Compounding in a Swedish Blog Corpus

Robert Östling and Mats Wirén

November 1, 2012

1 Introduction and background

1.1 The problem

Research in compounding for Swedish has a long tradition at Stockholm University, with Benny Brodda starting already in 1967/68 (Brodda, 1981, p. 102). One of the sources that he used was an electronic version of SAOL 9, the Swedish Academy word list in its 9th edition (1950), with compound borders indicated. However, in later work he also looked solely at the forms of words and syllables without any lexical resource at hand (Brodda, 1981).

Good compound analysis is highly needed for unrestricted text, especially for languages whose orthographies concatenate compound components (that is, juxtapose the components without an intervening space). This means that that every such concatenation corresponds to a word. This way of forming words is extremely productive in most Germanic languages (including Swedish, but with the exception of English) and, for example, Finnish, Hungarian and Greek, and provides an important reason why an exhaustive list of words remains impossible to construct in these languages. Also, in languages like this an unknown word will most likely be a compound (Stymne and Holmqvist, 2008). On a related note, compounding seems to be an area where a lot of the creativeness of language is put to work (Svanlund, 2009; De Smedt, 2012).

So what new is there to say about Swedish compounding that could not be said one or a couple of decades ago? To begin with, there has been an enormous increase in the amount of electronically available data. At the Department of Linguistics, we have collected corpora from several Internet sources during the last years, including 2.7 billion tokens of Swedish blog text. There are two reasons why we find working with data from the Internet in general and blogs in particular highly useful. First, the sheer amount of data means that we obtain new ways of studying systematically various marginal and low-frequency phenomena that previously were more or less out of reach. One such example concerns neologisms and creative compounding in Swedish, which we can find by looking among words that are extremely low-frequency in spite of the large data set. Secondly, as has frequently been pointed out, text found on the web often has a colloquial and spontaneous character. The umbrella term here is *user-generated content*, that is, text published predominantly by non-professionals in media such as blogs, forums, reviews, social networks and wikis. User-generated content, including blogs, can thus provide an effective window into language change

and variation.[1]

Another reason for us to look anew at compounding is that there are now good lexical resources with permissive "copyleft" licensing[2] for Swedish, notably SALDO (Borin and Forsberg, 2009) which contains about 115 000 entries. Earlier attempts at Swedish morphology and compounding (Karlsson, 1992; Sjöbergh and Kann, 2004, 2006) have used SAOL (the Swedish Academy word list) as a lexical back-end. While SAOL is a rich source (its 13th edition from 2006 comprises about 125 000 entries), its use is restricted and it is not freely available in any machine-readable format.[3]

The basic approach is thus to use the SALDO lexicon and to test the compound algorithm on our set of blog data. More specifically, the goal of this paper is threefold:

- to provide a downloadable, "free software"[4] tool for splitting Swedish compounds;
- to test the accuracy of this tool on our blog corpus (which is downloadable as a citation corpus called the Swedish Blog Sentences corpus);
- to show some uses of compound splitting with respect to this corpus, including quantifying the use of creative compounding.

1.2 What is a Swedish compound?

According to SAG, the Swedish Academy Grammar (Teleman et al., 2000), a Swedish compound is a word form that consists of a first and a last component. These components can be root morphemes, derivations or compounds. The fact that a component can itself be a compound introduces recursiveness into word formation, potentially giving rise to very long compound words. Typically, the last component is the main one, grammatically and semantically, with respect to the whole compound.

The most common types in our data (and also the types mostly discussed in SAG, Volume 2) are compound nouns, adjectives and verbs. By a "compound noun", etc., we mean a compound whose last component is a noun, etc., but whose first component may be another part of speech.

For noun compounds, SAG (Volume 2, Nouns, § 25–27) mentions most other parts of speech as possible first components, namely, nouns (*fågel-bok*, bird book), proper nouns (*Zorn-tavla*, Zorn painting), adjectives (*låg-pris*, low price), verbs (*köp-stopp*, buying stop), participles (*vuxen-gymnasium*, adult high-school), count words (*vi-känsla*, we feeling), pronouns (*vi-känsla*, we feeling), numerals (*andre-pilot*, second pilot), prepositions (*med-vind*, fair wind), and adverbs (*fram-sida*, front side). All

[1] We have already made preliminary work on dialectal word variation in the blog corpus; see `http://www.ling.su.se/english/nlp/tools/dialect-maps`

[2] This includes the freedom to use, modify and re-distribute the work. All of these are important in a project such as ours, where one needs to adapt a lexicon to specific purposes and then distribute the final software package.

[3] As far as we know, SAOL is available only through individual licenses with the Swedish Academy. Furthermore, although it is accessible on the Internet (at `http://www.svenskaakademien.se/ordlista`), the text is only available as images of the print edition.

