SAOL Plus—A New Swedish Electronic Dictionary

Sture Berg
Louise Holmer
Anki Hult
University of Gothenburg

In September 2007, a CD version of the Swedish Academy Glossary, SAOL Plus, was released. In SAOL Plus, all inflected forms are shown in full text and virtually every text fragment is searchable. Standard search functions include Search lemma, Search inflected forms and Search article text. Advanced searches can be made with the usual wild cards. It also has an advanced tool for fuzzy search based on pronunciation.

SAOL Plus can be an asset for the public as well as an efficient and functional utensil for linguists. Moreover, thanks to the fuzzy search, it is useful for people with reading and writing disorders, as well as secondary language users.

System requirements: Windows 98/NT/2000/XP/Vista.

1. Introduction

SAOL Plus, a CD version of the Swedish Academy Glossary (*Svenska Akademiens ordlista*, in short SAOL), was released in September, 2007. It contains inflected forms in greater number and detail than SAOL and has several advanced search options. It also has a tool for fuzzy search based on pronunciation (Oribi 2008). Intended at being useful for people with reading and writing disorders, as well as secondary language users, the user group is extended.

In this paper we describe SAOL, the underlying database SMDB and the construction of *SAOL Plus*. We focus on how to perform common and advanced searches and how these can be of interest for the public as well as for linguists.

2. The Swedish Academy Glossary (SAOL)

SAOL was first published in 1874. New editions appear approximately every tenth year, and the 13th edition was published in April, 2006, under the direction of Dr. Martin Gellerstam. SAOL comprises about 125,000 lemmas, which is about twice as many as other Swedish monolingual dictionaries; however, it provides fewer definitions. About one fifth of the lemmas are defined, commented on or syntactically exemplified (Gellerstam 2002). SAOL is widely used as a public reference work with regard to orthography and inflection, and to some extent also the pronunciation of Swedish. Approximately 10,000 new words have been added in the 13th edition and 5,000 obsolete words have been excluded in comparison with the 12th edition (1998).

3. How *SAOL Plus* was created

SAOL Plus is based on the Swedish morphological database (SMDB), which was developed at the Department of Swedish, University of Gothenburg (Språkdata). SMDB contains all the SAOL lemmas and their inflected forms, morpho-syntactically tagged (Berg & Cederholm 2001). Swedish has a rich and varied set of inflectional patterns. All lemmas in SMDB are divided into groups, with definitions of their inflected forms. On the basis of these groups, the inflectional paradigms of all lemmas can be generated from SMDB. *SAOL Plus* uses the forms generated, presented in full text together with grammatical information. This differs from SAOL, where only the most basic forms are included, in abbreviated form.

SAOL Plus provides the relevant part of the initially over-generated SMDB. While working on SAOL Plus, it became obvious that the contents of the database needed to be scrutinized.

The initial classification, consisting of 450 groups, had to be refined; some generated inflections were not semantically plausible or generally accepted. Nouns, verbs and adjectives were examined in depth—nouns as to whether they have plural and/or definite form; verbs as to whether they allow past participle and/or passive forms; and adjectives as to whether they take comparison and whether they may take the specific masculine ending in the definite singular form. For example, the non-comparable adjective *gravid* ("pregnant") originally belonged to a group also including adjectives that may undergo comparison, such as *rädd* ("afraid"). In total, the necessary modifications increased the number of groups to 850. The linguistic work consisted of extracting frequencies of inflectional forms using a tag program connected to a 200 million word corpus. These frequencies then provided a basis of our judgements when manually examining about one million inflectional forms. We especially focused on low frequencies as these indicated that the specific word form would be obsolete, rare or semantically implausible.

4. Search possibilities in *SAOL Plus*

In *SAOL Plus*, virtually every text fragment is searchable. Standard search functions include: *Search lemma* (in Swedish *Sök i uppslagsord*), *Search inflected forms* (*Sök i böjningsformer*) and *Search article text* (*Sök i artikeltext*). Advanced searches can be made with the usual wild cards ("*" and "?"). Search for specific affixes can be made using the vertical line ("|"). *SAOL Plus* also features *Crossword assistance* (*Korsordshjälp*, not further dealt with here).

