Improving Dictionaries by Measuring Atypical Relative Word-form Frequencies

Kristian Blensenius, Monica von Martens

University of Gothenburg, Dept. of Swedish, Lundgrensgatan 1B, Gothenburg (Sweden)
E-mail: kristian.blensenius@gu.se, monica.von.martens@gu.se

Abstract

In this article, we discuss and give examples of how word-form frequency information derived from existing corpora statistics can be used to improve dictionary content. The frequency information is used in combination with rule-based morphological data based on derivational and inflectional information from the Swedish Morphological Database compiled at the University of Gothenburg, and the lexical database owned by the Swedish Academy. The method currently used in the ongoing project for updating the monolingual Contemporary Dictionary of the Swedish Academy is described, and some examples of dictionary entries identified as candidates for update based on frequency measures are given. Different aspects of morphological dictionary content are discussed and highlighted by comparison between the above-mentioned definition dictionary and a learner's dictionary. The role of headword or lemma as well as cross-referencing methods in a digital dictionary as compared to a printed dictionary is also discussed. Finally, a few examples of suggested modifications and enhancements are given.

Keywords: morphology; frequency; word forms

1. Introduction

In this article, we discuss how dictionary content can be improved by (re)using morphological information and enhancing it with corpus frequency information. Two contemporary Swedish monolingual dictionaries are used to illustrate how this method can be used to enhance dictionary content.

Morphological frequency matters have been much discussed from various perspectives, for example from the point of view of learning and producing word forms (e.g. Bybee, 1995; Hay, 2001; Dąbrowska, 2008). Hay (2001) specifically targets the question of absolute vs. relative word-form frequency, and particularly the relation between the so-called base form and a derived word form, showing that relative frequency seems to be even more important than absolute frequency when it comes to morphological decomposition.

In lexicography, morphology is instead traditionally often discussed in terms of the amount, compactness, and type of inflectional and derivational information to be presented in dictionaries (e.g. Heuberger, 2018; Svensén, 2009: 124ff.), reflecting the fact that presenting morphological information often constitutes a central component in dictionary entries. Inflectional information serves receptive functions, e.g. for finding

the lemma form and, importantly, productive functions, e.g. for finding inflected forms based on a lemma form. Finally, the morphological form of the headword is a topic discussed in the lexicographic literature, because it is often important to distinguish between the "base" form and other forms of a lemma. For example, some plural nouns like *arms* 'weapons' have a distinct meaning and may need to be presented as distinct headwords from their singular counterparts (see Atkins & Rundell, 2008: 325).

Whereas frequency-based lemma selection has indeed been discussed (e.g. Trap-Jensen et al., 2012), the distribution of individual *word forms* (inflectional and derivational forms) in terms of their relative frequencies has not been much considered in the lexicographic field. For that reason, our aim is, as mentioned above, to fill this gap by presenting a study of word forms in a morphological database, which we will evaluate with respect to two dictionaries.

2. The dictionaries

We limit our study to two major dictionaries of contemporary Swedish, aimed at two different user categories.

The first one is the monolingual 'Contemporary Dictionary of the Swedish Academy' (*Svensk ordbok utgiven av Svenska Akademien*, abbreviated SO), which is a definition dictionary primarily aimed at native speakers and advanced learners of Swedish. It is mainly a reception dictionary, but it is also production-oriented (Sköldberg, 2017: 123). SO is an edited extract of a much larger database compiled at the University of Gothenburg. The printed version of SO was published in 2009, the corresponding mobile app in 2015 and, finally, a freely available on-line web version was released in 2017. The tools used in this study have been developed as part of the revision process aiming at publishing a new up-to-date online version of SO.

The second dictionary used in our study is the present on-line version of Lexin, which is primarily a learner's dictionary (see the Lexin introduction). Lexin consists of a monolingual Swedish core, compiled at the University of Gothenburg on behalf of the Language Council of Sweden, which is translated into a number of immigrant languages. Older versions of Lexin have been published as both monolingual and bilingual printed dictionaries.