[4] *Free software* denotes software whose source code is available under a copyleft license. This is not to be confused with software available free of charge but without source code or with restrictions on its use. The software used in this study is available under the GNU GPL license: `http://www.gnu.org/copyleft/gpl.html`

of these cases are present in our data; however, it is difficult to see how populated some of the groups are because of ambiguity. For example, should we regard *köp* in *köp-stopp* (buying stop) as a verb (which SAG suggests) or a noun (which is also possible)?

For adjective compounds, SAG (Volume 2, Adjectives, § 33) mentions nouns (*barn-vänlig*, children friendly), proper nouns (*falu-röd*, Falu red), adjectives (*mörk-grön*, dark green), verbs (*prat-glad*, literally "speak happy", talkative), pronouns (*all-svensk*, pan-Swedish, relating to the Swedish football league), prepositions (*mellan-stor*, middle-sized), and adverbs (*bak-tung*, tail-heavy). Again, all of these cases are present in our data.

For verb compounds, SAG (Volume 2, Verbs, § 19) mentions nouns (*hunger-strejka*, hunger-strike), adjectives (*kal-hugga*, clear-cut, with respect to woods), verbs (*bränn-märka*, brand), prepositions (*till-foga*, append), and adverbs (*ill-vråla*, yell terrifically). Yet again, all of these cases are present in our data.

SAG also mentions compounds belonging to most other parts of speech. However, in quantifying the different types, we shall limit ourselves to the most frequent ones, namely, compound nouns, adjectives and verbs.

1.3 Legal issues

In any corpus project, one would ideally want the consent of all the intellectual property right holders for the texts in the corpus. Unfortunately in our case, the rights to the blog corpus we use is owned by some half a million individual writers, or about seven percent of the entire population of Sweden!

Since reproducibility is a cornerstone of all research, and in addition Swedish universities as branches of the government are mandated by the constitution to share data with anyone requesting it, it is important that we find some way to make our corpus publicly available.

For some uses, it is sufficient to provide a search interface, so that users can search for sentences but not download the entire data set. This avoids the copyright issues, but in general is not useful for research in computational linguistics, where one wants to automatically analyze *all* of the data, preferably using one's own programs and not being limited by a search interface. One example of this is SAOL, whose search interface allows you to look for particular words, but for which only single book pages are returned as bitmaps, not in a machine-readable format.[5] Another example is Litteraturbanken[6], which only allows you to look at (or access) a single page at a time. Of course, both SAOL and Litteraturbanken are circumscribed by copyright protections.

Some projects, such as the WaCky corpus (Baroni et al., 2009), choose an "opt-out" model, where copyright holders are not notified by the project about the existence of the corpus, but can request to be excluded from it if they find out. Although a corpus may contain copyrighted material, it consists of material already published on the Internet for which WaCky's automatic routines cannot detect that the material is copyright-protected. The potential damage of an inclusion into WaCky may therefore not be very high. Also, this model is a very practical way of providing easy access for researchers to a valuable resource. On the other hand, such an approach may require

[5] `http://www.svenskaakademien.se/ordlista`

[6] `http://litteraturbanken.se`

the project to either stay under the radar, or to have sufficient financial resources to crush any opposition in a legal process.

Alternatively, if the basic unit of processing is at the sentence level or below, one could distribute a "citation" version of the corpus where the sentences are randomly shuffled around, making it impossible to reconstruct the original texts. This is the way that for instance the Swedish Språkbanken material[7] and the Danish KorpusDK[8] are distributed.

For the purposes of this study, we do not depend on information above the sentence level. The most practical approach thus appears to be the "citation corpus" approach, which we have adopted.

2 Data

The current study is based on two major resources: a large corpus of Swedish blog posts, and the SALDO lexicon of Swedish morphology. In this section we discuss these resources.

2.1 Blog corpus

Our blog corpus consists of Swedish blog posts obtained from Twingly[9] from November 2010 until September of 2012, in total about 2.7 billion tokens, spread over 220 million posts from 660 000 different blogs.

The set of sentences from the blog corpus, on which this study was based, is available for download at the Department of Linguistics website.[10]

A smaller amount of blog text has been annotated and published as the Stockholm Internet Corpus (Östling, 2012), available from the department website under a permissive copyleft license.[11] Although its annotation quality is higher than the full, automatically processed corpus, the small size (currently about 8 000 words) prevents us from using it in our current study.

Due to the large quantity of text, all processing is done automatically. The material contains some undesirable elements that may be difficult for the computer to detect and filter out accurately, such as:

- duplication, for instance due to quotations
- automatically generated or translated text
- text in foreign languages
- passwords and other random letter sequences

All of these may result in artifacts. For instance, a misspelling in an article quoted by many blogs may show up as a relatively common variant, even though it was only actually written once.

