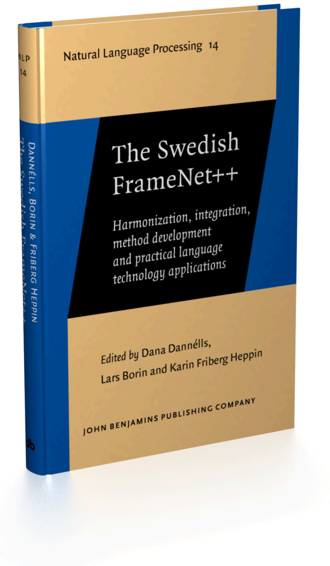
CHAPTER 3

Swedish FrameNet++ – lexical samsara*

Lars Borin | University of Gothenburg
Markus Forsberg | University of Gothenburg
Lennart Lönngren | Arctic University of Norway
Niklas Zechner | University of Gothenburg

https://doi.org/10.1075/nlp.14.03bor


Natural Language Processing 14
NLP 14
The Swedish FrameNet++
Harmonization, integration, method development and practical language technology applications
Dannélls, Borin & Friberg Heppin
Edited by Dana Dannélls, Lars Borin and Karin Friberg Heppin
John Benjamins Publishing Company

Pages 69–96 of
The Swedish FrameNet++: Harmonization, integration, method development and practical language technology applications
Edited by Dana Dannélls, Lars Borin and Karin Friberg Heppin
[**Natural Language Processing**, 14] 2021. xiv, 333 pp.

CHAPTER 3

Swedish FrameNet++ – lexical samsara[1]

Lars Borin[1], Markus Forsberg[1], Lennart Lönngren[2]
and Niklas Zechner[1]
[1]University of Gothenburg / [2]Arctic University of Norway

One of the main goals of the Swedish FrameNet++ initiative is to recycle and include as many existing modern Swedish lexical resources as possible into one unified lexical macroresource useful for automatic language processing. In this chapter we describe the structure of Saldo, the central resource of Swedish FrameNet++, the design of the formal interlinking mechanism keeping the lexical macroresource together, and our work on Swesaurus, a Swedish wordnet, and a Swedish Roget-style thesaurus as components of Swedish FrameNet++.

> ***samsara*** *[n] (Hinduism and Buddhism) the endless cycle of birth and suffering and death and rebirth*
> WordNet (v. 3.1): s.v. *samsara*

1. Introduction

The present chapter describes the modern lexical resources that have formed the foundation for Swedish FrameNet++ (SweFN++), both those already incorporated in SweFN++ and those where this work is still ongoing.

The original inspiration for the "++" part of the name Swedish FrameNet++ came from the assumption that formal interlinking of our existing digital lexical resources would be a necessary prerequisite for their usefulness in new language technology (LT) applications.

[1] Parts of this chapter build on and elaborate content previously presented in the following publications: Borin (2005); Borin et al. (2008, 2013); Borin & Forsberg (2009, 2010, 2011, 2014); Zechner & Borin (2020).

Linguistic examples in this chapter are glossed using the *Leipzig Glossing Rules* https://www.eva.mpg.de/lingua/resources/glossing-rules.php, with the following addition(s): a: adjective; n: noun; v: verb.

https://doi.org/10.1075/nlp.14.03bor 🔒 Available under the CC BY-NC 4.0 license

In part, the prerequisites for this interlinking were already in place, in the design of the central component of SweFN++, the lexical-semantic resource Saldo (see Sections 2 and 3), but most of the details still remained to be worked out, and other resources needed to be adapted to the envisioned structure. In this chapter we first describe the conceptual and formal structure of the central resource of SweFN++, Saldo (Sections 2 and 3), followed in Section 4 by a description of other modern resources already at least in part incorporated in the macroresource. Finally, in Section 5, some of the resources which are still in the pipeline for inclusion are mentioned, and we discuss how new types of lexical information have been added during the work on SweFN++.

2. Saldo: The heart of Swedish FrameNet++

2.1 Saldo in a nutshell

Saldo is a lexical-semantic resource, presenting an alternative information architecture to the best-known such resource, the Princeton WordNet (PWN; Fellbaum 1998). For a comparison between Saldo and PWN, see e.g. Borin & Forsberg (2009) and Borin et al. (2013).

Some of the salient characteristics of Saldo are:

1. The lexical-semantic relations in Saldo are non-classical. These are described and motivated in Section 2.3;
2. Saldo covers *all* parts of speech, not only open classes (Sections 2.3 and 2.4);
3. it reflects a principled approach to the description of multiword expressions (MWE; see Chapter 9 in this volume);
4. as well as a principled approach to content-model standardization with an explicit formal expression (Section 3).

2.2 The origin of Saldo

Svenskt associationslexikon (SAL; see Lönngren 1992) – 'The Swedish Associative Thesaurus' – which formed the basis for Saldo, is a relatively little known Swedish thesaurus with an unusual semantic organization (described in Section 2.3 below).

SAL was first published in paper form in two reports by Lönngren (1989, 1992). The ideas underlying SAL and its history have been documented by Lönngren (1988a, 1989, 1998) 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 3,000) of proper nouns found in SAL. Eventually, a list of the headwords from a large Swedish reference dictionary was included in SAL, and its second paper edition (Lönngren 1992) contained 71,750 entries.

The work described here first started in late 2003, when the three first authors of this chapter initiated a collaboration aiming at making SAL into a digital lexical resource for LT. In order to be practically useful in LT applications, it had to be complemented with inflectional morphological information. Saldo consequently has a lexical-semantic component inherited from SAL – described in the next section – and a morphological component providing information about part of speech, inflectional and compounding behavior, and some other information about entries, which comprise both single orthographic words and multiword expressions. Morphological processing – inflectional analysis, inflectional full-form generation, and compound analysis – is handled by a dedicated morphology component (see Section 2.4). At the time of writing, in January 2021, Saldo contains close to 150,000 entries, and it is growing continuously.[2]

2.3 The semantic structure of Saldo

As a semantic lexicon, Saldo is a kind of lexical-semantic network. The basic linguistic idea underlying Saldo is that, semantically speaking, the whole vocabulary of a language can be described as having a center – or core – and (consequently) a periphery. The notion of *core vocabulary* is familiar from several linguistic subdisciplines (see Chapter 6 in this volume). In Saldo this idea is consistently applied down to the level of individual word senses.

The basic lexical-semantic organizational principle of Saldo is hierarchical. Every entry in Saldo – representing a *word sense* – is supplied with one or more semantic descriptors, which are themselves also entries in the dictionary. All entries in Saldo (with the exception of the top node PRIM; see below) are actually occurring words or conventionalized or lexicalized multiword units of the language. One of the descriptors, called *primary descriptor*, is obligatory. The primary descriptor is the entry which better than any other entry fulfills two requirements: (1) it is a *semantic neighbor* of the entry to be described and (2) it is more *central* than it. Both these aspects need some clarification.

That two words are semantic neighbors means that there is a direct semantic relationship between them (such as synonymy, hyponymy, antonymy, meronymy, argument-predicate relationship, and so on). Here we are immediately faced with one problem: What about semantically empty words: how can we determine their neighbors? Saldo includes not only open-class words, but also pronouns, prepositions, conjunctions, etc., which are often considered to have only a syntactic

[2] See https://spraakbanken.gu.se/en/resources/saldo.

function, rather than semantic content. This is, however, a simplification. It is true that some of these words, such as *of, for, that*, are extremely empty, but so are, for instance, support verbs like *exert* (and the noun *exertion*), *undergo*, etc. In a phrase like *exert influence* the verb has a purely syntactic, i.e. connective, function. In such cases closeness must be determined with respect to function or syntagmatic connections, rather than (semantic) content. Fortunately, the majority of entries have semantic content, so we will continue to talk here about semantic closeness without loss of generality.

Centrality is determined by means of several criteria, the most important being frequency: a word with high frequency is more central than a word with low frequency. Although the compilers of Saldo rarely refer to explicit frequency data in making their decisions about primary descriptors, but rather rely on their lexicographical experience and linguistic intuition in making this judgement, in fact it turns out *a posteriori* that slightly over 90% of the Saldo entries have primary descriptors that are at least as frequent as the entries themselves in a large (billions of words) corpus of written Swedish.[3]

A second criterion is stylistic and emotive value: a stylistically and emotively neutral word is more central than one that is stylistically marked or carries an emotive connotation.