2.1 Morphology and headword policy in the dictionaries

Both SO and Lexin provide morphological information next to the headword. For example, the verb *köpa* 'buy' is paired with the following inflectional information:

SO: **köpa** *köpte köpt*, pres. *köper*

Lexin: **köper** att köpa, köpte, har köpt, är köpt, köp!

In SO, the inflectional paradigm of *köpa* is represented by the infinitival headword, the preterite and supine forms, and finally the present form of the verb. In addition, SO provides derivational information further down in the dictionary entry in the shape of two nominalizations, *köpande* 'buying' and *köp* 'purchase'. The learner's dictionary Lexin presents the headword, and then comes the infinitive (preceded by the infinitive marker *att*), the preterite, the supine (preceded by the perfect auxiliary in the present tense, *har*), the perfect participle (preceded by the passive auxiliary in the present tense, *är*), and finally the imperative form followed by an exclamation mark.

Svensén (2009) presents a list of "the grammatical forms most used as lemma forms", remarking that nouns are presented in the nominative singular form, verbs in the active infinitive, etc., provided that the structure of the language allows it. In the light of this, it could be noted that Lexin stands out in presenting the present form of verbs as the headword, in this case *köper*. This is not motivated by frequency[1] but by the assumption that the present form is the better basis for deriving the other forms of the verb (Gellerstam, 1999: 7f.). While on the subject, it could be pointed out that dictionaries for many other languages (e.g. Arabic and, as noted by Svensén, Latin) do not necessarily use the infinitive form as headword. Also, although the infinite has been the conventional lemma form for verb entries in Swedish dictionaries for approximately two hundred years, the present tense was commonly used in older dictionaries (e.g. Spegel, 1712, and Schenberg, 1739; see Hannesdóttir 1998: 148, 202). In older dictionaries in general, the choice of headword sometimes looks quite arbitrary to modern eyes for other word classes, too. Adjectives, for example, which take the suffix *-t* in the neuter singular form in Swedish, are frequently presented in this headword form in older dictionaries (e.g. Schenberg, 1739).

Dictionary entries in SO and Lexin commonly include special cross-referential headword forms, such as irregular verb forms like *gick* 'went' pointing to the base form *gå* 'go' (or the present *går* 'goes' in Lexin). Being electronic, both dictionaries should handle headword identification (Lew, 2012) automatically in these cases, either as redirections or links. This is the case for Lexin and the app version of SO, but regrettably these referential lemmas have at the time of writing gone missing in the web version of SO. This can be taken as a reminder of the fact that digitalization has not only upsides but also downsides; even a thorough proof-reading and testing phase on one or a few platforms cannot guarantee full functionality on all existing and upcoming environments, and it is seldom in the hands of the editors to decide about, and stipulate conditions for, the availability of the dictionary on new devices.

Another instance of morphological consideration affecting the choice of headword form for the dictionary entry is cases where the canonical lemma form is hardly ever, or

[1] Our frequency investigations also show that the infinitive and the present-tense forms are almost equally frequent for most verbs in our corpora, so frequency considerations can hardly be called on to favour one form over another as headword.

never, used. An example in English is the plural lemma form of nouns such as *scissors* (cf. Svensén, 2009: 105f.). A counterpart in SO is the active preterite auxiliary verb *torde* 'is probably, should' used as headword (this form is not included in Lexin).

Luckily, in digital dictionaries the choice of headword form for a lexical entry is typically not an either/or choice. As mentioned, by means of clickable links or redirection the user can often reach the desired entry regardless of which word form is entered in the search box. Still, one has to take caution not to give the user the impression he or she made some kind of mistake causing redirection. It has been reported by second language teachers that Lexin users sometimes believe a redirection was caused by misspelling, when the redirection was in fact caused by a void in the dictionary. Also, behind the scenes, in the database, it is strongly advised to attach inflection information in a standardized manner to a standard base form even if that form is not the one used to head the entry as shown to the public.

3. Problems

In Section 2 above we reviewed some cases of well-known morphologically induced problems a lexicographer needs to address, such as words for which the expected base form of a lemma is out of use, or almost out of use, and the case of verbs with irregular inflection which creates a need for several "entry points". In this section we address a couple of more intricate problems, for example how to deal with cases where a "base" headword form is indeed used but another word form is much more frequent and may have a slightly different meaning.