[7] http://spraakbanken.gu.se/eng/node/1587/
[8] http://ordnet.dk/
[9] http://www.twingly.com/
[10] http://www.ling.su.se/sbs
[11] http://www.ling.su.se/sic

For all its problems we choose to use this corpus because the sheer amount of text, from a wide variety of different sources, puts us in a very good position to investigate uncommon compounds and compounding patterns in general usage.

2.2 Preprocessing

All text has been tokenized, and annotated with parts of speech (POS tagged) and citation forms (lemmatized), using the Stagger POS tagger (Östling, 2012). While elaborate algorithms have been developed for tokenizing Internet text, we found that a simple regular expression-based tokenizer provided satisfactory results after it was modified to handle some typical Internet phenomena, such as emoticons.

The automatic annotation process is not perfectly accurate. In a sample of (poorly written) blog posts, about 8% of tokens were analyzed incorrectly. In the more conventional written Swedish of the Stockholm–Umeå Corpus (Gustafson-Capková and Hartmann, 2008), this figure is close to 3%.

This preprocessing step provides two important pieces of information for the future analysis: whether a word belongs to the four parts of speech considered for compound analysis (noun, adjective, verb or adverb), and the citation form of the word, on which the compound analysis is based.

In section 3.2, we show that although errors in this step do propagate into the final compound analysis, the impact is relatively minor for noun and adjective compounds. Many non-words are however mistaken for verbs or adverbs during preprocessing, which makes it difficult for us to accurately analyze verb and adverb compounds.

2.3 Lexicon

We use the SALDO lexicon of Swedish morphology (Borin and Forsberg, 2009), containing 115 661 words[12] as the basis of our algorithm.

Compounding forms are divided into initial and middle forms, the latter typically end with the linking morpheme *-s-*. While the compounding form of a word is often simply the word's stem, there are exceptions, and using the SALDO lexicon makes it trivial to include these. The same applies to the sometimes idiosyncratic behavior of the linking morpheme.

The final segment of a compound is identical to the corresponding word in isolation, and also behaves like it in terms of inflectional morphology. Thus, we can simply use the inflectional forms in SALDO for final compound segments.

3 Methods

3.1 Compound analysis

Our basic approach to analyzing a compound is very simple:

[12]This is the number, at the time of writing, of *lemgrams* in the SALDO morphological lexicon. Each lemgram represents a word, an abbreviation or a multi-word expression, and contains its citation form as well as inflectional and compounding morphological forms.

For a given word of a given part of speech, find the shortest possible sequence of compound segments in the SALDO lexicon, which put together form the word in question with the correct part of speech.

For instance, given the word *kändisgala* (*celebrity gala*), which we know to be a noun from the preprocessing, a dictionary look-up gives us four initial hypotheses:

1. *kändis-gala* (noun, *celebrity gala*)
2. *kändis-gala* (verb, *to celebrity-crow*)
3. *känd-is-gala* (noun, *famous ice gala*)
4. *känd-is-gala* (verb, *to famous-ice-crow*)

Hypotheses (2) and (4) can be eliminated because the part of speech of the compound as a whole and its final segment do not agree, and hypothesis (3) can be eliminated since it contains three segments whereas (1) only contains two.

There are however several complications to this conceptually simple method, which we will now discuss.

Letter pairs Swedish orthography does not allow more than two identical letters in a row, which means that a letter pair may be split in three different ways. For instance, *glasskål* could be *glass-skål* (icecream bowl), *glas-skål* (glass bowl) or *glass-kål* (icecream cabbage). This special rule generates some more ambiguity, which however does not seem to be a problem in practice. One could also consider introducing a similar rule for s/ss ambiguity, since it is fairly common (although prescriptively frowned upon) to omit the *-s-* linking morpheme when it is followed by another *s*, as in *blåbär(s)soppa* (blueberry soup). Sequences that would consist of more than three letters are very rare (one example is *Råå-å-ål*, *eel from the Råå river*), and we do not consider these.

Words in SALDO The SALDO lexicon contains non-compounds, highly lexicalized compounds, as well as mostly compositional compounds, but does not provide any information about which category a given word belongs to. In order to avoid the difficult problem of how to demarcate the borders between these categories, we could simply treat all in-lexicon words as atomic and refuse to split them. Unfortunately, this would leave us unable to answer one important question: *how often are compounds used?*

For this reason, we do try to split words that occur in SALDO. Since a naive approach results in erroneously splitting many non-compounds, we use a simple semantic filter to extract plausible analyses. An analysis is only accepted if the compound shares an ancestor within two levels in SALDO:s semantic hierarchy, with the proposed segments. For instance, the compound *gastryck* is found in SALDO and has two analyses: *gas-tryck* (*gas pressure*) and the less probable *gast-ryck* (*ghost twitching*). The former analysis is accepted, since the compound is directly related to both *gas* and *tryck* in the semantic hierarchy. Because no such connection exists for the second analysis, it is rejected. If all analyses are rejected, we do not consider the word a compound.