4.1. *Search lemma*

The standard option is to search in all parts of speech, but the search can also be limited to a certain word class. Figure 1 shows the search result for the noun *dans* ("dance"). In the left field, the list of lemmas is shown, where the new words in the 13th edition are marked with a plus. In the middle field, the article text is shown in the way it is presented in the printed version. In the right field, one of the new features of the CD is shown; all the inflected forms of the lemma in full text, together with grammatical information. For nouns, as is the case in Figure 1, information about indefinite and definite base and genitive forms is given, first in the singular and then in the plural.

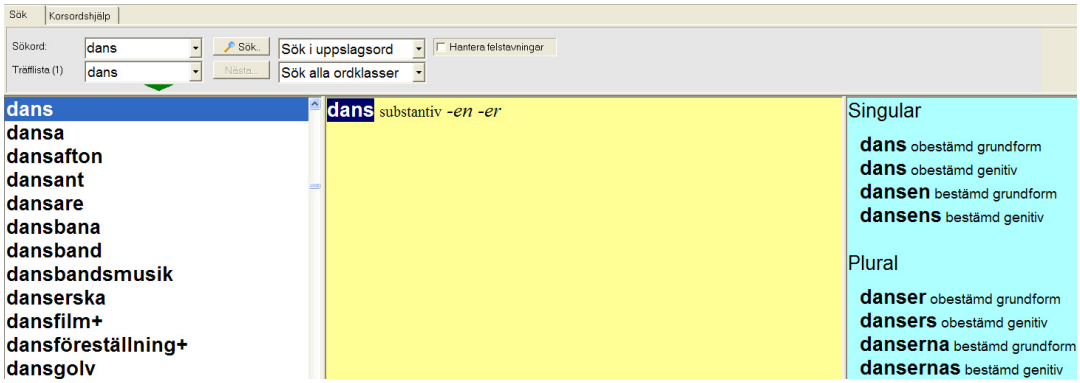

Figure 1. A screenshot from *Search lemma* with *dans* ("dance").

Wild cards enable searches to be extended. This is particularly useful for the large number of fixed compounds in Swedish. One can easily search for every word that ends with, for example, *-dans,* by typing an asterisk followed by "dans", i.e. **dans*, and get all the relevant words presented in the middle field. This kind of search is almost impossible to perform in a paper version, unless the compounds are already known.

4.2. *Search inflected forms*

In the printed version of SAOL, only a few inflectional forms are given for the base form of each lemma. Fixed compounds, which are very common in Swedish, provide no information about inflection or conjugation; the reader is referred to the base form of the compound for such information. For example, when looking up the verb *avgå* ("resign") the reader is referred to *gå* ("go" or "walk"). Often, however, there are semantic differences between simple words and

compounds, and consequently the inflectional patterns may differ to a greater or lesser extent (as in the case of *avgå*). In *SAOL Plus*, all inflectional and conjugational forms, also for fixed compounds, are shown in full text. By choosing the function *Search inflected forms*, they are also searchable.

Swedish has a wide variety of inflected forms and many of them are homographs. In a large group of nouns, for example, the indefinite plural has the same inflectional form as the present tense of many verbs (*-ar*). The word form *bottnar* is both indefinite plural of the noun *botten* ("bottom", "ground") and present tense of the verb *bottna* ("touch the bottom"). The result of this kind of search is shown in Figure 2 and 3. Lemmas affected by homography are listed in the middle field. The inflected (searched) form for each lemma is marked in the right field.