3.1 Word forms with a slightly different meaning than the base form

Sometimes the frequency distribution differs between word senses. Looking at, say, the plural form *blommor* 'flowers', we find that this form is much more frequent in several corpora than the singular form *blomma* 'flower', which is the headword form in SO and Lexin. Now, this does not necessarily mean that the plural *blommor* should be considered as a headword, not even for cross-referencing. Instead, the fact that the plural form is much more frequent than the singular form should make the lexicographer attend to the structure and content of the dictionary article. In this case, it is quite clear from corpus inspection that the plural form in most cases refers to flowers in the sense 'flower plants' (i.e. including stems and leaves) while the core sense, which is far less common in everyday language, has a more regular distribution of word-form frequencies. This structure is reflected in the article structure in Lexin, where the 'plant' meaning is given as the first sense. However, in SO, the first sense given for *blomma* only refers to the often brightly-coloured reproductive part of a plant. This is probably motivated by etymology (the 'brightly-coloured flower' is older than the 'flower plant'), as well as by a tradition of trying to identify and present the core meaning of a word. What makes a plant a flower (in the second sense) is having flowers (in the first sense).

Another example of sense shifting with word form is the Swedish word *pengar* 'money', which morphologically is a regularly formed plural of the word *peng*, 'coin'. The learner's dictionary Lexin has two separate entries, one for *pengar* and one for *peng*. The latter entry lists *pengar* as the plural, i.e. 'coins', without any reference to the alternative meaning of this word form. SO only gives the entry *peng*, together with the core definition 'coin or note', and the usage information 'mostly plural'. The meaning 'money' is given as a sub-sense of the core meaning.

3.2 Frequency of inflected forms varying with orthography of the headword

A much debated issue in Swedish from a language-planning perspective is the use of English spelling. In particular, the English plural suffix *-s* is counteracted by the normative 'Swedish Academy Glossary' (*Svenska Akademiens ordlista*, abbreviated SAOL). This approach can also be found in the current edition of SO. For example, the headword *skanner* 'scanner', spelled with *k*, is provided together with the recommended indefinite plural form *skannrar* 'scanners'. SO also gives the *c*-spelling variant, *scanner*, as an alternative, and the recommended plural form *scannrar*. The *-s* plurals *skanners*/*scanners* are given as optional plural forms. Focusing on the relation between the variant spelling and the two plural forms *-rar* and *-s*, the frequency tool shows that the distribution of plural suffixes is far from even between the variant spellings. It seems that people using the more "Swedish-looking" *k*-spelling *skanner* also use the Swedish plural suffix *-rar*, whereas the (more frequent) spelling *scanner* tends to be combined with the *-s* plural suffix. This is not reflected in the article in SO. (Note that in Lexin, the plural *-s* suffix is not included as a plural variant.)

3.3 Very frequent derivations

Lexicographers' decisions about which items to be included as lexical items with individual main entries and which ones to be registered as derivatives are often unclear (Battenburg, 1992: 69). Using a frequency test can provide some interesting results.

Creating nominalizations is a conventional way of deriving Swedish verbs. Adding *-ande* to an arbitrary Swedish verb theoretically yields both nominalizations and present participles (the latter are usually adjectival or verbal). For example, from the verb *springa* 'run', one can derive *springande*, which means both 'running' and 'the act of running'. In SO, *-ande* forms are often included as words forms in the verb entries, to indicate nouns carrying the semantics of the verbal headword. However, the adjectival (participial) *-ande* forms are often missing in the dictionary even though the adjectival use of the word might be much more common than the nominal use. These "missing adjectives" can be found using frequency information, by examining words which have a high frequency of the *-ande* form compared to the frequency of the infinite headword

form. One example is the "nominalization" *ambulerande* 'moving from place to place' in relation to the verb headword **ambulera** 'move from place to place'. Looking at this word more in detail using the corpus tool Korp (Borin et al., 2012), it can be noted that the *-ande* form is primarily an adjectival form used as a modifier in noun phrases like *ambulerande tjänsteman* 'travelling administrator', *ambulerande tivoli* 'travelling amusement park', etc.; see Figure 1 below.