No short segments Many non-compounds may be interpreted as a compound consisting of very short segments, such as the English *hate* which could be interpreted as *ha-te* (*having-tea*). For this reason, unless a hyphen is used, compound segments shorter than three letters are not allowed. This unfortunately means that we will not detect the (relatively few) such compounds that actually exist, and others may want to make a different decision.

No more than four segments Compounds with more than four segments are very rare in practice, Sjöbergh and Kann (2006) found only two cases among 3400 compounds manually checked. Like the previous case with very short segments, allowing very long compounds tends to introduce many errors where non-compounds receive some nonsensical segmentation.

Fewest-segments heuristic (Karlsson, 1992, p. 29) uses and defends this heuristic, but lists a few cases when it may give unintended results. While none of his examples are citation forms (which is what our algorithm analyses), one can construct such examples where using the heuristic leads to missing an actual compound. For example, *finska* (*Finnish language*) would always be preferred to *fin-ska* (*high-class ska music*), even though the latter may be intended in some cases.

3.2 Accuracy

To be able to say anything about the occurrence of compounds in our corpus, we first need to find out how accurate the algorithm is at identifying and analyzing compounds.

In order to investigate the accuracy of the compound splitting algorithm, we perform a manual inspection of the compound analyses delivered. There are two measures we are interested in: precision, the proportion of compounds identified by the algoritms that are actually Swedish compounds, and recall, the proportion of the total number of compounds in the data that were found by the algorithm.

In total, there were 2 253 395 noun-tagged unique words identified by the algorithm as compounds, and 2 483 827 as non-compounds of which 1 474 were in the SALDO lexicon. We now take a closer look at what the two categories contain.

3.2.1 Precision

Since the total number of compounds is very large, we randomly selected smaller samples of compounds and their analyses by the algorithm.

Generally speaking, rare words are more difficult to analyze, since they are more often misspellings or use non-standard morphology than frequently occurring words. While using frequent words in the evaluation would probably give better-looking accuracy figures, most of the compounds in our corpus are quite rare: the majority (63.3%) of noun compounds occur only once, and another 13.5% just twice.

Unique noun compounds Table 1 summarizes the result of an error analysis of 256 randomly selected nouns that occur only once each in the corpus, and which received at least one analysis as a compound by the algorithm. In 188 cases (73.4%), the system output only one analysis, which turned out to be correct. The second most common

Table 1: Unique noun-tagged words interpreted as compounds by the algorithm.

Count	*Percentage*	*Result*
188	73.4%	Unique correct analysis
5	2.0%	Ambiguous, but at least one correct analysis
14	5.5%	No correct analysis
40	15.6%	Non-compound incorrectly analyzed as compound
9	3.5%	Preprocessing error
256	**100.0%**	**Total**

Table 2: Unique adjective-tagged words interpreted as compounds by the algorithm.

Count	*Percentage*	*Result*
52	52.0%	Unique correct analysis
0	0.0%	Ambiguous, but at least one correct analysis
3	3.0%	No correct analysis
24	24.0%	Non-compound incorrectly analyzed as compound
21	21.0%	Preprocessing error
100	**100.0%**	**Total**

case (40, 15.6%) was that the word was not a Swedish compound at all (but e.g. a misspelling or a foreign word), although the algorithm managed to find some analysis for it. For instance, *trangeldrama* was interpreted as *tran-gel-drama* (whale-oil gel drama), whereas it is most likely a misspelling of *triangel-drama* (triangle drama).

Just 14 cases (5.5%) were acceptable Swedish compounds that did not receive any correct analysis. Only in these few cases, the compound splitting algorithm alone is responsible for failing to deliver a correct analysis.

In spite of the fact that the correct part of speech tagging and lemmatization of blog text is a difficult problem, particularly for words that only occur once and are typically not in the lexicons used by the preprocessing tool, only nine cases (3.5%) of the incorrectly split compounds were due to errors in these preprocessing steps.

It is possible for a compound to have several analyses, which may be more or less probable. For instance, in this sample we observed the compound *flyttemperatur*, which could be interpreted as either *flytt-temperatur* (moving temperature) or as *flyt-temperatur* (floating temperature). In total, five of the correctly analyzed compounds were ambiguous and received at least one correct analysis.

Although the 73.4% precision figure may seem low, the largest share of the errors come from non-compounds (frequently misspellings). If non-compounds and preprocessing errors are excluded, the algorithm finds a unique, correct analysis for 191 of 207 compounds (90.8%).

Unique adjective compounds Table 2 shows the corresponding precision figures for unique adjectives. The main difference from the nouns is that the number of prepro-

Table 3: Unique verb-tagged words interpreted as compounds by the algorithm.

