From the People's Synonym Dictionary to fuzzy synsets – first steps

Lars Borin and Markus Forsberg

Språkbanken, University of Gothenburg, Sweden
lars.borin@svenska.gu.se, markus.forsberg@gu.se

Abstract

We present our ongoing work on creating fuzzy synsets for Swedish using the lexical resources Synlex and SALDO. Synlex is a graded synonym list 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. SALDO is a full-scale Swedish lexical-semantic resource with non-classical, associative relations among word and multiword senses, identified by persistent formal identifiers. We discuss two approaches for mapping Synlex synonym pairs to SALDO senses – transitive closure and clique formation – as well as our planned work for including other kinds of classical lexical-semantic relations from various existing free lexical resources, into Swesaurus, a multi-faceted resource for Swedish combining classical wordnet-type relations with the associative thesaurus relations from SALDO.

1. Introduction

The Princeton WordNet (WN; Fellbaum 1998) and other wordnets created in its image are standard items in any modern language technology resource toolkit. Notwithstanding their widespread use and general popularity in language technology research and applications, some of the decisions that shaped WN are debatable at least from a lexicographical and linguistic point of view. Most often, the unclear theoretical status of the notion of synonymy is pointed out (e.g., Ci 2008; Piasecki et al. 2009).

Since the synonymy relation is the basis of the whole wordnet endeavor, defining as it does the central entity of Princeton-type wordnets, the *synset*, any flaw in this concept will call into question the foundations of the whole wordnet enterprise. Relevant in this connection, there is 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 organization is based on the notion of synonymy almost by definition cannot present a faithful reflection of our lexical knowledge, at least not from a linguistic point of view.

WN synonyms, as originally defined, should be interchangable in some contexts, but not necessarily in all contexts (Miller, 1998, 24); in fact, even one context is enough (Alonge et al., 1998, 22). This indicates that synonymy in the WN sense may not correspond exactly to how linguists and lexicographers understand this term, and further that it may be a matter of degree – for instance expressible as the number of possible substitution contexts of a particular synonym pair. To the best of our knowledge, this interesting notion has never been explored with respect to wordnets. But if this is the case, and if we had some practicable means of quantifying the degree of synonymy among words, then we could actually define a kind of wordnet based on this, where synsets could grow or shrink, depending on the degree of synonymy that we require for a particular purpose.

The work described below represents an attempt to accomplish exactly this. In Språkbanken, a language technology R&D unit at the University of Gothenburg, We have started work on a 'fuzzy wordnet' for Swedish, understood here as a wordnet based on 'fuzzy', or graded, synsets. This endeavor is made feasible by the previous existence of a number of freely available lexical resources on which we can draw in our work. The work is in its initial stages, so what we can offer in this paper are some preliminary results of automatic merging of two unique lexical resources, together with a discussion of a number of interesting theoretical issues that arise from this work.

The two lexical resources under discussion here are Synlex and SALDO.

2. Synlex

Graded synonymy relations for part of the Swedish vocabulary are available in *Synlex* (the People's Synonym Lexicon; Kann and Rosell 2006). This lexical resource 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 a large number of judgements is available for download under an open-source license. The latest version of the list at the time of writing is

dated 2009-05-29, and contains 18,607 graded synonym pairs (37,214 when symmetry of synonymy is taken into account).

The members of these pairs are words (i.e., text word forms) – not even part of speech (PoS) is indicated – mainly dictionary base forms (lemmas), but sometimes inflected forms, and in some cases multiword units (MWUs). One problem then becomes, in the case of a word having as synonyms several other words – because of homonymy and polysemy – to determine how many synsets we are dealing with. Also, for those familiar with WN, we should add that Synlex contains words of all PoS, and synonymy relations are sometimes between words with different PoS, just as in EuroWordNet.[1]

3. SALDO

SALDO (Borin, 2005; Borin and Forsberg, 2009; Borin et al., 2008; Borin and Forsberg, 2008), or SAL version 2, is a free modern Swedish semantic and morphological lexicon intended for language technology applications. The lexicon is available under a Creative Commons Attribute-Share Alike license and LGPL 3.0.