While the first two criteria could be applied to any pair of word senses, a third, more limited criterion is based on word formation: of two derivationally related words the one with lower complexity is more central than the one with higher complexity.

Finally, there are certain semantic relationships that themselves are asymmetric in such a way that one member must be considered superordinate. Thus a hyperonym is generally more central than a hyponym (but see below), a word signifying ‘part’ is less central than the word signifying the corresponding whole object.

Given that only one primary descriptor can be assigned, the compilers of Saldo will sometimes need to choose among several candidates which fulfill some, but not all, criteria. In such cases, the emergent topology of the whole lexical-semantic network becomes crucial (see below).

An entry in Saldo may in addition have an arbitrary number of *secondary descriptors*. A secondary descriptor is an entry which, once the primary descriptor is chosen, helps to specify the sense of the entry word, notably to differentiate it in relation to other entries with the same primary descriptor.

The requirement that Saldo form a hierarchy means that there must be at least one sense at the top which of course itself cannot have another word sense as

[3] The corpus collection is browsable through https://spraakbanken.gu.se/korp (Borin et al. 2012), and downloadable at https://spraakbanken.gu.se/en/resources.

primary descriptor. In fact, it turns out that there is a small number of senses for which so far no suitable primary descriptor has been found. In order to make Saldo into a single hierarchy, an artificial entry, called PRIM, is used as the primary descriptor of 41 semantically unrelated entries at the top of the hierarchy, making all of Saldo into a single rooted tree. These “semantic primitives” of Saldo are shown in Figure 1. See also Chapter 6 in this volume for a description of these primitives in the context of a discussion of core vocabularies in linguistics and related disciplines.

all^{1} ‘all’
annan^{1} ‘other’
bara^{1} ‘only’
bra^{1} ‘good’
fort^{1} ‘quickly’
framme^{1} ‘in/at front’
färg^{1} ‘color’
för^{2} ‘for’
före^{1} ‘before’
göra^{1} ‘do, make’
ha^{1} ‘have’
hur^{1} ‘how’
hända^{1} ‘happen’
i^{2} ‘in’
ja^{1} ‘yes’
just^{1} ‘exactly’
ljud^{1} ‘sound’
ljus^{1} ‘light’ (n)
med^{2} ‘with’
men^{1} ‘but’
mycken^{1} ‘much’
måste^{1} ‘must’
namn^{1} ‘name’
natur^{1} ‘nature’
när^{1} ‘when’
om^{1} ‘if’
om^{2} ‘about’
på^{1} ‘on’
rak^{1} ‘straight’
röra^{1} ‘move’
säga^{1} ‘say’
till^{1} ‘to’
tänka^{1} ‘think’
vad^{1} ‘what’
var^{1} ‘where’
vara^{1} ‘be’
varm^{1} ‘warm, hot’
vem^{1} ‘who’
veta^{1} ‘know’
vid^{1} ‘by’
vilja^{1} ‘want’

Figure 1. Saldo’s 41 “semantic primitives”: The top-level word senses

Below, we give a few examples of entries, found under the letter “L” in Saldo (superscript numbers differentiate lexical items – i.e. word senses – with the same lemma; the index “1” is sometimes omitted):

> *läkare*^{1} : *bota*^{1} ‘physician’ : ‘cure (v)’
> *lexikon*^{1} : *ordbok*^{1} ‘lexicon’ : ‘dictionary’
> *lexikon*^{2} : *ordförråd*^{1} ‘lexicon^{2}’ : ‘vocabulary’
> *lie*^{1} : *slå*^{2} ‘scythe’ : ‘mow’
> *liga*^{1} : *grupp*^{1} + *brottslig*^{1} ‘gang’ : ‘group’ + ‘criminal (a)’
> *Lucretius*^{1} : *filosof*^{1} + *romersk*^{1} ‘Lucretius’ : ‘philosopher’ + ‘Roman (a)’

We have often characterized Saldo as an “associative thesaurus”, but in this connection it is important to note that the intended associations are *lexical-semantic* ones and by no means the free word associations elicited in a psychological or psycholinguistic test.[4] Nor do they coincide with the kinds of relations among words referred to as “non-classical” (Morris & Hirst 2004), “evocation” (Boyd-Graber

[4] For instance, in the *Dictionary of Russian associative norms* (see Chapter 6 in this volume for the concept of psycholinguistic “norms”) by Leont’ev (1977), the adjective *lošadinaja* ‘of a horse, equine’ is listed among the strongest associations for the word *familija* ‘family name’, which comes from the title of a short story by Anton Chekhov (*A horsey name*), but does not reflect any kind of linguistic association in the sense intended here.

et al. 2006), or "association" (Hill et al. 2015) in the literature, even if there is some overlap with all of these. The only valid criteria for appointing a primary descriptor to an entry are *maximal semantic closeness to the entry and higher centrality than the entry*. These criteria are partly *extrinsic*, in the sense that they need to draw on information that is not to be considered part of the word senses themselves.

For each entry the primary descriptor is selected manually, according to the mixed relationship mentioned above. Is this an objective method? Yes, we believe so. This method corresponds to that used in traditional lexicographic work, which generally proceeds on an entry-by-entry basis. Thus, the global structure emerges from numerous local decisions. In traditional lexicons and in Saldo alike, systematicity and objectivity are achieved by having a clear set of principles and highly qualified people – i.e. with appropriate training in linguistics and lexicography – to do the work. The objectivity is further ensured by thorough subsequent examination of the resulting network configuration surrounding each entry. Groups of entries sharing the same primary descriptor should be semantically coherent, and the occasional odd ones out are relatively easy to spot, so that their description can be checked and revised if needed. The constant revision of previous analyses is extremely important.

As was mentioned above, the predominant relations are synonymy and hyponymy. Synonymy, being non-directional, requires additional criteria to ensure that the entry is less central than its primary descriptor, for example frequency, stylistic and emotive value, and morphological complexity. Examples are *tjej* 'gal, chick' : *flicka* 'girl'; *sannolikhet* 'probability' : *sannolik* 'probable'; *liv* 'life' : *leva* 'live (v)'.[5]

Hyponymy is a directional relation, the hyperonym being, in principle, more central than the hyponym, for example *kanot* 'canoe' : *båt* 'boat'. In compounds, so productive in Swedish, the second part is, as a rule, superordinate in relation to the whole word, and can often serve as primary descriptor. The first part of the compound then is typically used as secondary descriptor, for example *stenhus* 'stone house' : *hus* 'house' + *sten* 'stone'.

It turns out that in many cases neither a synonym nor a hyperonym can be used as a descriptor. Possible candidates may be less central than the keyword, as has long been claimed about natural kinds, where central, prototypical concepts often reside at some intermediate taxonomical level (the "basic level" of Rosch et al. 1976). For example, for *häst* 'horse' the zoologically correct taxonomical hyperonym *hovdjur* 'ungulate' must be rejected. There may also exist a more central,

[5] Many, but not all construals of synonymy make it to be non-directional; see Murphy (2003: Chapter 4) for a discussion. Note also that the pairs *sannolikhet* 'probability' : *sannolik* 'probable' and *liv* 'life' : *leva* 'live (v)' would be considered synonyms on some definitions of this term, despite their different parts of speech.

but too distant, hyperonym, for example *person* 'person' in words like *erövrare* 'conqueror'. Similarly, in many cases *del* 'part (n)' is rejected as primary descriptor.

In the absence of a suitable synonym or hyperonym other relations correspond to the primary descriptor. For example, in *balkong* 'balcony' : *hus* 'house' we observe meronymy, a directional relation. The primary descriptor may be a predicate, for which the entry would be a prototypical argument, for example *damm* 'dust' : *torr* 'dry'; *streck* 'line' : *rita* 'draw (e.g. a picture)'. Particularly commonly occurring arguments are agents and instruments, for example *vinnare* 'winner' : *vinna* 'win'; *astronom* 'astronomer' : *astronomi* 'astronomy'; *kniv* 'knife' : *skära* 'cut'.