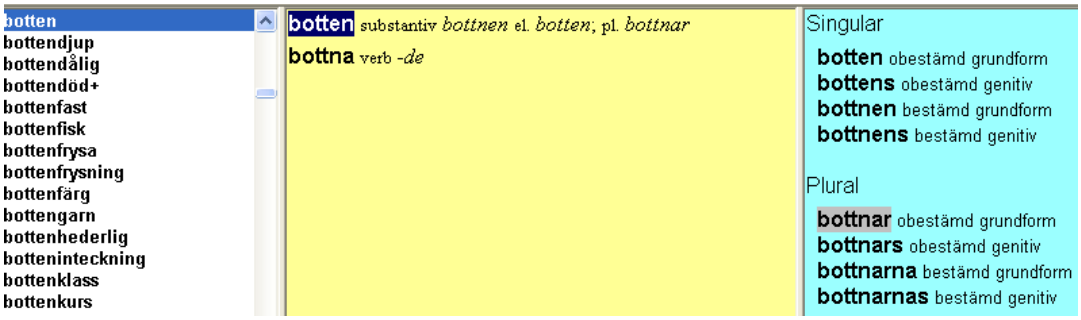

Figure 2. The result of the search *bottnar* in *Search inflected forms*.
The word form *bottnar* is marked in the right field. In this figure, it is the plural form of the noun *botten*.

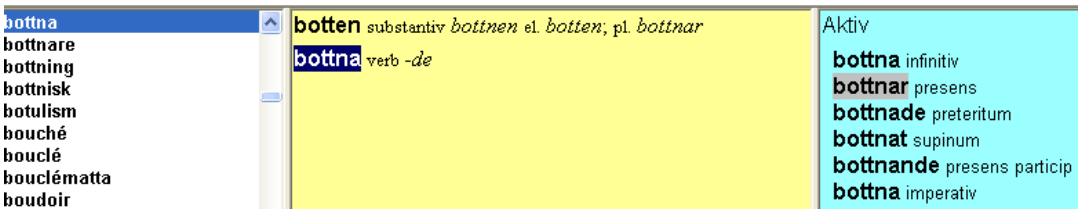

Figure 3. The result of the search *bottnar* in *Search inflected forms*.
The word form *bottnar* is marked in the right field. In this figure it is the present tense form of the verb *bottna*.

4.3. *Search article text*

Although SAOL is primarily a word list, many articles have a great deal of information besides inflection and pronunciation. This information is searchable in *SAOL Plus* thanks to the option *Search article text*.

Subject fields, regional variants and stylistic level are examples of comments that some lemmas may have. For example, if one is interested to know what lemmas are related to sports (*sport*), or what lemmas are often used in a humorous way, or what words are considered old-fashioned (*skämts*., *åld.*), these can be extracted through these comments on register and usage. In addition, since new words are marked with a plus in the left field, it is possible to see what new words have been added to a certain subject field. The result of such a search is presented in the middle field, where the search string is marked in a different colour.

Another search option worth mentioning is to search for a certain paradigm, since this information is included in the article text. A typical inflectional paradigm among nouns is *-en -er* (*elev – eleven – elever*, "pupil" – "the pupil" – "pupils"). It is also possible to extract the more unusual noun paradigm *pl. -s* (plural ending *-s*, typical of English loan words) and get all words inflected this way listed in the middle field. This kind of search is shown in Figure 4.

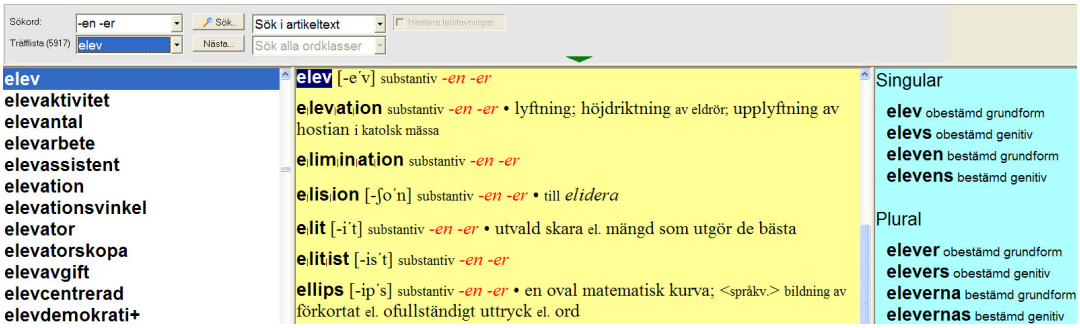

Figure 4. The result of the search for the noun paradigm *–en –er* in *Search article text*.