Count	*Percentage*	*Result*
32	32.0%	Unique correct analysis
1	1.0%	Ambiguous, but at least one correct analysis
5	5.0%	No correct analysis
35	35.0%	Non-compound incorrectly analyzed as compound
27	27.0%	Preprocessing error
100	**100.0%**	**Total**

Table 4: Noun-tagged words interpreted as compounds by the algorithm.

Tokens	*Percentage*	*Result*
7 293 527	90.7%	Unique correct analysis
9 131	0.1%	Ambiguous, but at least one correct analysis
484	0.0%	No correct analysis
713 747	8.9%	Non-compound incorrectly analyzed as compound
27 384	0.3%	Preprocessing error
8 044 273	**100.0%**	**Total**

cessing errors is larger, at the expense of the correctly analyzed compounds. In most cases, this is due to the part of speech tagger mistaking nouns for adjectives. Excluding non-compounds and preprocessing errors, 52 of 55 (95%) of the proper adjective compounds in the sample receive a unique, correct analysis by the algorithm.

Unique verb compounds For verbs (table 3), the situation is even worse. 62% of the identified compounds are in fact non-words or mis-tagged words from other parts of speech, frequently nouns ending with the common verb infinitive suffix *-a*.

Unique adverb compounds Only 10% of unique words identifies as adverb compounds are in fact such, since actual adverb compounds seem to be rare, while many non-words and words of other parts of speech are mistaken for adverbs in the preprocessing.

Noun compounds overall We now turn from unique compounds to compounds as they occur in the corpus, with a heavy bias towards a few, common words (see figure 1). For a randomly selected word token in the corpus that the algorithm identifies as a compound, what is the rate of success?

Table 4 shows the result of a manual analysis of the algorithm's performance on 256 randomly chosen such words. Since the analysis is independent of the context in which a word is found, we can simply count a correct analysis of a word occurring 1000 times as that many correct instances, and similarly for incorrect analyses.

While unique compounds are nearly always compositional (for natural reasons), we now run into issues of lexicalization and language change. The word *frukost* (breakfast) makes up 22% of the total instances in this sample, so a great part of the accuracy figure depends on whether or not we accept the analysis *fru-kost*. Historically, this is indeed a compound meaning *early meal*, but in current use the prefix *fru/fro-* (*early*) is obsolete.

In table 4 we have counted the analyses *fru-kost* and *kär-lek* (*love*) as correct for etymological reasons. If one decides to count these two cases as incorrect, the percentage of unique correct analyses immediately drops to 61.9%, so these figures should be interpreted with great care.

The question of when a compound stops being a compound is a matter of definitions, but in practice this is of little consequence since very lexicalized compounds tend to be in the lexicon, and the main purpose of a compound analyzer is to find a way to interpret words *not* in the lexicon.

3.2.2 Recall

Out of 128 randomly selected unique words that were tagged as nouns and were neither in SALDO nor identified as compounds, only 18 (14.1%) were considered to have an analysis as a correctly spelled Swedish compound.

The remaining 85.9% were mainly random letter sequences (e.g. passwords and web addresses), misspellings and parts of sentences written without white space, that the part of speech tagger had (often erroneously) tagged as nouns.

Since about 14.1% of all unique nouns classified as non-compounds are in fact compounds, we see that about 82.5% of all unique noun compounds were correctly identified by the algorithm. Assuming that we really want to discard misspellings, this shows us that our algorithm does not miss very many noun compounds, even for the difficult to analyze unique words.

For adjectives, verbs and adverbs, we also found that few actual compounds are missed by the algorithm.

3.2.3 Discussion

Since very rare words in Swedish are typically compounds, a good compound analyzer would in theory be able to separate "real" words from misspellings, random letter sequences, foreign language words, and similar types of Web Noise.

Misspellings[13] are a fact of life, and of Internet life in particular. By inspecting a random sample of the unique noun-tagged words in the blog corpus, we found that about one fourth of them are misspellings, and quite often misspellings of compounds.

One way to improve the compound analysis algorithm would be to allow for spellings not in the lexicon. However, this has to be done with great care. It is already too easy for a non-word to be interpreted as a (weird) compound, if one also allows *misspelled* compounds, this risk increases further.

[13] Since our software is based on the SALDO lexicon, we define the spelling(s) listed there as "correct", and alternative spelling as misspellings. For words not in the lexicon, the subjective opinions of the authors are used. This is a practical decision to enable automatic analysis, and should not be interpreted as prescriptivist fundamentalism.

3.3 Related work

Sjöbergh and Kann (2004) studied algorithms for splitting Swedish compounds, and report that with their best method 99% of compounds were split, and of these 97% were analyzed correctly. This is substantially higher than our figures, but the difference can be explained in part by the following factors:

- They use the Stockholm–Umeå Corpus of published (and presumably proofread) professional prose (Gustafson-Capková and Hartmann, 2008), as opposed to the unedited blog texts by non-professional writers we used. In particular, this means that they did not have to deal with the large amount of non-words observed among the unique letter sequences in the blog corpus. The manual annotation of the Stockholm–Umeå Corpus also means that they did not suffer to the same extent from the preprocessing errors we describe.