If a synonym is used as the primary descriptor, there is no need for a secondary descriptor. If a hyperonym is used, one or more secondary descriptors are often added, with a specifying function. Beyond this, there are no particular semantic requirements on a secondary descriptor. It can be a holonym or a predicate, for example *bordskant* 'table edge' : *kant* 'edge' + *bord* 'table'; *såg* 'saw (n)' : *verktyg* 'tool' + *såga* 'saw (v)'. But also other semantic relations are possible here, for example material, as we saw in *stenhus* 'stone house'.

In many cases the predicate–argument relation holds between two descriptors, for example *glasblåsning* 'glass-blowing' : *blåsning* 'blowing' + *glas* 'glass'; *snabbseglande* 'fast-sailing' : *segla* 'sail (v)' + *snabb* 'fast'. Certain words are very frequently used as modifying secondary descriptors, for example: *inte* 'not', *utan* 'without', *(o)möjlig* '(im)possible', *hon* 'she', *motsats* 'opposite (n)', *mycket* 'very', *alltför* 'too', *igen* 'again', *mot* 'against', *väl* 'well', *nyss* 'recently', *lätt* 'easy', *likna* 'be similar', and *som* 'like, as'.

While every entry in Saldo has exactly one primary descriptor, it may have an arbitrary number of secondary descriptors, e.g. we find multi-part compounds like *femrumslägenhet* 'five-room apartment' (and derivations like *femrummare*, lit. 'five-roomer'), with two secondary descriptors: *lägenhet* 'apartment' + *rum* 'room' & *fem* 'five'.

As was already pointed out, each distinct sense of a word constitutes a separate entry in Saldo. Distinguishing such senses is a difficult task, which is solved in different ways by traditional lexicographers and in different lexicographical traditions. To single out an autonomous sense in Saldo means accepting this distinction pervasively; for instance, wherever the word occurs in a compound it must be assigned the correct sense index.

Saldo has on the average a 1.17 senses per entry lemma, and the most polysemous entry has 19 senses. Approximately 13% of all base forms are polysemous. In PWN the most polysemous entry is the verb *break* with 59 senses. Both lexicons display the same kind of – distinctly Zipfian – distribution of senses over headword lemmas (see Chapter 1 in this volume).

The considerable difference in word-sense granularity between Saldo and PWN is partly because of lexicographic tradition, partly because of the way sense distinctions must be propagated through the Saldo hierarchy, and partly due to the special construal of synonymy determining the structure of PWN. Word sense distinctions can be made arbitrarily fine-grained (Murphy 2003), among other things depending on how much the understanding of words in context is believed to rely on the same general principles of interpretation and inference which can also be invoked to account for novel word usages, such as creative metaphor (Cruse 2000; Hanks 2013). However, there seems to be a (fuzzy) limit beyond which the usefulness of further subdivision diminishes, at least from a practical text-processing perspective (Kilgarriff 1997; Hanks 2000; Erk 2010).

Saldo is built completely bottom-up, by means of a large number of local decisions. No global plan is imposed. There are no metalinguistic categories or features. Every sense is given a description, and is also itself a potential descriptor. Of course, a great part of the senses have not – yet – been used as descriptors. At present, two thirds of the word senses in Saldo are leaf nodes in the tree.[6]

By logical necessity, every sense in Saldo has a certain depth, by which we mean distance from the top node PRIM in terms of a chain of successive primary descriptors. For instance, starting from *myror i huvudet* [ants in head.DEF.SG] 'puzzle-headedness', which has depth 13, we arrive at PRIM through the following steps: *sätta myror i huvudet* [put ants in head.DEF.SG] 'puzzle (v)', *bryderi* 'perplexity, puzzlement', *förvirring* 'confusion', *förvirra* 'confuse', *oreda* 'disorder', *oordnad* 'disordered', *ordna* 'order (v)', *ordning*^{2} 'order (n), sequence', *följd* 'succession', *följa* 'follow', *efter* 'after', *före* 'before'. This entry is close to the maximal depth in Saldo, which is 16, while the mean depth is only 6, and the median depth is 7.

Interestingly, comparing Saldo's topmost word senses – the children of PRIM – with the "semantic primes" of Wierzbicka and Goddard's *Natural Semantic Metalanguage* (NSM) (Wierzbicka 1996; Goddard 2008), we may note that the Swedish counterparts of the NSM primes (Goddard & Karlsson 2008) are generally found close to the top node in Saldo.[7] This again indicates that the numerous local decisions by which Saldo's lexical-semantic hierarchy emerges are, on the whole, sound.

[6] Thus, the average number of children of a non-leaf node in Saldo is between two and three.

[7] The NSM semantic primes have undergone many revisions through the years. For a current version, see the *Proposed semantic primes (2014)* on the NSM homepage: https://intranet.secure.griffith.edu.au/schools-departments/natural-semantic-metalanguage/what-is-nsm/semantic-primes.

2.4 Morphological information in Saldo

The original SAL was a pure onomasiological lexicon – i.e. a lexicon organized by meaning – with no 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. If we want to offer a full lexical component for LT applications, inflectional morphology must also be addressed. Swedish inflection – although still fairly simple from a global perspective – is an order of magnitude more complex than that of English. There are more forms in the paradigms for all the open parts of speech. In addition, there are 7 noun declensions, 2 adjective declensions, and 3 verb conjugations – all with further subdivisions – plus a number of irregular items.

The morphological component of Saldo has been implemented using *Functional Morphology* (FM; see Forsberg 2007; Borin et al. 2008). The size of Saldo's morphology is ~2M word forms compiled from 128k morphological specifications, i.e. an average of almost 16 word forms per specification.[8]

At present there are more than one thousand different inflectional patterns represented in the lexicon. Among these are many singleton patterns. In many cases, these are the irregular words of traditional grammar. Surprisingly often, however, the source of plenty is another, viz. variation. We often find that a particular combination of morphosyntactic features – a particular position in a paradigm – for a word or small group of words can be filled by more than one form, i.e. realized in more than one way. Such cases are legion, but each case is also restricted to individual items, while seemingly analogous words will not show the same behavior, e.g. the three alternative forms *himmeln, himlen, himmelen* [heaven.SG.DEF.NOM] (citation form *himmel*), of a word which in all other respects follows the inflectional pattern designated as `nn_2u_nyckel`,[9] which includes words like *nyckel* 'key', *åker* 'field', *öken* 'desert', *hummer* 'lobster'. This pattern allows only for the first of the three variants shown above for the singular nominative definite form of himmel, viz. the form made by affixing an *-n* to the citation form.

[8] Recall that Saldo's inflection tables include compounding forms – since these are not predictable in the general case – so that even many uninflected items end up with at least two forms in their paradigm.

[9] The inflectional pattern identifiers are designed to be mnemonic to a human user familiar with Swedish inflectional morphology. Thus, `nn_2u_nyckel` indicates a second-declension ("2") non-neuter ("u") noun ("nn") inflected like the word *nyckel* '(door)key', i.e. dropping the last *e* before a vowel-initial suffix.

A practically useful computational lexicon should in any case specify the morphological behavior of individual words as accurately as possible. In Saldo, this behavior is encoded uniformly for all words, in the form of a unique identifier for each inflectional pattern. In other words, in the lexicon we do not make a distinction between inflectional classes and individual cases in the sense of, e.g. Wurzel (1989: 57). This task is relegated to the computational morphological component, where a mapping is made between Saldo's inflectional patterns and regular, sub-regular and idiosyncratic inflectional descriptions. However, it is not difficult to get a picture of which inflectional patterns are general and which idiosyncratic. As with many other linguistic phenomena, a small number of patterns account for the majority of entries. There are 28 inflectional patterns with more than 1,000 members each in Saldo, which together account for 76% of all entries.

