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 fre-

quent, 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 pre-specifying 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 cut-off 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.