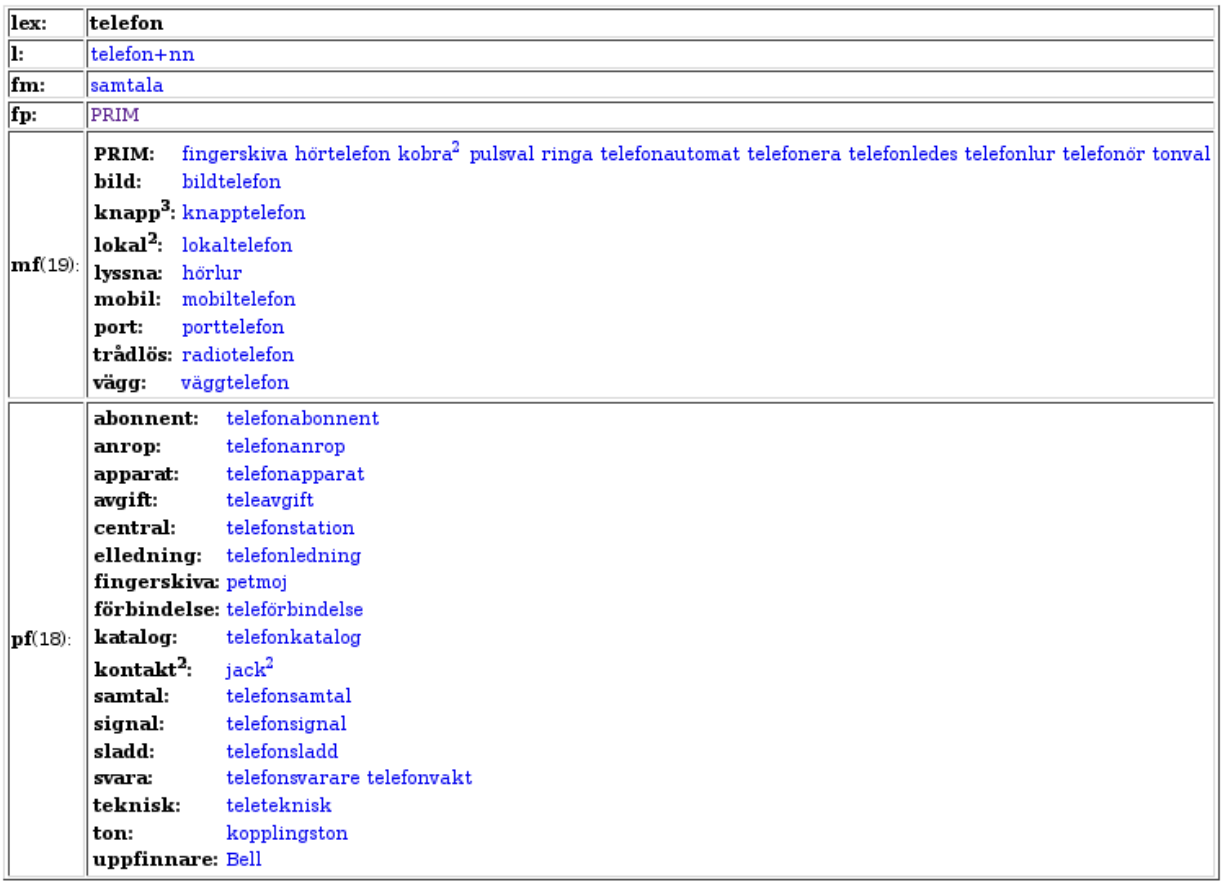
SALDO started its life as *Svenskt associationslexikon* (Lönngren, 1992) – 'The Swedish Associative Thesaurus' – a so far relatively unknown Swedish thesaurus with an unusual semantic organization, reminiscent of, but different from that of WordNet (Borin and Forsberg, 2009). SAL has been published in 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 names 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 76,200 entries, the increased number being because a number of 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.