3. Persistent identifiers: The glue of Swedish FrameNet++

Any entities to be manipulated in the infrastructure should have unique *persistent identifiers* (PIDs). This was an additional reason for choosing Saldo as the central resource of SweFN++. In Saldo, there are PIDs for word senses, lemgrams,[10] parts of speech, paradigms, and lexical-semantic relations. The most important ones in the larger scheme of things are the word sense identifiers, which form the links between Saldo and other resources, and by extension, links among all the other resources through Saldo. For pragmatic reasons, the identifiers were designed both to be mnemonically human-readable[11] and to be valid XML names. A difference in Saldo sense identifiers conveys only that there is a word sense difference, but says nothing about its character or magnitude; this information must be provided separately, e.g. as a lexical-semantic relation, possibly accompanied by a degree (e.g. "A and B have a degree of synonymy of 0.8").[12]

Since Saldo is the pivot among our lexical resources, work on the other resources takes Saldo's sense and lemgram sets as their point of departure. This is not

[10] Since more commonly encountered terms such as *lexeme* and *lexical item* are ambiguous in practice, we have coined the term *lemgram* (*lem*ma+*gram*matical characteristics) used in Saldo about the combination of a lemma and a specified set of formal features, including pronunciation, part of speech, inflectional paradigm and compounding form(s).

[11] This has turned out to be a very sensible design decision. Non-human-readable PIDs – such as handle identifiers or DOIs – are of course formally equivalent. However, with well-designed mnemonic identifiers for key entities, all kinds of manual work become much more efficient.

[12] Degrees of synonymy are a feature of the Synlex resource, a crowdsourced Swedish synomym lexicon (Kann & Rosell 2006); see Section 4.1.3.

supposed to be a straitjacket, however. Often, this work reveals the need for new lexical units, which are added to Saldo – possibly after some discussion among the team members – and from there in turn propagated to the other resources.

4. Branching out: Lexical semantics galore

One of the main reasons for embarking upon the SweFN++ endeavor has been the methodological hypothesis that its component resources could be both harmonized and synergistically enriched in the process, yielding a whole that is more than the sum of its parts. Hence our focus in this chapter is on the extraction, harmonization and refinement of resource-internal linguistic information,[13] in order both to enrich and to the individual component resources.

4.1 The Swesaurus component of Swedish FrameNet++

4.1.1 *Towards a Swedish wordnet*

The PWN is a de facto standard for lexical-semantic resources in language technology, because of its size and even more because of its open license. PWN has inspired numerous wordnet projects for other languages, and of course we would like one for Swedish as well, if nothing else, in order to compare a wordnet and Saldo as resources for various LT applications.

The lexical-semantic information in Saldo is not an alternative to that found in a wordnet, but should be seen as a complement. Ideally, we would like both kinds of information for all lexical entries. However, at present, there is no generally available full-size wordnet for Swedish (see Chapter 5).

One of the emerging components of SweFN++ is Swesaurus, a Swedish "protowordnet" under active development.[14] Swesaurus is being constructed mainly by recycling and refining lexical-semantic information from a number of existing lexical resources. This is ultimately made possible by the design principles of SweFN++, as outlined above in Section 3 and in more detail in Chapter 1.

[13] A concrete example of a refinement which has often been requested by researchers using SweFN++ concerns definitions: Saldo lacks definitions (due to its history), and the descriptors are sometimes less than ideal indicators of semantic differences among colexified lemmas. However, definitions are present in some of the resources slated for inclusion in SweFN++, e.g. the Lexin dictionary (see Section 4.1.5).

[14] Chapter 5 in this volume situates Swesaurus in the context of international wordnet-creation efforts (including linkage to Princeton WordNet), while the focus of the present description is on the methodology used to extract and infer classical lexical-semantic relations from extant lexical resources.

Even if the goal of the Swesaurus endeavor is to create a Swedish wordnet, it exhibits some noteworthy differences from a true wordnet.

Firstly, all the lexical-semantic relations in Swesaurus are between word senses only; there are no synsets.[15] Synonymy is simply one of these relations among many others. This design feature is partly due to tradition, but in this way we also avoid having to define synonymy. Even though synonym dictionaries are among the oldest products of lexicography – even the Sumerians and Akkadians compiled them (Civil 1990) – in practice synonymy has turned out to be a most slippery notion: while synonyms are self-evidently a central feature of language according to Lieber (1841: vii), they are "morally impossible" to Döderlein (1863: xii).

Occam's razor also enters into the picture: since word senses seem to be needed in any case, and to be in some sense more basic than synsets – more than half (54%) of the synsets in PWN have only one member, arguably a word sense[16] – we see no pressing need to adopt the synset as basic notion.

Secondly, Swesaurus covers all parts of speech, unlike PWN, which contains only the open parts of speech, the "content words", specifically nouns, verbs, adjectives, and adverbs. However, because of what seems to be a specific Anglo-Saxon lexicographical practice (Apresjan 2002), numerals are also included in WordNet, classified as nouns (cardinals) or adjectives (ordinals).

Thirdly, from one of its constituent resources (Synlex; see Section 4.1.3), Swesaurus inherits the notion of *graded relations*, primarily *degree of synonymy*.

The classical lexical-semantic relations present in Swesaurus are listed in Table 1. The basic information unit in Swesaurus is the (word-sense) *relational triple*, whose three components are: (1) a source word sense; (2) a *graded* (in the interval [0..100]) lexical-semantic relation; and (3) a target word sense. In addition, each triple has provenance information, i.e. from which resource it originates and whether it is primary or derived. All relations except `related-sense` are generally taken to hold only within a part of speech, i.e. source and target word senses must belong to the same part of speech. A concrete example:

abrupt^1 IS-A-SYNONYM-OF:80 *plötslig*^{1} synlex

This triple (extracted from Synlex; see Section 4.1.3) conveys the information that (the Saldo word sense) *abrupt*^{1} 'abrupt' is 80% synonymous with *plötslig*^{1} 'sudden'.

[15] *Synsets* are the fundamental units constituting WordNet, defined as "sets of synonyms that serve as identifying definitions of lexicalized concepts" (Miller et al. 1990: 240). This means that synonymy holds a special place in WordNet, different from and more basic than other classical lexical-semantic relations, where the latter are understood in WordNet to hold among synsets, not word senses.

[16] This holds for example for the noun *samsara* used in the title of the present chapter.

Table 1. Lexical-semantic relations used in Swesaurus and their logical properties (used for inferring missing links among lexical items)

Relation	Logical properties
`synonymy`	symmetric, transitive
`antonymy`	symmetric
`related-sense`	symmetric, transitive(?)
`hyponymy/subordinate sense`	transitive, inverse of `hyperonymy`
`hyperonymy/superordinate`	transitive, inverse of `hyponymy`
`cohyponymy`	symmetric, transitive
`partonymy`	transitive(?), inverse of `holonymy`
`holonymy`	transitive(?), inverse of `partonymy`

4.1.2 *Mining Saldo for classical lexical-semantic relations*

The foremost of the resources utilized for compiling Swesaurus is Saldo itself.

We capitalize on the fact, already mentioned (in Section 2), that the primary descriptor of a Saldo entry will in practice quite often be either a hyperonym or synonym of the entry. Thus, Saldo was mined for Swesaurus candidates by extracting all same-POS entry–primary descriptor pairs. In the process, some important special cases were recognized which require very little manual post-processing, such as noun compound entries where the form of the primary descriptor corresponds to the last member of the compound, e.g. *livförsäkring : försäkring* 'life insurance' : 'insurance' – about a third of all noun entries in Saldo (see Chapter 9 in this volume) – and where the entry in the overwhelming majority of cases is a hyponym of the primary descriptor. Using this method, a large number of synonyms, near-synonyms, hyperonyms, antonyms, and related senses could be extracted from Saldo, representing all parts of speech.

From the initial set of relations mined in this way from Saldo (and other resources described below), we can derive additional relations which follow from the logical properties (e.g. transitivity) which obtain among relations. For example, if we know that A `is-a-synonym-of` B and B `is-a-synonym-of` C, we can infer that A `is-a-synonym-of` C even in the absence of explicit information to this effect. More subtle inferences are also possible, for example, if A `is-a-hyponym-of` B and C `is-a-hyponym-of` D and A `is-a-cohyponym-of` C, we can infer that B `is-a-synonym-of` D.