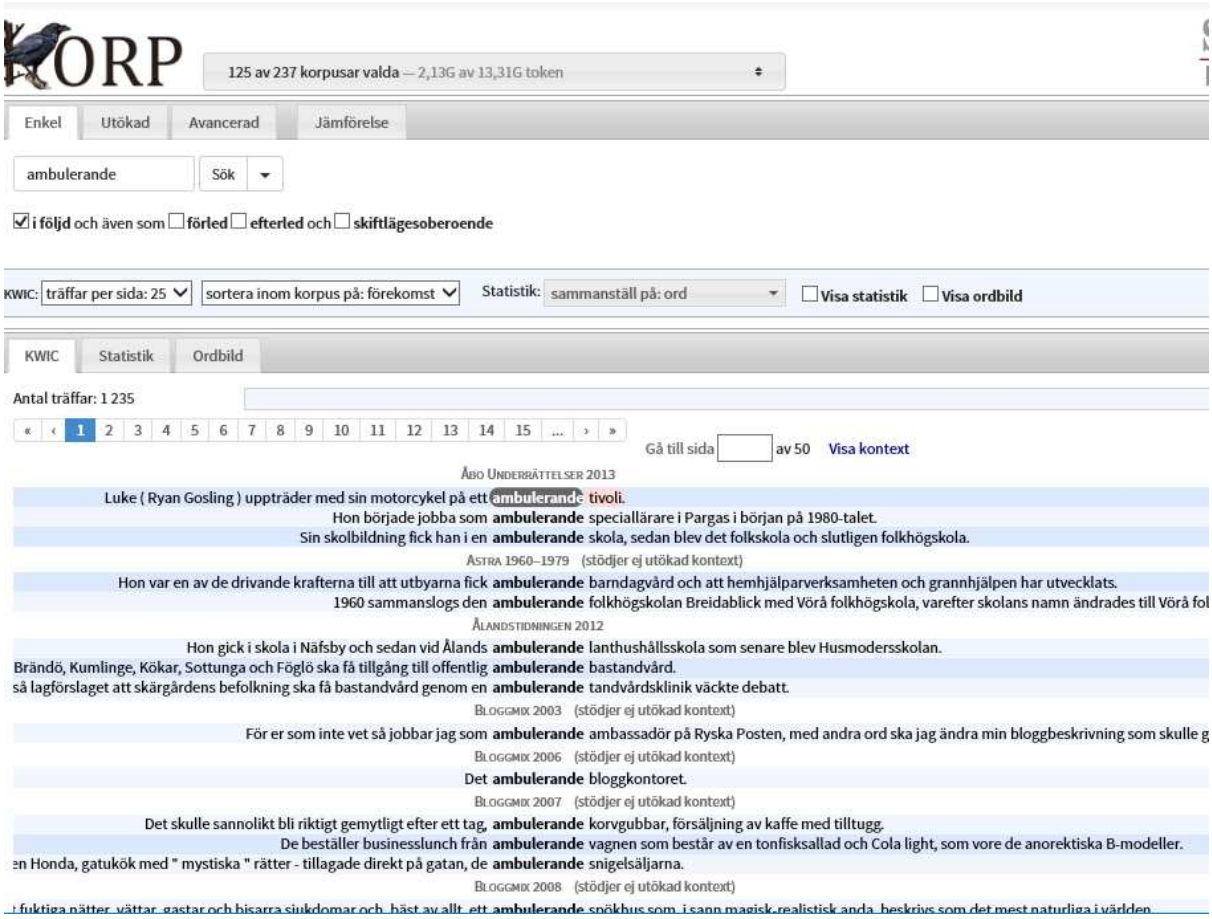


Figure 1: Korp corpus concordance search for *ambulerande*

Comparing with the learner's dictionary Lexin, it could be noted that only the verb form *ambulera* is included, although this form is infrequent in use.