- They split word tokens in context, which means that the compound *types* evaluated are biased towards common ones, according to the Zipfian distribution of compounds such as the one shown in figure 1. We draw our evaluation sample from unique or rare word tokens, and as we showed in section 3.2.1, more common words are in general easier to analyze.

As we mentioned towards the end of section 3.2.1, excluding non-compounds and preprocessing errors we also arrive at a figure of 97% precision. This figure is also not directly comparable to that of Sjöbergh and Kann (2004), among other things due to different standards in judging the correctness of analyses, where they are more strict. In short, while a direct comparison with previous work is difficult, it seems that the accuracy of our method is fair.

4 Results

Having established that our methods can be trusted in the majority of cases, at least for some parts of speech, we now turn to look at how the overall distributions of compounds look. Later, we will also look at how compounding patterns can be used to extract semantic information.

One should keep in mind that the blog corpus contains much noise, which may be misidentified as compounds. This is a particularly large problem for adverb compounds, since actual compounds are quite uncommon and easily disappear in the noise. For this reason we do not consider adverb compounds at all, only briefly discuss adjective and verb compounds, and focus on noun compounds which are the most frequent and where our system produces the most accurate results.

4.1 How common are compounds?

We want to know how common compounds are in the blog corpus. First, we will have a look at compounds that are *not* in the SALDO lexicon, and which are typically quite rare and compositional. Then, we will look at compounds in the SALDO lexicon, which are more common but often highly lexicalized.

Table 5: Frequencies in the blog corpus of words in SALDO and compounds not in SALDO.

	Words in SALDO		**Compounds not in SALDO**	
	Types	*Tokens*	*Types*	*Tokens*
Noun	63 197	334 731 534	3 563 928	30 176 761
Adjective	13 394	111 174 488	206 715	3 429 193
Verb	7 075	426 829 335	187 898	1 082 911

Table 6: Frequencies in the blog corpus of non-compounds in SALDO and compounds.

	Non-compounds in SALDO		**Compounds**	
	Types	*Tokens*	*Types*	*Tokens*
Noun	29 154	295 235 054	3 593 728	67 730 894
Adjective	10 485	106 970 680	209 967	7 581 560
Verb	5 986	422 812 258	189 799	5 104 810

Recall from section 3.1 that SALDO does not contain information about which words are compounds, so we are forced to use an imperfect method to attempt to deduce this.

Compounds not in SALDO Table 5 shows the number of different words (*types*) and the number of instances (*tokens*) in the blog corpus. Perhaps the most striking fact is that on the one hand, the vast majority of words in a text are found in the lexicon, but on the other hand, the vast majority of word *types* are not.

7.6% of noun-tagged tokens are interpreted as compounds by our algorithm, compared to 2.6% of adjective-tagged tokens and 0.24% of the verb-tagged.

Figure 1 shows the frequency of words as a function of their rank, that is, how many times we find the most common word, the second most common, the third most common, and so on. Although the total number of compounds varies between different parts of speech, the compound rank/frequency plots all follow *Zipf's law*, according to which the frequency of the n:th word is roughly proportional to $1/n$.

Compounds in SALDO In table 6, we show how these figures change if one also counts compounds in the SALDO lexicon. As expected, the number of compound types and tokens increases. 17.0% of noun-tagged tokens are now identified as compounds, compared to 5.6% of adjective-tagged tokens and 0.46% of verb-tagged tokens.

Comparing table 6 to the number of unique noun compounds (see section 3.2), we find that about 0.6% of all nouns in the blog corpus are *unique* compounds, which demonstrates the level of creativity present in Swedish noun compounding.