According to the website of the *Global WordNet Association*,[17] "resources that follow the wordnet design" must include

- links to WordNet (Princeton or others that are linked to PWN)
- WN structure (minimally: synset, hyponymy)

[17] http://globalwordnet.org/resources/wordnets-in-the-world/

Swesaurus marginally fulfills the first criterion – only one of its components (the Core WordNet; see Section 4.1.5) is linked to PWN – although we acknowledge the usefulness of such a linking, and are planning to extend it to the other components of Swesaurus. It also fails the second criterion, since there are no synsets at all in Swesaurus. However, as we have argued and shown above, a PWN-style wordnet – in fact, many different PWN-style wordnets, in the form of "wordnetified" versions of Swesaurus – can be completely mechanically derived from Swesaurus through the transitive closure of the synonymy relation, and to some extent also utilizing other relations.

4.1.3 *Synlex*

Synlex (the People's Synonym Lexicon; Kann & Rosell 2006)[18] is a lexical resource that has been created by asking members of the public – users of an online Swedish-English dictionary – to judge the degree of synonymy of a random, automatically generated synonym pair candidate, on a scale from 0 (not synonyms) to 5 (fully synonymous). A synonym pair list containing all pairs that average 3.0 or more on three or more judgements is available for download under an open-source license. The version selected for inclusion in Swesaurus contains almost 20,000 synonym pairs, selected and graded by crowdsourcing, and subsequently curated in the SweFN++ project.

The members of these pairs are words (i.e. text word forms) – not even part of speech is indicated – mainly dictionary base forms (lemmas), but sometimes inflected forms, and in some cases multiword expressions. One problem then becomes, in the case of a word having as synonyms several other words – because of colexification – to determine how many senses we are dealing with.

We could use the synonymy degree information at arbitrary cut-off points to create virtual "fuzzy wordnets" for Swedish (Borin & Forsberg 2010). This would partly address an oft-heard criticism of the PWN concept, invoking a postulated universal linguistic principle of *synonymy avoidance* (Carstairs-McCarthy 1999; Murphy 2003). 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 PWN-like lexical resource where we could exactly quantify the 'near-synonymy' that is sometimes said to define PWN synsets.

Graded relations complicate this picture, and it is not completely clear how to best use the degree information in computing derived relations. Consequently, we must be careful when deriving new synonym pairs in Synlex, especially if we iterate over already derived ones.

[18] http://folkets-lexikon.csc.kth.se/synlex.html

4.1.4 *Wiktionary*

Wiktionary is an undertaking similar to Wikipedia, but for collaborative writing of dictionaries rather than encyclopedias. The Swedish Wiktionary,[19] is a downloadable free resource that, among other things, contains some lexical-semantic relations. The work of extracting such relations from Wiktionary is hampered by the fact that the data set is only partially encoded with a formal structure. It is the responsibility of the author of each lexical entry to encode the different information categories in it in the correct wiki format that was intended by the creator of Wiktionary, but no automatic check of the encoding is actually done. Since the result of a faulty encoding may actually look correct to the human eye, there are in practice a number of errors in Wiktionary that complicate the automatic information extraction.

We have experimented with extracting synonymy relations between words, with a resulting set of 10,529 synonymy pairs, of which 3,857 of the word pairs have members with only one sense in Saldo. Hence, no manual disambiguation is needed, so they may be incorporated immediately into Swesaurus. Some of the pairs are wrong, since some lexical entries contain information from more than one language. This results in a few cases where, e.g. a Swedish word is linked to a Polish one. In practice, this is rarely an issue since such a word will almost never coincide with a lemma in Saldo.

The synonymy relations in Wiktionary are in general of higher quality than those in Synlex, which is to be expected since the author of a lexical entry in Wiktionary makes a conscious choice when assigning synonyms to a word, but Synlex, on the other hand, builds upon automatically generated word pairs, with the consequence that words that are not normally judged synonymous are sometimes assigned a degree greater than zero. For example, consider the pair *förlovning : förpliktelse* 'engagement to be married' : 'obligation', the members of which are normally not considered to be synonymous, but when presented together and you are asked to quantify their synonymy degree, you may be tempted to give them at least a small degree of synonymy.

4.1.5 *Core WordNet*

As part of the EC-funded META-NORD project (2011–2013), a linking of the Princeton Core WordNet (CWN) to Swedish was completed and included in Swesaurus. The linkage was bootstrapped by using the *Lexin* basic Swedish-English dictionary (~25,000 entries).[20] Swedish lemmas in Lexin were automatically linked, in an overgenerating manner, to Saldo sense identifiers, giving us a set of senses

[19] sv.wiktionary.org

[20] https://spraakbanken.gu.se/en/resources/lexin

for every lemma. The glosses of CWN were subsequently linked to these sense sets via Lexin. CWN has 5,000 entries, of which around 89% were covered by Lexin. Furthermore, 23% had a unique link to one Saldo sense, and the remaining an average ambiguity of 4.4 (a rather high ambiguity, but not unexpected for a core vocabulary).

4.1.6 *The Gothenburg semantic database*

The Gothenburg Semantic Database (SDB; Järborg 2001) is a lexical database for modern Swedish covering 61,000 entries with an extensive description of inflection, morphology and meaning. Originally based on a lexicographical database that has been used in producing two modern Swedish reference dictionaries, SDB has been enriched with a deeper semantic description where many of the verb senses have been provided with semantic valency information using a set of about 40 general semantic roles and linked to example sentences in a corpus.

SDB holds two kinds of relevant lexical-semantic information: (1) explicit lexical semantic relations cross-referencing among different lexical entries (lemmas); and (2) relations implicit in its hierarchical organization of lexical entries into main senses and subsenses, typically corresponding to a superordinate– hyponym relation.

The linking of SDB senses to Saldo sense identifiers is ongoing. Some of its explicit lexical semantic relations have been included in Swesaurus, and some of the derived relations calculated (see Section 4.1.2). In the process, it has become clear that the explicit relations are not consistent, and will need a good deal of manual curation, which is ongoing.

4.1.7 *The state of Swesaurus*

All the activities listed in the preceding sections are ongoing to various degrees. In summary, approximate current numbers of primary and derived relational triples in the different Swesaurus components are as follows:

Component	**Primary**	**Derived**
Synlex	19,000	9,500
Wiktionary	4,000	–
CWN	4,500	–
SDB	10,000	13,500
Saldo	32,500	–

All counts are for *normalized relational triples*, which means that symmetric relations are counted only once for a given word-sense pair, and that for relations with an inverse, only one of the two is present in the data. Thus, A `is-an-antonym-of` B will exclude the presence of B `is-an-antonym-of` A, and A `is-a-hyperonym-of` B will be transformed into B `is-a-hyponym-of` A.

4.2 Towards a thesaurus component of Swedish FrameNet++

4.2.1 *Roget's* Thesaurus *and language technology*

While wordnets completely dominate the LT field, in most other contexts the most well-known lexical-semantic resource for English is without doubt Roget's *Thesaurus* (Roget 1852; Hüllen 2004), which appeared in its first edition in 1852 and has since been published in numerous editions all over the English-speaking world. The digital version of Roget offers a valuable complement to PWN (Jarmasz & Szpakowicz 2004), which has seen a fair amount of use in LT as an alternative source of lexical-semantic information, and which can be used both to address other kinds of LT tasks than with a wordnet, and even be more effective for some of the tasks where wordnets are normally used, e.g. *lexical cohesion, synonym identification, pseudo-word-sense disambiguation*, and *analogy problems* (Morris & Hirst 1991; Jobbins & Evett 1995; Jarmasz & Szpakowicz 2004; Kennedy & Szpakowicz 2008, 2014).

4.2.2 *Including Bring's Swedish thesaurus in Swedish FrameNet++*

One of the available existing lexical resources to be included in SweFN++ is Sven Casper Bring's Swedish adaptation of Roget's thesaurus, which appeared in 1930 under the title *Svenskt ordförråd ordnat i begreppsklasser* 'Swedish vocabulary arranged in conceptual classes' (Bring 1930).

Bring's thesaurus is made available online in two digital versions (under a CC BY license):

1. *Bring* (v. 1), providing the full contents of the original 1930 book version (148,846 entries)[21]
2. *Blingbring* (v. 0.3), a version of Bring which has been curated to remove obsolete items. This version contains 126,911 entries[22]

Bring is the resource where we have most systematically explored various automatic methods for linking its vocabulary to Saldo, thereby integrating Bring into SweFN++. This warrants a more detailed account of these efforts, to which we now turn.