3.4 Word forms in phrases which special syntactic functions

In SO, adjectives are typically illustrated as modifiers in noun phrases and as subject complements. For example, an adjective like *gul* 'yellow' is illustrated with examples such as *torrt gult gräs* 'dry yellow grass' and the subject-complement clause *bladen var gula redan i slutet av september* 'the leaves were yellow already at the end of September'. Adjectives that are mostly used serving other syntactic functions are usually marked e.g. 'typically used adverbially', while adjectives frequently used both as attributes and

as adverbs are described using syntactic examples of both use cases, with a parenthesis, "(adv.)", added to the latter.

When comparing the frequency of the headword form of the adjective with the form used in adverbials, normally a form ending in *-t* (roughly corresponding to English *-ly*), the headword form is typically more frequent than the *t*-form (see an illustration based on our material in figure 2 below)[2]. But for some adjectives, the *t*-form is much more frequent than the "base" form. Examples include **undantagslös** 'without exception' (lit. *exceptionless*), a form used only three times in our corpus, compared to 291 times for the *t*-form *undantagslöst*, and the more frequent **avsevärd** 'considerable', which is used 1,220 times in our corpus, compared to 7,018 for the *t*-form.

```
select sortering,f.tagg,concat("'",rules),ifnull(sum(FLASHBSAMH_frek),0)
from  tagg_expl t,bklasser b, fullform_utf8 f
left join korp_freks_utf8 k  on f.sortform=k.ordform_bin
where f.bklass like '41%'  and t.tagg=f.tagg and b.bklass='41'
and b.tagg=f.tagg
group by sortering,tagg
order by 1,2
```

```
select sortering,f.tagg,f.sortform,ifnull(sum(FLASHBSAMH_frek),0)
from  tagg_expl t, fullform_utf8 f
left join korp_freks_utf8 k
on f.sortform=k.ordform_bin
where f.bklass like '41%'
and t.tagg=f.tagg
and f.l_nr=350400
group by sortering,tagg
order by 1,2
```

Totalfrekvens *gemena* ordformer som ingår i böjningsmönstret

A01a	AQPUSNI	'=	9282942	**37,164**	
A01b	AQPUSGI	'=+/s	476952	**1,909**	
A01c	AQPNSNI	'=+t	7445697	**29,809**	
A01d	AQPNSGI	'=+ts	4659	**0,019**	
A02a	AQP**ND	'=+a	5751308	**23,025**	
A02b	AQP**GD	'=+/as	69772	**0,279**	
A02c	AQPMSND	'=+e	184946	**0,740**	
A02d	AQPMSGD	'=+/e/s	7303	**0,029**	
A05a	AQC**N*	'=+/are	816061	**3,267**	
A05b	AQC**G*	'=+/are/s	291	**0,001**	
A06	AQS**NI	'=+/ast	177919	**0,712**	
A07a	AQS**ND	'=+/aste	162549	**0,651**	
A07b	AQS**GD	'=+/ast/e/s	14	**0,000**	
N01a	NCUSNI	'=+/het	331674	**1,328**	
N01b	NCUSGI	'=+/het/s	1898	**0,008**	
N02a	NCUSND	'=+/het/en	128715	**0,515**	
N02b	NCUSGD	'=+/het/en/s	6526	**0,026**	
N03a	NCUPNI	'=+/het/er	103144	**0,413**	
N03b	NCUPGI	'=+/het/er/s	997	**0,004**	
N05a	NCUPND	'=+/het/er/na	23261	**0,093**	
N05b	NCUPGD	'=+/het/er/na/s	1619	**0,006**	
			24978247	**100**	

Exempelord: svensk (homograft)**

sortering	tagg	hom	ordform	antal	procent
A01a	AQPUSNI	*	svensk	94240	**33,807**
A01b	AQPUSGI	*	svensks	78	**0,028**
A01c	AQPNSNI		svenskt	20073	**7,201**
A01d	AQPNSGI		svenskts	0	**0,000**
A02a	AQP**ND	*	svenska	162172	**58,176**
A02b	AQP**GD	*	svenskas	10	**0,004**
A02c	AQPMSND		svenske	1813	**0,650**
A02d	AQPMSGD		svenskes	0	**0,000**
A05a	AQC**N*		svenskare	52	**0,019**
A05b	AQC**G*		svenskares	0	**0,000**
A06	AQS**NI		svenskast	1	**0,000**
A07a	AQS**ND		svenskaste	62	**0,022**
A07b	AQS**GD		svenskastes	0	**0,000**
N01a	NCUSNI	*	svenskhet	157	**0,056**
N01b	NCUSGI	*	svenskhets	2	**0,001**
N02a	NCUSND	*	svenskheten	92	**0,033**
N02b	NCUSGD	*	svenskhetens	9	**0,003**
N03a	NCUPNI	*	svenskheter	1	**0,000**
N03b	NCUPGI	*	svenskheters	0	**0,000**
N05a	NCUPND	*	svenskheterna	0	**0,000**
N05b	NCUPGD	*	svenskhetern as	0	**0,000**
				278762	**100**