The central semantic relations of SALDO are based on *association*, a "non-classical" lexical-semantic relation (Morris and Hirst, 2004). SALDO describes *all* words semantically, not only the open word classes. It is organized by two primitive semantic (association) relations, one obligatory and one optional. Every entry must have a *mother* (or *main descriptor*), a semantically closely related entry which is more central, i.e., semantically and/or morphologically less complex, probably more frequent, stylistically less marked and acquired earlier in first and second language acquisition, etc. The mother will in practice often be either a hyperonym or synonym of the headword. However, it need not be either: Sometimes it is an antonym, and quite often it is a different part of speech from the headword, which takes us outside the realm of traditional lexical-semantic relations. An artificial most central entry, PRIM, is used as the mother of 50 semantically unrelated entries at the top of the hierarchy, making all of SALDO into a single rooted tree. An entry may also have an additional descriptor, the *father* (or *supplementary descriptor*), which serves to further characterize the entry semantically. The mother and father relations can then form the basis of any number of derived relations. Thus the m-sibling relation – 'having a common mother' – is very interesting, as such sibling groups tend to correspond to natural semantic groupings. Figure 1 shows how the SALDO entry for the Swedish noun *telefon* 'telephone' is associated to a number of other words: *samtala* 'hold a conversation' is the mother of *telefon*, while *telefonledes* 'by phone', *ringa* 'call v.', *mobiltelefon* 'mobile phone', *pulsval* 'pulse dialling' and a number of others are m-siblings having *telefon* as their mother. In the p-sibling group of *telefon* (senses having *telefon* as their father), we find *telefonkatalog* 'phone directory', *telefonsvarare* 'answering machine', the proper name *Bell* and a number of others.

We soon realized that in order to be useful in language technology applications, SAL would have to be provided at least with part-of-speech and inflectional morphological information – both entirely absent from SAL in its original form – and SALDO was created. The morphological component of SALDO has been defined using Functional Morphology (FM) (Forsberg and Ranta, 2004; Forsberg, 2007), a tool that provides a development environment for computational morphologies. It is a tool with a flexible language for

[1] Although in EuroWordNet this kind of synonymy is still formally distinct from within-PoS synonymy, bearing the label XPOS_NEAR_SYNONYM (Alonge et al., 1998, 25ff).



lex:	**telefon**
l:	telefon+nn
fm:	samtala
fp:	PRIM
mf(19):	**PRIM:** fingerskiva hörtelefon kobra[2] pulsval ringa telefonautomat telefonera telefonledes telefonlur telefonör tonval **bild:** bildtelefon **knapp**[3]**:** knapptelefon **lokal**[2]**:** lokaltelefon **lyssna:** hörlur **mobil:** mobiltelefon **port:** porttelefon **trådlös:** radiotelefon **vägg:** väggtelefon
pf(18):	**abonnent:** telefonabonnent **anrop:** telefonanrop **apparat:** telefonapparat **avgift:** teleavgift **central:** telefonstation **elledning:** telefonledning **fingerskiva:** petmoj **förbindelse:** teleförbindelse **katalog:** telefonkatalog **kontakt**[2]**:** jack[2] **samtal:** telefonsamtal **signal:** telefonsignal **sladd:** telefonsladd **svara:** telefonsvarare telefonvakt **teknisk:** teleteknisk **ton:** kopplingston **uppfinnare:** Bell

Figure 1: Semantic (associative) relations for *telefon* 'telephone n.' in SALDO (rendered in blue/non-bold)

defining morphological rules together with a platform for testing, which is used to minimize the risk of resource degradation during development. Furthermore, it has a rich export system, targetting around 20 formats, and supports both (compound) analysis and synthesis.

SALDO is, as one of its distribution channels, published as web services, updated daily. Web services provide clean interfaces and instant updates, but are restricted to small amounts of data because of network latency. Presently available web services include incremental fullform lookup, semantic lookup, compound analysis, and an inflection engine service. See <http://spraakbanken.gu.se/eng/saldo>.