[21] https://spraakbanken.gu.se/en/resources/bring

[22] https://spraakbanken.gu.se/en/resources/blingbring

The initial linking to Saldo senses in Blingbring did not involve a disambiguation step. Rather, as described in Chapter 1 in this volume, we have capitalized on the Zipfian distribution of word senses over lexemes. Hence, the linking was made between matching lemma-POS combinations from the two resources. As expected, most linkages are unambiguous: Blingbring includes slightly over 21,000 entries with more than one Saldo sense (~17%), or about 4,800 ambiguous word sense assignments (out of about 43,000 unique lemma-POS combinations: ~11%).

By and large, the formal structure of Bring is taken over from Roget. At the highest level, there are 1,015 numbered *conceptual classes*. Each class comes with a heading, a label indicating a broad semantic characterization of the words and multiword expressions listed in the class, e.g. classes #981: *himmel* 'heaven' and #982: *helvete* 'hell'. These semantic fields are often quite abstract, with the consequence that a particular Saldo word sense may be included in more than one Bring class. For example, the word sense *nirvana*^{1} 'nirvana' is listed in class #981, but also in the following two other classes: #2: *intighet* 'inexistence' and #360: *död* 'death'.[23]

Classes are further subdivided into parts of speech, with one division each for nouns, verbs and a third category containing words of other parts of speech, mainly adjectives and adverbs, but also idioms and some function items.

Finally, the lowest-level complex unit in Bring is the *group* (marked by a final semicolon in the printed version). The lemmas in a group – including multiword lemmas – referred to here as the *(lexical) items* (or *entries*) of Bring, are often further arranged by semantic proximity, so that synonym clusters can be discerned within groups, although these clusters are not formally indicated in any way.

As mentioned above, Bring was published in 1930, and its most recent entries were first attested in print around 1920, although the influx of new words starts to peter out already around 1914 (Lange 2007: 11), i.e. its vocabulary is over a century old.

In order to integrate Blingbring more fully into SweFN++, we need to accomplish two objectives: (1) to disambiguate the ambiguous linkages; and (2) to develop good methods for adding modern vocabulary to Bring from Saldo or some other SweFN++ component resource, placing each item in its most appropriate Bring class(es), thereby hopefully producing a modern Swedish Roget-style resource for the LT community. For both objectives we have pursued automatic, LT-based methods.

[23] Only about 40% of the unambiguous lemma-POS combinations in Blingbring appear in only one class, and at the other end of the spectrum, one item appears in 28 classes: *ytlig*^{1} 'superficial'.

Regarding the first objective, this is a closed task, since the number of ambiguous items is fixed, but the effort spent on automating this task will hopefully also take us some ways toward fulfilling the second objective.

The second objective is more difficult. Rather than just a small number of options, we now need to distinguish between a very large number of target classes, and an even larger number of groups within the classes. This is also an open-ended objective, in that we would ideally like any new sense added to SweFN++ to also be assigned its proper class(es) in Blingbring.

In an initial set of experiments applying machine learning approaches both a corpus-based and a lexicon-based classifier were applied to the first objective, the disambiguation problem, reaching accuracies of 69% and 78%, respectively (Borin et al. 2015). However, simply choosing the first listed sense in Saldo results in 63% accuracy, making the corpus-based method barely viable. Following up on the more promising lexicon-based approach, which utilized only one of several possible aspects of the lexical structure of Saldo, we have conducted a more detailed investigation of if and how more of Saldo's structure could be used for this purpose. The hypothesis was that the macrostructure of Saldo will correspond to that of Bring at some level.

The evaluation data used for the experiments consisted in 1,308 manually disambiguated gold-standard entries.[24] The degree of ambiguity in this gold standard data is shown in the second column of Table 2, while the third column shows the degree of ambiguity in the full Blingbring dataset containing 44,615 unique lemma-POS combinations (out of which 39,275 are unambiguous).

Table 2. Word-sense ambiguity in the gold standard data and in Blingbring (entries = lemma-POS combinations)

# senses/entry	Gold standard: # entries	Blingbring: # entries
2	739	4,006
3	304	873
4	147	286
5	71	102
6	11	31
7	13	18
8	15	10
9	6	3
10	2	6
11	0	5

[24] For details about how the gold-standard data was prepared, see Borin et al. (2015).

The terminology for lexical-semantic relations in Saldo's predecessor SAL (see Section 2.2) was based on a family metaphor: the primary descriptor was called "mother", the secondary descriptor referred to as "father" and their dependents were consequently known as "children" with "sibling" relations among them (Lönngren 1988b, 1989, 1992, 1998). Here we extend this analogy further in order to talk comfortably about the topology of Saldo's lexical network. A sense which has a particular other sense as its "mother" (primary) or "father" (secondary) is its "daughter" or "son", respectively. Senses sharing a primary or secondary descriptor are "sisters" or "brothers", respectively. In the otherwise rare case where the mother of one sense is the father of another, we will call them "cross siblings". Terms like "parent", "aunt", etc. should follow by analogy.

In the experiments described below, we investigate assignment of Saldo word senses both to coarser Bring (conceptual) classes, and to more fine-grained groups within classes, referring to both granularities jointly as Bring *groupings*.

For the first experiment, we make the assumption – possibly overly simplistic – that only one of the senses of an ambiguous Bring entry belongs in a particular class. We will refer to this sense as the *correct* sense and the other senses as *incorrect* senses.

For each ambiguous entry and for each of its possible senses, we investigate which (if any) of its closely related senses are present in the same Bring grouping. It quickly becomes clear that some of the relations are stronger indicators than others that the investigated sense is the correct sense. For example, if a parent of the sense is present in the group, that is a very strong indicator, but on the other hand, it only happens in a small percentage of the cases. Conversely, a sense with a sibling appearing in the class is much more common, but this is a less strong indicator that this is the correct sense. For a detailed description of the procedure, see Zechner & Borin (2020).

Figure 2 shows the results of applying such indicators disjunctively in order of decreasing strength, defaulting to the sense listed first in Saldo as a last resort. We can either spot a small number of entries with high accuracy, or a larger number of entries with lower accuracy.

A manual error analysis of the cases where this method has resulted in an incorrect sense being chosen reveals that generally, most of the failed items are closely related senses, often with quite subtle differences, such as *samling*^{1} 'collection', *samling*^{2} 'arrangement', and *samling*^{3} 'group', and sometimes including metaphors, such as *tomhänt*^{1} [empty.handed] 'with empty hands' and *tomhänt*^{2} 'with nothing to offer'.

For the second experiment, the task is to determine if a Saldo sense which is not present in Bring should be placed in a particular grouping or not. If the new

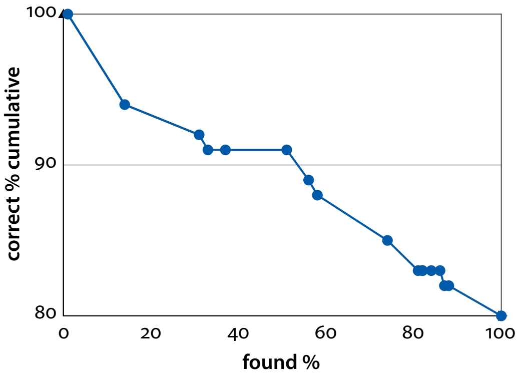


Figure 2. Coverage and accuracy for different methods of disambiguation

sense belongs in the current grouping, we call this grouping a *true* grouping (class or group) for the sense. Otherwise, it is a *false* grouping. As remarked above, there may be more than one true class for any word sense (but only one group within a class).

Again, we attempt to map the structure of Saldo's network onto that of Bring, the hypothesis being that a sense will have more relatives in its true grouping than in false groupings. Testing this is straightforward using the unambiguous lexical items already present in Bring, and we have conducted such a test on the full set of unambiguous entries. The hypothesis is confirmed. For example, a group that contains a given sense *x* will contain the mother sense of *x* in 13% of cases, but a group that does not contain *x* contains its mother sense in only 0.14% of cases, and at the granularity of Bring classes, the corresponding figures are 36% and 0.21%.