Blad1 Blad2 Blad3

Figure 2: Frequency distribution for common adjectives and a homographic example

3.5 Word forms in multi-word expressions

Sometimes an unusual frequency distribution for the word forms of a lemma is a signal indicating that the word is primarily used in collocations. In SO, such words often lack

[2] On the worksheet illustration, the upper part shows the SQL queries (see Section 4) which provide the frequency information provided below the queries. Row 16 includes information about the headword forms (tag: AQPUSNI), and row 18 gives information about the *-t* forms (tag: AQPNSNI). The left side of the figure provides frequency information about a large class of adjectives in the morphological database, whereas the right side includes information about a single adjective, in this case the adjective *svensk* 'Swedish'. Figures in bold refer to the percentages of the word forms.

a definition and are introduced by the formula "I frasen …" ('in the phrase …'). An example is the word *aftonkvisten* (lit. *the branch of evening*), which is principally only used in the definite singular in the prepositional phrase *på aftonkvisten* 'between afternoon and evening' (lit. *on the branch of evening*). An examination of word-form frequencies confirms this.

Another example is the lemma noun *sort* 'kind', where the genitive singular form, *sorts*, is almost 600% more frequent than the nominative headword form (the singular nominative is normally much more frequent than the singular genitive). A concordance study reveals that a great number of the genitive forms make up the very frequent classifying construction *en sorts* + NOUN, as in *en sorts frukt* 'a kind of fruit' (lit. *a kind.*GEN *fruit*). The fact that the genitive form is included in a collocation and the fact that only the genitive form is allowed in this collocation is not really clear, neither in SO, nor in Lexin.

4. The morphological corpus frequency tool

The morphological frequency tool stores information on word forms and frequencies for a number of corpora in a format which is easily combined with dictionary information on headwords, inflection groups, and inflected forms. The frequency information is retrieved from the Korp corpus tool and "de-lemmatized", i.e. stripped of lemma information before being stored in a relational database which can be accessed using standard tools like MySQL Workbench. The frequency information is used in combination with rule-based morphological data based on derivational and inflectional information from Svensk Morfologisk Databas ('The Swedish Morphological Database', Berg & Cederholm, 2001) compiled at the University of Gothenburg and the lexical database owned by the Swedish Academy.