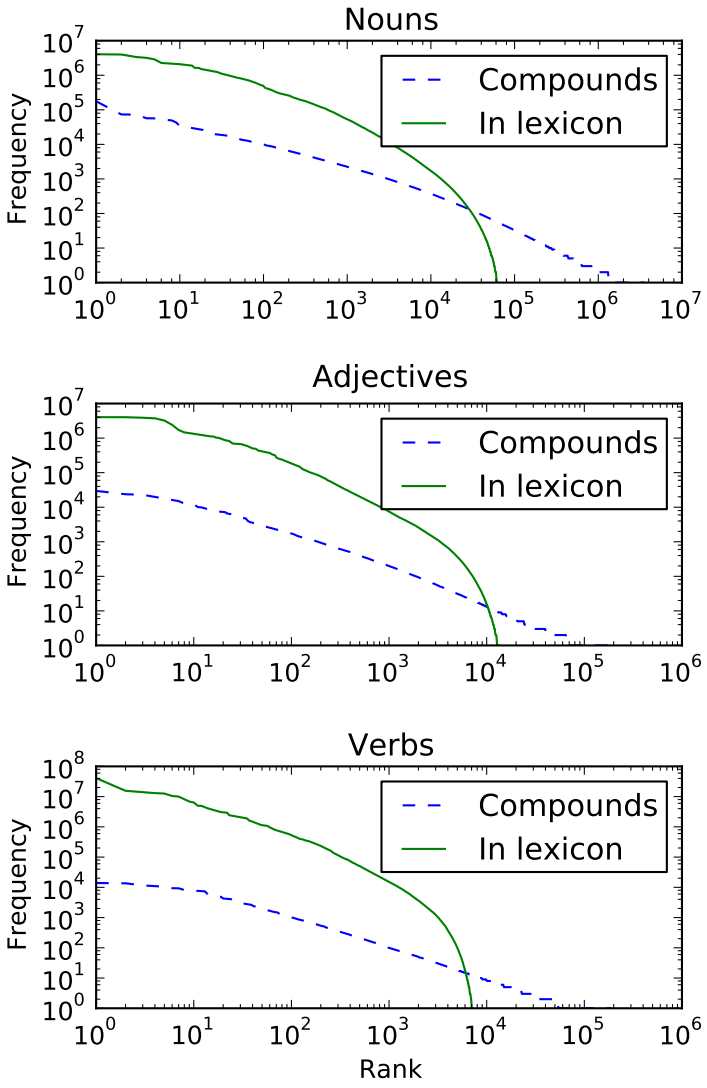

Figure 1: Rank/frequency distribution of different parts of speech.

4.2 Mapping attitudes through compounds

In order to demonstrate what can be done by computer analysis of our large database of compounds, we will try to create a map of attitudes towards different activities, based on a set of compound constructions with different connotations, denoting different kinds of enthusiasts.

This is in instance of *distributional semantics*, which exploits the correlation between the meaning of a word and its distribution in a language. Often one uses as wide a variety of contexts as possible in order to explore all aspects of a word's meaning, but here we look at compounds with similar denotations, and try to find the connotations associated with each prefix.

We consider five specific compound suffixes: *-bög*, *-fantast*, *-freak*, *-nörd* and *-tönt*. These can all be used to denote an enthusiast of a particular thing or activity, but their connotations are different. *-bög* (a homosexual) conveys a negative judgment of the activity in question, often as being snobbish. *-fantast* (fan) is fairly neutral, while *-freak*, *-nörd* (nerd) and *-tönt* (dork) have the particular (mostly negative) connotations of their English counterparts.

Lexicalized compounds and polysemous suffixes could in theory cause trouble, by breaking our basic assumption that *X-suffix* means exactly "an enthusiast of X" and that the choice of suffix reflects only which attitude towards the activity the writer intends to express. It turns out that most prefixes tend to occur with several different suffixes, and only in exceptional cases are they associated with a single suffix, which is some indication that this is not a problem in practice.

We collect all prefixes that occur at least 25 times in total with the suffixes mentioned, in total 279 prefixes. Since the suffixes are not equally frequent, we normalized the distribution of compound frequencies within each suffix. Next, we normalized the frequencies of each *prefix*, in order to obtain a measure of how the suffixes would be distributed for each prefix, if suffixes were equally frequent.

Finally, we used Principal Component Analysis[14] to reduce the 5-dimensional vectors of each prefix into two dimensions, which are shown for a selection of the most common prefixes in figure 2.

At the bottom, we see typical nerdy activities, such as *data-* (computer) and *språk-* (language). Towards the upper left, there are prefixes associated with *-bög* and *-tönt*, such as *pryl-* (gadget) and *iphone-*. In the upper right, we have prefixes associated with *-fantast*, such as *skräck-* (horror fiction) and *deckar-* (crime fiction). Finally, prefixes that show no clear preferences towards any suffix are grouped near the middle of the figure, for instance *musik-* (music) and *teknik-* (technology).

5 Summary

We have used the largest corpus that, to the best of our knowledge, has ever been used for analyzing Swedish compounds. Through this, we have been able to quantify the use of different types of compounds used on Swedish blogs. Among other things,

[14] We used the implementation of the Modular toolkit for Data Processing (MDP) project. Please refer to the documentation of this software for further information:
`http://pypi.python.org/pypi/MDP`