There are almost 7,000 candidate groups in slightly over 1,000 candidate classes for each new sense to be added, and preliminary experiments using similar methods as for the disambiguation step have showed that the accuracy is too low for automatic group or class assignment using the Saldo network structure.

Taking the mother sense as example, it is almost 88 times more likely to be found in a true group than in a false group, but a group containing the mother sense is still 4 times more likely to be false than to be true. With classes, the figures look somewhat more promising. The mother sense is about 170 times more likely to be found in a true class than a false class, but a false class containing the mother sense is twice as probable as a true class.

However, other close relatives (in particular daughters and sisters) are also good indicators. Exploring this fact, we have experimented with a scoring system, counting multiple close relatives in the same grouping, under the hypothesis that the highest-scoring grouping will be the true grouping. For groups, we find this to be true in 30% and for classes in 52% of the cases.

Since the accuracy of the automatic methods is barely over 50% (for classes), we could perhaps try semi-automatic methods. We can for example choose to list the suggested groupings in order of decreasing score, and see how many groupings we would on average need to look at to find a true grouping. Figure 3 shows the result both at the group and the class level.

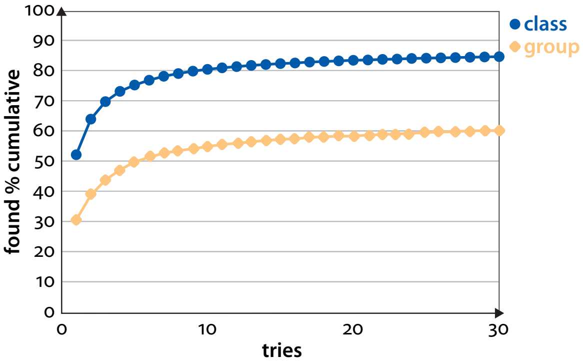


Figure 3. Percentage of entries for which a true grouping (class or group) is found within a given number of groupings, starting from the highest-scoring grouping

We see that while 30% of the true groups are found in the first guess, 50% are found in the first 5, and 55% in the first 10. These are still not very impressive numbers, but the results for classes look quite good in comparison. The highest-scoring class is the true class in 50% of the cases, and the true class will be among the top five in 75% of the cases and among the top ten 80% of the time.

The backbone of Bring is made up by the (1,015) conceptual classes, and even finding the correct class would help considerably, leaving only a presumably much more manageable task to human judgement – because it requires dealing only with a very local context instead of the whole lexicon – viz. that of the proper placement of the new word sense inside the class.[25]

Summing up, using the relations from Saldo to disambiguate or classify words in Bring is viable as a tool, even if the accuracy is not high enough to rely solely on this method. For disambiguation of already existing entries, we can get an accuracy of 80% for the entire list, and higher for a subset; this may be considered acceptable in itself, or it can be seen as a starting point for manual annotators. For classifying

[25] And in fact, placement inside the correct POS category is automatic, since POS is a defining characteristic of Saldo's word senses, further narrowing the number of groups that need to be considered once the proper class has been determined.

new senses, the accuracy is not good enough for automatic annotation, but it can reduce the number of classes (or groups) a manual annotator will have to consider by a large factor.

An added value from conducting these experiments – especially pursuing the second objective (adding new senses to Bring) – is an increased understanding of how the structure of Saldo can be used for measuring semantic distances among lexical items. This knowledge is applicable also in other cases of resource integration, e.g. the wordnet-building efforts described in Section 4.1 above, as well as the work described in Chapter 7 of this volume.

5. Looking forward: New directions up ahead

Over the years that the Swedish FrameNet++ initiative has been active, the initial design parameters have by and large stood the test of time. In addition to the originally foreseen resources – whose properties arguably have influenced the original design process to a great extent – we have also successfully included new kinds of lexical information in SweFN++.

One concrete example of this is the new Swedish sentiment lexicon SenSaldo, developed in the *Knowledge-based culturomics* project (Rouces et al. 2018),[26] where the whole development process was considerably aided by being able to draw on the conveniently and suitably structured lexical-semantic and morphological information already available in Swedish FrameNet++.

This bodes well for the prospects of including similar kinds of information in SweFN++, notably the kinds found in the lexical databases often referred to as "norms" by psycholinguists and psychologists (see Chapter 6 in this volume), i.e. text words or lemmas with information about, e.g. concreteness, sentiment, frequency, age of acquisition, spontaneous response, etc.

As noted above (especially in Section 4.2), the network structure of Saldo can be used to calculate semantic relationships among words, although with low recall. On the one hand, this confirms that the effort put into developing SweFN++ – to a very large extent manual and drawing on expert knowledge – has been well spent (cf. the comparison with the completely automatically produced resource BabelNet in Chapter 1). On the other hand, the low recall does point to a need for additional means of extending SweFN++ (beyond the purely manual). A reasonable route forward would be one where we would be able to continue to draw on the rich lexical

[26] https://spraakbanken.gu.se/en/resources/sensaldo

knowledge already present in SweFN++, combining it with the rapidly evolving corpus-based machine-learning methodologies – in particular deep-learning approaches – which are the methods of choice for tackling natural language understanding problems in LT at present.

Funding

Saldo and the other lexical resources described in this chapter have been developed with public funding. After 2003, the University of Gothenburg, through its support of the Centre for Language Technology and of Språkbanken, has financed the main part of the work on Saldo. During 2006–2008, the development of Saldo was partly supported by the Swedish Research Council project *Library-Based Grammar Engineering* (grant 2005–4211). From 2008 onward part of the funding has come from the Swedish Research Council, through the projects *Safeguarding the future of Språkbanken* (grant 2007–7430), *Swedish FrameNet++* (grant 2010–6013), *Knowledge-based culturomics* (grant 2012–05738), *Corpus-driven induction of linguistic knowledge* (grant 2013–04944), *Swe-Clarin* (grant 2013–02003), and *The National Swedish Language Bank* (grant 2017–00626), and by the European Commission through the project *META-NORD* (grant 270899).

References

Apresjan, Yuri D. 2002. Principles of systematic lexicography. In Marie-Hélène Corréard (ed.), *Lexicography and natural language processing: A Festschrift in honour of B. T. S. Atkins*, 91–104. Grenoble: EURALEX.

Borin, Lars. 2005. Mannen är faderns mormor: *Svenskt associationslexikon* reinkarnerat [The man is the grandmother of the father: The *Swedish Associative Thesaurus* reincarnated]. *LexicoNordica* 12: 39–55.

Borin, Lars & Markus Forsberg. 2009. All in the family: A comparison of SALDO and WordNet. In *Proceedings of the Nodalida 2009 workshop WordNets and other lexical semantic resources – between lexical semantics, lexicography, terminology and formal ontologies*, 7–12. Odense: NEALT.

Borin, Lars & Markus Forsberg. 2010. From the People's Synonym Dictionary to fuzzy synsets – first steps. In *Proceedings of the LREC 2010 workshop Semantic relations: Theory and applications*, 18–25. Valletta: ELRA.

Borin, Lars & Markus Forsberg. 2011. Swesaurus – ett svenskt ordnät med fria tyglar [Swesaurus – a Swedish wordnet with free reins]. *LexicoNordica* 18: 17–39.

Borin, Lars & Markus Forsberg. 2014. Swesaurus; *or*, The Frankenstein Approach to Wordnet Construction. In *Proceedings of GWC 2014*, 215–223. Tartu: University of Tartu Press.

Borin, Lars, Markus Forsberg & Lennart Lönngren. 2008. The hunting of the BLARK – SALDO, a freely available lexical database for Swedish language technology. In Joakim Nivre, Mats Dahllöf & Beáta Megyesi (eds.), *Resourceful language technology: Festschrift in honor of Anna Sågvall Hein*, 21–32. Uppsala: Uppsala University, Department of Linguistics & Philology.

Borin, Lars, Markus Forsberg & Lennart Lönngren. 2013. SALDO: A touch of yin to WordNet's yang. *Language Resources and Evaluation* 47(4): 1191–1211. https://doi.org/10.1007/s10579-013-9233-4

Borin, Lars, Markus Forsberg & Johan Roxendal. 2012. Korp – the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012*, 474–478. Istanbul: ELRA.

