistics
University of Iceland
whelpton@hi.is