Moreover, comments about grammatical information can be searched for, such as verbs mostly used in the past participle form (*mest i perf. part.*, "mostly used in past. part.") or which two-word expressions can also be written as one word (*hopskr. äv.*, "also one word"), etc.

Finally, *Search article text* can provide valuable information about frequency, e. g. how frequent a certain inflectional paradigm is, or how many nouns have the English plural *-s*. Frequencies are shown in brackets next to the list of hits (*Träfflista*). Since SAOL reflects the contemporary Swedish lexicon, *SAOL Plus* can be used as a tool to obtain a representative lemma selection for linguistic studies. In this way, *Search article text* can be of great use and interest for linguists.

5. The tool *Fuzzy Search*

One of the most interesting tools of *SAOL Plus* is *Fuzzy search* (*Hantera felstavningar*), which handles misspellings. This tool helps the user get to the target word, even if he or she doesn't know how it is spelled. The program presents a list of suggestions of words that the user might have meant. *Fuzzy search* can be used not only together with *Search lemma* but also with *Search inflected forms*, which is very useful since some inflected forms differ a great deal from the base form of the lemma. The module for handling misspellings has been specially developed for all lemmas in *SAOL Plus* by Oribi Ltd (see further Oribi 2008). It is based on both pronunciation and transposition of letters. Hence, it is very useful both for those who misspell words due to lack of knowledge or negligence and for users with reading and writing disorders.

Figure 5 shows the result when the definite plural form *läxorna* ("the homework") is misspelled as **lexerna*. This kind of misspelling could be made for several reasons. First, the short sound [ɛ] can be spelled with one of the alternants <e> or <ä>, but, they are usually pronounced in the same way. Second, the ending *-orna* is usually pronounced with an initial sound [ɛ] in colloquial speech. In the middle field, six suggestions of lemmas are presented, of which the top suggestion is the lemma searched for.

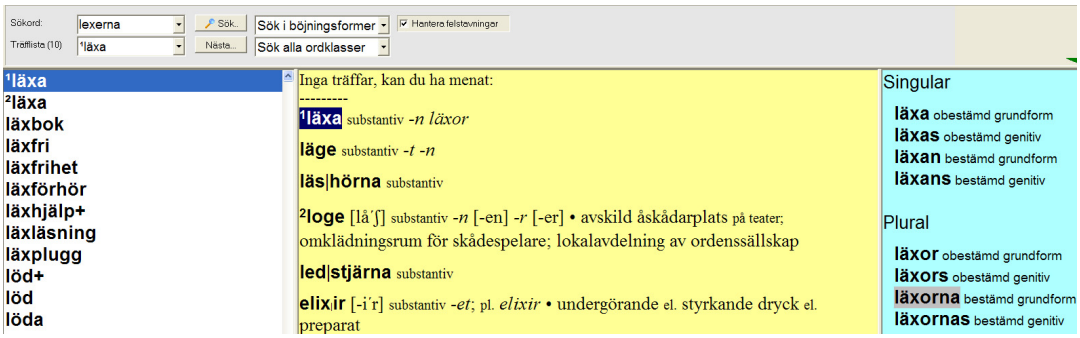

Figure 5. The result of the search *lexerna* (correctly spelled *läxorna*) in *Search inflected forms* together with *Fuzzy search*. In the middle field the suggestions of words are presented, and in the right field the inflected form is marked.