4. From Synlex and SALDO to fuzzy synsets

Importantly to our purposes here, the basic units of SALDO are uniquely identified *word senses*. The current version of SALDO contains some 73,400 senses. Consequently, it is easy to find an answer to the question: "How many senses does a particular base form have?"[2] We can simply make an automated comparison between words in Synlex and word senses in SALDO via the Synlex words. From the point of view of Synlex, such a comparison yields five interesting sets, for a word w_i in Synlex (on the assumption – simplifying but largely correct – that Synlex contains pairs of base forms, including MWUs):

1. w_i is not a base form in SALDO
2. w_i occurs once in Synlex and it has one sense in SALDO
3. w_i occurs once in Synlex and it has several senses in SALDO
4. w_i occurs in several pairs in Synlex and it has one sense in SALDO
5. w_i occurs in several pairs in Synlex and it has several senses in SALDO

Calculating the set of Synlex pairs, such that each member of every pair is in one of set 2 or set 4 above – i.e., pairs where both members have only one SALDO sense – should then allow us to go on to calculate fuzzy synsets of various degrees from this set. Performing the first calculation yielded an initial set of 9,236 pairs, i.e., a bit less than half of Synlex (see the lower half of table 1). In the result set we replace each Synlex word form w_i with the corresponding SALDO word sense identifier l_i. For convenience we also mul-

[2]Base forms in SALDO include multi-word units.

set	size (Synlex entries)
1: 1 – 0	4,909
2: 1 – 1	4,779
3: 1 – many	645
4: many – 1	20,784
5: many – many	6,097
total 1–5	**37,214**

pair type	size (number of pairs)
set 2 – set 2	1,254
set 2 – set 4	2,144
set 4 – set 2	2,144
set 4 – set 4	6,097
total	**18,472**

Table 1: Connecting Synlex to SALDO

tiply the synonymy degree by a factor 20 in these pairs, making it range from 60 to 100.

In this paper, we report on our work on this subset of the Synlex entries, as part of a recently initiated project with the aim of bootstrapping a fuzzy wordnet for Swedish from Synlex and other available lexical resources (see section 9 below).

We have experimented with two different methods for constructing fuzzy synsets from Synlex: transitive closure (next section) and cliques (section 6).

5. Synset construction by transitive closure

Our first algorithm for building fuzzy synsets is a straightforward computation of the transitive closures of the word sense pairs, as follows. For every graded word sense pair with a degree higher or equal to d_{cutoff}, we check membership of the word senses in the current result set of synsets $Synsets$, and make the necessary adjustments to this set based upon their membership; in pseudocode:

$$
\begin{array}{l}
Synsets = \{\} \\
for\ \langle\langle l_i, l_j\rangle, d_k\rangle \in Synlex_{saldo} \\
\quad d_k \geq d_{cutoff} \\
\quad case\ membership(\langle l_i, l_j\rangle, Synsets)\ of \\
\quad\quad \langle S_1, S_2\rangle \Rightarrow Synsets.merge(S_1, S_2) \\
\quad\quad \langle S_1, \{\}\rangle \Rightarrow Synsets.add(l_j, S_1) \\
\quad\quad \langle \{\}, S_2\rangle \Rightarrow Synsets.add(l_i, S_2) \\
\quad\quad \langle \{\}, \{\}\rangle \Rightarrow Synsets.new(\{l_i, l_j\}) \\
return\ Synsets
\end{array}
$$

In other words, we calculate the synsets of degree d_{cutoff} by collecting in the same set all l_i that are connected by some path of graded synonymy relations where no relation has a degree less than d_{cutoff}.

The calculation of the transitive closure carries the hidden assumption that implicitly derived pairs are valid at the same degree as the pairs they are derived from. This assumption turns out to be rather problematic (see section 7 below).