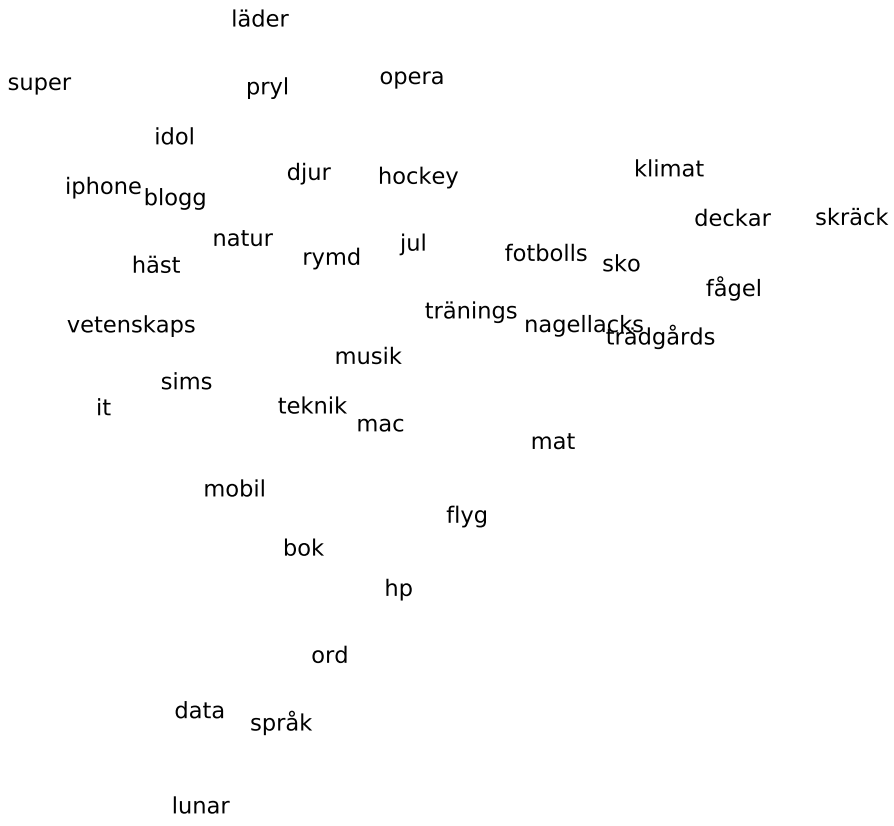

Figure 2: Attitude map based on the suffixes *-bög*, *-fantast*, *-freak*, *-nörd*, *-tönt*.

we have shown that nearly one in every thousand words is a unique noun compound, clearly demonstrating the level of creativity in Swedish compounding.

Furthermore, we have applied our large database of Swedish compounds to create a map over the overall attitude towards different activities, based on the connotations of the words they form compounds with. This is merely a small demonstration of what kind of information can be extracted from compounding patterns. We hope that future studies studies will be able to use our data and methods to extract more of the information present in Swedish compound constructions.

The program used for analyzing compounds in this work is Free Software, and can be downloaded as part of the SPyRo package from the website of the Department of Linguistics[15].

[15] `http://www.ling.su.se/english/nlp/tools/spyro`

References

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources & Evaluation*, 43:209–226.

Borin, L. and Forsberg, M. (2009). All in the family: A comparison of SALDO and WordNet. In *Proceedings of the Nodalida 2009 Workshop on WordNets and other Lexical Semantic Resources – between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*, Odense.

Brodda, B. (1981). Yttre kriterier för igenkänning av sammansättningar. In *Förhandlingar vid trettonde sammankomsten för att dryfta frågor rörande svenskans beskrivning*, pages 102–114. Meddelanden från Institutionen för nordiska språk och nordisk litteratur vid Helsingfors universitet.

De Smedt, K. (2012). Ash compound frenzy: A case study in the norwegian newspaper corpus. In Andersen, G., editor, *Exploring Newspaper Language. Using the web to create and investigate a large corpus of modern Norwegian*. John Benjamins, Amsterdam.

Gustafson-Capková, S. and Hartmann, B. (2008). *Manual of the Stockholm Umeå Corpus version 2.0*. Stockholm University.

Karlsson, F. (1992). Swetwol: A comprehensive morphological analyser for swedish. *Nordic Journal of Linguistics*, 15:1–45.

Östling, R. (2012). Stagger: A modern POS tagger for Swedish. In *Proceedings of the Swedish Language Technology Conference (SLTC)*.

Sjöbergh, J. and Kann, V. (2004). Finding the correct interpretation of swedish compounds, a statistical approach. In *In Proc. 4th Int. Conf. Language Resources and Evaluation (LREC)*, pages 899–902.

Sjöbergh, J. and Kann, V. (2006). Vad kan statistik avslöja om svenska sammansättningar? *Språk och stil*, 1:199–214.

Stymne, S. and Holmqvist, M. (2008). Processing of swedish compounds for phrase-based statistical machine translation. In *Proceedings of the 12th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 180–189, Hamburg, Germany. European Association for Machine Translation.

Svanlund, J. (2009). *Lexikal etablering : En korpusundersökning av hur nya sammansättningar konventionaliseras och får sin betydelse*. Number 52 in Stockholm studies in Scandinavian philology. 1 edition.

Teleman, U., Hellberg, S., and Andersson, E., editors (2000). *Svenska Akademiens Grammatik*, volume 1–4. MediaPrint, Uddevalla, Sweden.