Borin, Lars, Luis Nieto Piña & Richard Johansson. 2015. Here be dragons? The perils and promises of inter-resource lexical-semantic mapping. In *Proceedings of the Workshop on semantic resources and semantic annotation for natural language processing and the digital humanities at Nodalida 2015*, 1–11. Linköping: LiUEP.

Boyd-Graber, Jordan, Christiane Fellbaum, Daniel Osherson & Robert Schapire. 2006. Adding dense, weighted connections to WordNet. In *Proceedings of GWC 2006*, 29–35. Brno: Masaryk University.

Bring, Sven Casper. 1930. *Svenskt ordförråd ordnat i begreppsklasser* [Swedish vocabulary arranged in conceptual classes]. Stockholm: Hugo Gebers förlag.

Carstairs-McCarthy, Andrew. 1999. *The origins of complex language*. Oxford: Oxford University Press.

Civil, Miguel. 1990. Sumerian and Akkadian lexicography. In Oskar Reichmann Hausmann Franz Josef and, Herbert Ernst Wiegand & Ladislav Zgusta (eds.), *Dictionaries: An international encyclopedia of lexicography. Second volume*, 1682–1686. Berlin: Walter de Gruyter.

Cruse, D. Alan. 2000. Aspects of the micro-structure of word meanings. In Yael Ravin & Claudia Leacock (eds.), *Polysemy: Theoretical and computational approaches*, 30–51. Oxford: Oxford University Press.

Döderlein, Ludwig. 1863. The author's preface. In *Döderlein's hand-book of Latin synonymes. Translated by rev. H.A. Arnold, B.A., with an introduction by S.H. Taylor, LL.D*, ix–xvi. Andover: Warren F. Draper.

Erk, Katrin. 2010. What is word meaning, really? (And how can distributional models help us describe it?) In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, 17–26. Uppsala: ACL.

Fellbaum, Christiane (ed.). 1998. *WordNet: An electronic lexical database*. Cambridge: MIT Press. https://doi.org/10.7551/mitpress/7287.001.0001

Forsberg, Markus. 2007. Three tools for language processing: BNF converter, Functional Morphology, and Extract. Göteborg University & Chalmers University of Technology. (PhD thesis).

Goddard, Cliff (ed.). 2008. *Cross-linguistic semantics*. Amsterdam: John Benjamins. https://doi.org/10.1075/slcs.102

Goddard, Cliff & Susanna Karlsson. 2008. Re-thinking *think* in contrastive perspective: Swedish vs. English. In Cliff Goddard (ed.), *Cross-linguistic semantics*, 225–240. Amsterdam: John Benjamins. https://doi.org/10.1075/slcs.102.14god

Hanks, Patrick. 2000. Do word meanings exist? *Computers and the Humanities* 34(1–2): 205–215. https://doi.org/10.1023/A:1002471322828

Hanks, Patrick. 2013. *Lexical analysis: Norms and exploitations*. Cambridge: MIT Press. https://doi.org/10.7551/mitpress/9780262018579.001.0001

Hill, Felix, Roi Reichart & Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics* 41(4): 665–695. https://doi.org/10.1162/COLI_a_00237

Hüllen, Werner. 2004. *A history of Roget's Thesaurus: Origins, development, and design*. Oxford: Oxford University Press.

Järborg, Jerker. 2001. *Roller i Semantisk databas* [Roles in the Semantic database] *(research reports from the department of swedish no. gu-iss-01-3)*. Research report. Gothenburg: University of Gothenburg, Dept. of Swedish.

Jarmasz, Mario & Stan Szpakowicz. 2004. *Roget's Thesaurus* and semantic similarity. In Nicolas Nicolov, Kalina Bontcheva, Galia Angelova & Ruslan Mitkov (eds.), *Recent advances in natural language processing III. Selected papers from RANLP 2003*, 111–120. Amsterdam: John Benjamins. https://doi.org/10.1075/cilt.260.12jar

Jobbins, Amanda C. & Lindsay J. Evett. 1995. Automatic identification of cohesion in texts: Exploiting the lexical organization of Roget's Thesaurus. In *Proceedings of Rocling VIII*, 111–125. Taipei.

Kann, Viggo & Magnus Rosell. 2006. Free construction of a free Swedish dictionary of synonyms. In *Proceedings of Nodalida 2005*, 105–110. Joensuu: University of Joensuu.

Kennedy, Alistair & Stan Szpakowicz. 2008. Evaluating *Roget's* thesauri. In *Proceedings of ACL/HLT 2008*, 416–424. Columbus: ACL.

Kennedy, Alistair & Stan Szpakowicz. 2014. Evaluation of automatic updates of *Roget's Thesaurus*. *Journal of Language Modelling* 2(2): 1–49. https://doi.org/10.15398/jlm.v2i1.78

Kilgarriff, Adam. 1997. I don't believe in word senses. *Computers and the Humanities* 31(2): 91–113. https://doi.org/10.1023/A:1000583911091

Lange, Sven. 2007. *Thesaurus Lex – ett hyperlexikon med rötter hos Locke, Roget och Bring* [Thesaurus Lex – a hyperlexicon with roots in Locke, Roget, and Bring]. http://www.thesauruslex.se/artiklar/Roget.pdf

Leont'ev, Aleksej A. 1977. *Slovar' associativnyx norm russkogo jazyka* [Dictionary of Russian associative norms]. Moscow: Izdatel'stvo moskovskogo universiteta.

Lieber, Francis. 1841. Preface of the translator. In *Dictionary of Latin synonymes, for the use of schools and private students, with a complete index. by Lewis [Ludwig] Ramshorn. From the German by Francis Lieber*, iii–viii. Charles C. Little & James Brown.

Lönngren, Lennart. 1988a. Lexika, baserade på semantiska relationer [Lexicons based on semantic relations]. In *Proceedings of Nodalida 1987*, 229–236. Copenhagen: Copenhagen Business School.

Lönngren, Lennart. 1988b. *Svenskt associationslexikon*. Research report UCDL-R-88-2. Uppsala: Uppsala University, Center for Computational Lingistics.

Lönngren, Lennart. 1989. *Svenskt associationslexikon: Rapport från ett projekt inom datorstödd lexikografi*. Research report UCDL-R-89-1. Uppsala: Uppsala University, Center for Computational Lingistics.

Lönngren, Lennart. 1992. *Svenskt associationslexikon. Del I–IV*. Research report. Uppsala: Uppsala University, Dept. of Lingvistics.

Lönngren, Lennart. 1998. A Swedish associative thesaurus. In *Proceedings of EURALEX 1998, Vol. 2*, 467–474. Liège: University of Liège.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross & Katherine J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography* 3(4): 235–245. https://doi.org/10.1093/ijl/3.4.235

Morris, Jane & Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics* 17(1): 21–48.

Morris, Jane & Graeme Hirst. 2004. Non-classical lexical semantic relations. In *Proceedings of the Workshop on computational lexical semantics at NAACL/HLT 2004*, 46–51. Boston: ACL. https://doi.org/10.3115/1596431.1596438

Murphy, M. Lynne. 2003. *Semantic relations and the lexicon*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511486494

Roget, Mark Peter. 1852. *Thesaurus of English words and phrases*. London: Longman.

Rosch, Eleanor, Carolyn B. Mervis, Wayne D. Gray, David M. Johnson & Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive Psychology* 8(3): 382–439. https://doi.org/10.1016/0010-0285(76)90013-X

Rouces, Jacobo, Nina Tahmasebi, Lars Borin & Stian Rødven Eide. 2018. SenSALDO: Creating a sentiment lexicon for Swedish. In *Proceedings of LREC 2018*, 4192–4198. Miyazaki: ELRA.

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

Wurzel, Wolfgang Ulrich. 1989. *Inflectional morphology and naturalness*. Dordrecht: Kluwer.

Zechner, Niklas & Lars Borin. 2020. Towards a Swedish Roget-style thesaurus for NLP. In *Proceedings of the Globalex workshop on linked lexicography*, 53–60. Marseille: ELRA.