$d \geq$	$\lvert SS\rvert$	$\lvert S\rvert = 2$	$2 < \lvert S\rvert \leq 25$	$\lvert S\rvert > 25$	$\max \lvert S\rvert$
60	1,485	951	530	4	3,893
70	1,641	1,026	602	13	1,245
80	1,640	1,066	566	8	441
90	1,068	800	268	0	18
100	416	362	54	0	6

Table 2: Synsets computed with transitive closure

$d \geq$	$\lvert SS\rvert$	$\lvert S\rvert = 2$	$2 < \lvert S\rvert \leq 25$	$\lvert S\rvert > 25$	$\max \lvert S\rvert$
60	1,533	956	560	17	1,598
70	1,650	1,034	602	14	921
80	1,609	1,047	556	6	397
90	1,016	761	255	0	18
100	394	347	47	0	5

Table 3: Synsets computed with PoS-constrained transitive closure

Applying the transitive closure algorithm to our 9,236 Synlex pairs[3] yields the results presented in table 2, to be read as follows: d is the degree; $|SS|$ is the number of synsets; $|S| = 2$ is the number of synsets of size 2; $2 < |S| \leq 25$ is the number of synsets of a reasonable size; $|S| > 25$ is the number of synsets of a suspiciously large size; and $\max |S|$ is the size of the largest synset.

The largest synset in this table is salient – at degree 60 it is as large as 3,893. This is an indication of a couple of things: first that the basic assumption about the implicit pairs is too strong, but also, that Synlex contains word senses that are missing in SALDO or Synlex has pairs that are simply wrong. In all cases we have pairs that merge reasonable synsets into a huge one. Interestingly, the smaller sized synsets are reasonable on manual inspection, which indicates that Synlex generally provides us with good information.

As a heuristic filter, we kept only same-PoS pairs, which resulted in the removal of 484 pairs, and repeated the experiment with the new set of 8,752 pairs. The result is presented in table 3, where the size of the largest synset at degree 60 has been halved by this filtering process, but the size is still significant.

6. Clique-based synset construction

A *clique* is a graph theoretic notion that describes a subgraph of a graph where all nodes are connected to all other nodes in the subgraph. If we require that all synsets are cliques, then we avoid the assumption about the implicit pairs.

The algorithm for creating cliques is simple: for every SALDO sense occurring in Synlex we create synsets by iteratively adding new lexemes that are connected to all previous ones.

[3]The number of SALDO word sense identifiers in this set is 8,594, i.e., the average synset size is slightly above 2.

$d \geq$	$\lvert SS\rvert$	$\lvert S\rvert = 2$	$2 < \lvert S\rvert \leq 25$	$\lvert S\rvert > 25$	$\max \lvert S\rvert$
60	6,933	5,687	1,246	0	22
70	4,931	4,313	618	0	16
80	4,024	3,673	351	0	16
90	1,582	1,542	40	0	11
100	484	482	2	0	7

Table 4: Synset cliques

$$
\begin{array}{l}
Synsets = \{\} \\
for\, l_i \in SALDO_{Synlex} \\
\quad SS = \{\{l_i\}\} \\
\quad while\, SS.exts\_exists(Synlex_{saldo}, d_{cutoff}) \\
\quad\quad for\, S\, in\, SS.has\_exts(Synlex_{saldo}, d_{cutoff}) \\
\quad\quad\quad ES = S.extensions(Synlex_{saldo}, d_{cutoff}) \\
\quad\quad\quad SS.extend(S, ES) \\
\quad Synsets.add(SS) \\
return\, SS
\end{array}
$$

A comment is in order here: Since it is possible that l_i is in more than one synset, with the operation `extend` we build a new synset for every possible extension. A natural optimization of the algorithm would be to divide the extension set ES into cliques to avoid rebuilding the same synset.

The result of running the algorithm on our material is given in table 4, and looks initially very promising, since there are no oversized synsets.

However, this algorithm has a hidden assumption, namely that all relevant pairs have been graded. This is not true for Synlex, which has the effect that the synsets end up being small, and worse, that some senses appear in many synsets, which strictly speaking is a contradiction in terms, since synsets and senses are two sides of the same coin on the WN view of things, so that a particular sense of a lemma should appear in only one synset. In some cases this is an indication that Synlex has a more fine-grained sense description than SALDO (see below), but in many cases it is an invalid split caused by a missing pair, i.e., a synonymy judgement missing from Synlex.