In Swedish, words pronounced with the phoneme [ʃ] are hard to spell for many people, since this phoneme can be spelled in many ways. The spelling of a word such as *journalist*

("journalist"), which is pronounced [ʃʊɳaliṣt] in Swedish, can therefore cause a lot of trouble. In *SAOL Plus*, this word can be spelled *schonnalisst* or *cshonalist*, and one will still get *journalist* as the first hit in the suggestion list, as shown in Figure 6. Since the *Fuzzy search* can be combined with *Search inflected forms*, even a drastically misspelled word like *schonalistenas* (plural definite genitive, should be spelled *journalisternas*) will yield a successful result.

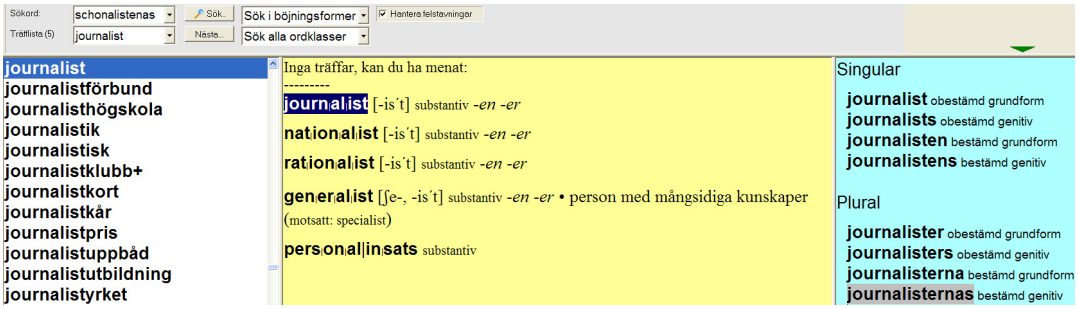

Figure 6. The result of the search *schonalistenas* (correctly spelled *journalisternas*) in *Search inflected forms* together with *Fuzzy search*.

Fuzzy search also works well with loan words like Eng. *mainstream*, which has kept both its pronunciation and spelling in Swedish. The *Fuzzy search* can easily be used as a spell-checker together with a word processor like Microsoft Word (see also Fontenelle 2006).

To the best of our knowledge, this *Fuzzy search* option based on pronunciation, has few rivals in its field.

6. Final comments

In this paper we have shown the facilities of the CD *SAOL Plus*, the electronic version of the Swedish Academy Glossary. Our examples show how the CD can be an asset for the public as well as an efficient and useful tool for linguists. First, *SAOL Plus* provides all inflected forms in full text together with grammatical information. The inflected forms are searchable, i.e. the user can get to any lemma by searching each of its inflected forms. Second, all kinds of information in the article text can be extracted and listed, such as frequency of inflectional paradigms, style comments, subject fields, etc. Third, with the help of *Fuzzy search,* specially developed for *SAOL Plus*, the potential group of users is widely extended.

References

Berg, S.; Cederholm Y. (2001). “Att hålla på formerna. Om framväxten av Svensk morfologisk databas.” In *Gäller stam, suffix och ord. Festskrift till Martin Gellerstam den 15 oktober 2001*. Göteborg: Meijerbergs arkiv för svensk ordforskning 29. 58-69.

Fontenelle, T. (2006). “Developing a Lexicon for a New French Spell-checker.” In *Proceedings XII Euralex International Congress. Torino, Italia, September 6th-9th, 2006.* Turin: Alessandria. 151-158.

Gellerstam, M. (2002). “Norm och bruk i SAOL”. I *LexicoNordica* 9. 21-30.

Oribi [on line]. *http://www.oribi.se/Eng/index_eng.htm*. Lund: Oribi Ltd. [Access date: 25 March 2008].

SAOL Plus [cd-rom]. Stockholm: Norstedts Akademiska Förlag, 2007.

Svenska Akademiens ordlista. [12] Stockholm: Norstedts Ordbok, 1998.

Svenska Akademiens ordlista. [13] Stockholm: Norstedts Akademiska Förlag, 2006.