7. Degree computation for implicit pairs

A natural question at this point is: Is it possible to calculate new degrees for the implicitly derived pairs computed by transitive closure, i.e., given (l_1, l_2, d_1) and (l_2, l_3, d_2), could we not calculate a reasonable degree for (l_1, l_3) from d_1 and d_2?

In general, this is not a simple problem, and we will illustrate why with two pairs taken from Synlex:

integration	anpassing	60
anpassning	integrering	60

Here, *integration* 'integration' is related with the lowest Synlex degree to *anpassning* 'adaptation', which in turn is related to *integrering* 'integrating n.', again with the lowest Synlex degree.

What would be a reasonable degree for the derived pair *integration – integrering*? As already indicated by the glossing, they are completely synonymous, or nearly so, but there is no way to calculate this information from these two pairs.

Naturally, we have discussed various possible ways of performing this calculation. Since synonymy is normally considered both symmetric and transitive, there ought to be a standard way of calculating the degree of transitively derived synonymy even with graded synonyms. Interpreting degrees as distances in the plane and calculating the Euclidian distance under the assumption that the two synonymy links are at right angles to each other could possibly yield good results, on average (but not in this particular case, of course).

Another possibility, proposed here, is to annotate every synset with some standard statistical measures that reflect the composition of the grades, which can be used as basis for the calculation of the implicit links in the synset. Currently, we use *mean*, *standard deviation*, *min*, and *max*, as defined in this figure:

$$
\begin{array}{ll}
\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_i - \mu)^2} & \mu = \frac{1}{n}\sum_{i=1}^{n} d_i \\
min = \min_i\{d_i\} & max = \max_i\{d_i\}
\end{array}
$$

As indicated above in section 5, we use the min measure in calculating the transitive closure of Synlex pairs, but we calculate all the measures for the resulting synsets (see figure 2, from <http://spraakbanken.gu.se/eng/swefn>)

8. Discussion

As we saw above, both methods for fuzzy synset construction have some drawbacks. However, the clique method has the more serious drawback – at least in our view – that in order for it to be used for synset construction, where a sense only occurs in one synset, we would need to somehow add information that is missing from Synlex, namely about the synonymy of unjudged pairs. For this reason, we have decided to adopt the transitive closure approach in our continued work. In fact, this approach offers a way of computing the synonymy of such pairs, as we discussed in the previous section, so that clique computation could be added as a refinement on top of it.

In working with the initial set of Synlex pairs and the fuzzy synset candidates automatically derived from this set by the transitive closure approach,

avg: 100 % **dev**: 0 % **min**: 100 % **max**: 100 %	anlita..1 leja..1
avg: 72 % **dev**: 0 % **min**: 72 % **max**: 72 %	annars..1 eljest..1
avg: 60 % **dev**: 0 % **min**: 60 % **max**: 60 %	annonsering..1 reklam..1
avg: 70 % **dev**: 14 % **min**: 60 % **max**: 90 %	integrering..1 integration..1 adaptation..1 anpassning..1
avg: 72 % **dev**: 8 % **min**: 64 % **max**: 80 %	tillreda..1 anrätta..1 tillaga..1
avg: 78 % **dev**: 16 % **min**: 64 % **max**: 100 %	nylle..1 nuna..1 anlete..1 ansikte..1
avg: 72 % **dev**: 0 % **min**: 72 % **max**: 72 %	anskri..1 stridsrop..1

Figure 2: Some synsets at threshold 60 (from <http://spraakbanken.gu.se/eng/swefn>)

we have been subjecting the result to constant manual evaluation, drawing upon our long experience of Swedish lexicography.

In its subdivision of lemmas into senses, SALDO reflects a well-established Swedish lexicographical tradition, which for practical reasons has tended to avoid excessively fine-grained sense distinctions. In paper dictionaries compiled in this tradition, definitions tend for this reason to be couched in very general terms, in order to cover as many different usages as possible.

WN, on the other hand, is all about different usages. The practice of defining synonymy as substitutability in at least one context, and then for all practical purposes defining senses using synsets, has the practical consequence that whenever we find that a word w_i can be substituted for another word w_j in one particular context, we will probably have to postulate a new sense both for w_i and w_j. Hence, WN lemmas by design will tend to have many senses (Vossen, 1998, 9). It seems that if we want manageable synsets, we have to accept a fine granularity of senses (as they are understood in WN).

The result presented above actually comes out of an iterative process where we have tried to identify problematic pairs using two simple diagnostics:

1. We examined senses with many connections to other senses, since many connections may be an indication that two or more senses have been collapsed into one in SALDO.
2. We inspected pairs that merged already large synsets, as a result of lowering the threshold for synset inclusion. The pairs that connect two large synsets together are not problematic in themselves, but they have the potential gain of reducing the size of the synsets drastically.

In both cases, the action to be taken is one of: 1. add a word sense to SALDO; 2. remove the pair from Synlex; 3. do nothing. In practice, all three have been necessary. The work with Synlex has thus allowed us to refine the semantic structure of SALDO in the direction of actual usage, which should be beneficial in a resource intended to be used in language processing. Reconciling the senses of SALDO – reflecting deep lexicographic thinking about words – with those of Synlex – a noisy and 'anarchistic' resource – will certainly raise many tricky and theoretically interesting problems, for lexicography and language technology alike.

9. Conclusion: Towards Swesaurus – a fuzzy wordnet for Swedish

A few thousand synsets do not a wordnet make. The work described above represents the first steps towards a Swedish lexical-semantic resource which we call Swesaurus, where our goal is to add classical lexical-

semantic relations and fuzzy synsets to the existing associative thesaurus structure of SALDO, thus combining classical and non-classical lexical-semantic relations in one resource.[4]

In this ongoing work, we can draw upon a number of other existing lexical resources, e.g.:

- Thus, we have extracted the lexical-semantic relations encoded in a conventional print dictionary, or rather, the database underlying this dictionary, where we find about 12,000 sense pairs explicitly labeled with one of five classical lexical-semantic relations: synonymy, hyponymy, hyperonymy, antonymy, cohyponymy, plus the 'lexicographic' relation often rendered as "see" in dictionaries. Computing the transitive closure of these sense pairs yields some 20,000 additional pairs, i.e., about 30,000 in total.
- The Swedish Wiktionary (close to 50,000 entries) provides lexical-semantic relations – e.g., synonymy, antonymy, "related words" – for a subset of its entries. This free resource also contains definitions of the senses, which we cannot get from other sources.
- We have further a lexical resource consisting of pairings of word senses from the same dictionary database mentioned above, with automatically extracted headwords of their dictionary definitions. Even though there are many invalid items in this extensive list (52,800 pairs), we believe that we can clean it with mostly automatic processing, using the other resources that we have at our disposal.

In fact, a decided advantage for our work is the fact that we can utilize several lexical resources containing overlapping information. This means that one resource can be used to disambiguate ambiguous information in another resource. For example, as mentioned above, the mother, or primary descriptor of a SALDO sense is in practice often a synonym or hyperonym of the sense. For those synonym pairs in Synlex – almost half – which have not been mapped to SALDO sense identifiers, because of a one-to-many or many-to-many mapping between the Synlex string and SALDO senses, we will use the heuristic that if the pair can be matched to a child-mother configuration in SALDO, the corresponding sense(s) will be chosen for the ambiguous member(s) of the pair, the reasoning being that if (the lemmas of) the senses of a SALDO child-mother relation map to a Synlex pair, then it is very likely that this particular mother happens to be a synonym of the child (although we could be wrong).[5]

Using such a strategy, we believe that we will be able to bootstrap a wordnet-like extension to SALDO, achieving a decent-quality resource with a fairly modest amount of manual work, because we are in the fortunate position of actually being able to reap the fruits of a large collected human effort that has gone into the creation of the existing resources we have at our disposal.[6]

10. Acknowledgements

The research presented here was supported by the Swedish Research Council (the project *Safeguarding the future of Språkbanken* 2008–2010, VR dnr 2007-7430) and the University of Gothenburg through its support of the Centre for Language Technology (the *Swedish FrameNet++* project) and through its support of Språkbanken (the Swedish Language Bank).

We would like to express our gratitude to the anonymous referees for their constructive remarks and insightful questions, which have contributed substantially to improving the presentation in this paper.

11. References

Antonietta Alonge, Nicoletta Calzolari, Piek Vossen, Laura Bloksma, Irene Castellon, Maria Antonia Marti, and Wim Peters. 1998. The linguistic design of the EuroWordNet database. In Piek Vossen, editor, *EuroWordNet: A multilingual database with lexical semantic networks*, pages 19–43. Kluwer, Dordrecht.

Lars Borin and Markus Forsberg. 2008. Saldo 1.0 (svenskt associationslexikon version 2). Språkbanken, Göteborgs universitet.

Lars Borin and Markus Forsberg. 2009. All in the family: A comparison of SALDO and WordNet. Odense.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2008. The hunting of the BLARK – SALDO, a

[4] Incidentally, this is the reverse of what is going on with the Princeton WordNet at this moment, where associative relations are being added (called "evocation" on the WN website).

[5] Another possibility is to first map the print dictionary database synonym pairs (and possibly its "see" relation pairs as well, since we are dealing with graded synonymy in Synlex) to SALDO, and then match this list against the remaining Synlex pairs. Again this would be an example of 'synergistic' merging of several existing lexical resources.

[6] Unfortunately, an existing embryonic Swedish wordnet is not one of those resources, since it has not been developed since about 2002, and further its licensing format prohibits its inclusion in the open source resource that we wish to develop.

freely available lexical database for Swedish language technology. In Joakim Nivre, Mats Dahllöf, and Beata Megyesi, editors, *Resourceful language technology. Festschrift in honor of Anna Sågvall Hein*, number 7 in Acta Universitatis Upsaliensis: Studia Linguistica Upsaliensia, pages 21–32. Uppsala University, Department of Linguistics and Philology, Uppsala.

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

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

Jiwei Ci. 2008. Synonymy and polysemy. In Patrick Hanks, editor, *Lexicology: Critical concepts in linguistics. Vol. III: Core meaning, extended meaning*, pages 191–207. Routledge, London. Reprinted from Lingua 72 (1987): 315–331.

Christiane Fellbaum, editor. 1998. *WordNet: An electronic lexical database*. MIT Press, Cambridge, Mass.

Markus Forsberg and Aarne Ranta. 2004. Functional morphology. In *ICFP'04. Proceedings of the ninth ACM SIGPLAN international conference of functional programming*, Snowbird, Utah. ACM.

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

Viggo Kann and Magnus Rosell. 2006. Free construction of a free Swedish dictionary of synonyms. In *Proceedings of the 15th NODALIDA conference, Joensuu 2005*, pages 105–110. Department of Linguistics, University of 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.

George A. Miller. 1998. Nouns in WordNet. In Christiane Fellbaum, editor, *WordNet: An electronic lexical database*, pages 23–46. MIT Press, Cambridge, Mass.

Jane Morris and Graeme Hirst. 2004. Non-classical lexical semantic relations. In Dan Moldovan and Roxana Girju, editors, *HLT-NAACL 2004: Workshop on Computational Lexical Semantics*, pages 46–51, Boston. ACL.

Maciej Piasecki, Stanisław Szpakowicz, and Bartosz Broda. 2009. *A wordnet from the ground up*. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław.

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

Piek Vossen. 1998. Introduction to EuroWordNet. In Piek Vossen, editor, *EuroWordNet: A multilingual database with lexical semantic networks*, pages 1–17. Kluwer, Dordrecht.