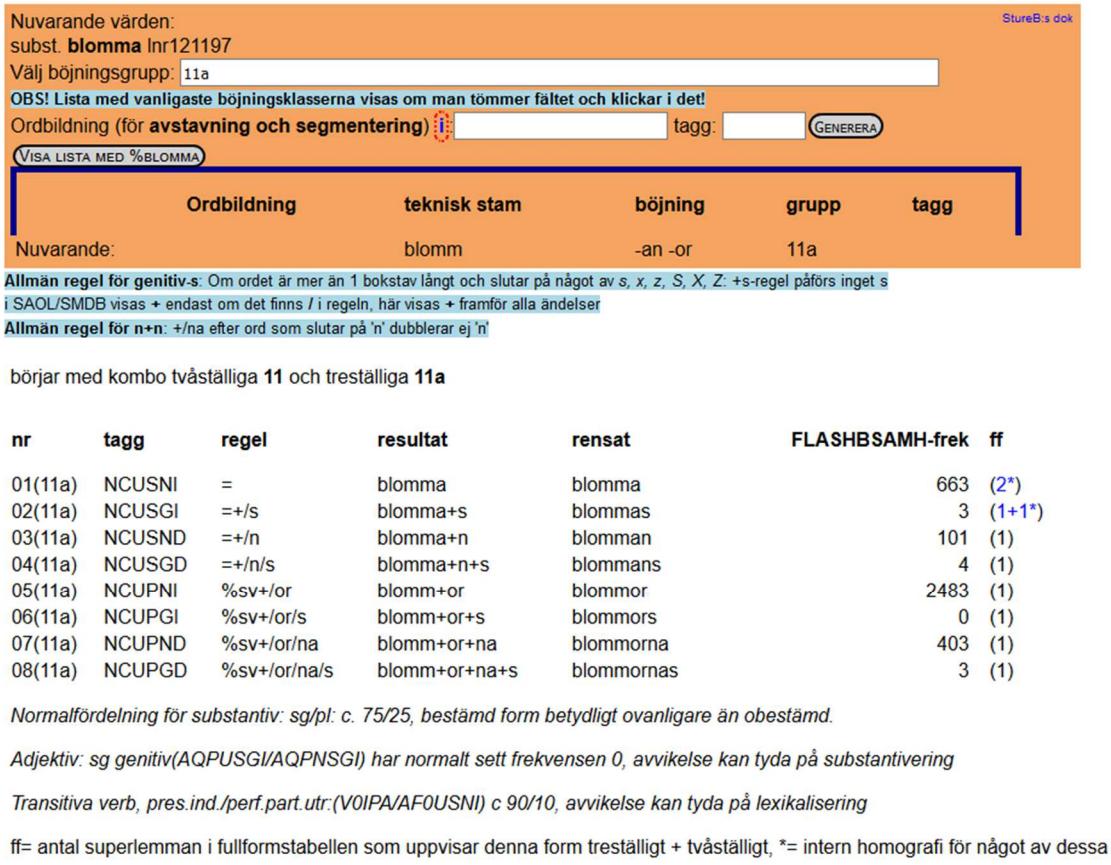
Part of the information is integrated in the editorial interface, for the convenience of the editors, while such tasks as retrieving lists of candidates for closer examination are carried out with the help of stored procedures.

A stored procedure in a relational database management system serves as a means to store a group of SQL statements with an assigned name, which can be called using parameters. We use stored procedures to create and examine word-form distribution tables based on joining the inflectional information from the dictionary database with frequency data from the corpus frequency database. For each word in the dictionary database the editor can enter a code indicating the inflectional paradigm (see subSection 4.1). Entering or changing the code generates a "blow-up" of all word forms with associated tags – up to over 20 forms for some verbs, including derived participles – which are stored in a table. This table, at the moment holding information on approximately 1,800,000 word forms, is immediately available for joining with corpus-frequency information for presentation in the editor interface, and it provides up-to-date information for the stored procedures. The inflectional paradigm code system and

associated tags and rules for word-form generation are a development of the system used for Svensk Morfologisk Databas.

4.1 Editorial interface

Each word in the dictionary is classified as belonging to an inflectional group, and in the process of this classification the editor is presented with frequency information for the actual word forms, which can be compared to some basic metrics indicating normal frequency distribution.



Nuvarande värden: StureB:s dok
subst. **blomma** lnr121197
Välj böjningsgrupp: 11a
OBS! Lista med vanligaste böjningsklasserna visas om man tömmer fältet och klickar i det!
Ordbildning (för **avstavning och segmentering**) tagg: GENERERA
VISA LISTA MED %BLOMMA

	Ordbildning	teknisk stam	böjning	grupp	tagg
Nuvarande:		blomm	-an -or	11a	

Allmän regel för genitiv-s: Om ordet är mer än 1 bokstav långt och slutar på något av *s, x, z, S, X, Z*: +s-regel påförs inget s
i SAOL/SMDB visas **+** endast om det finns **/** i regeln, här visas **+** framför alla ändelser
Allmän regel för n+n: +/na efter ord som slutar på 'n' dubblerar ej 'n'

börjar med kombo tvåställiga **11** och treställiga **11a**

nr	tagg	regel	resultat	rensat	FLASHBSAMH-frek	ff
01(11a)	NCUSNI	=	blomma	blomma	663	(2*)
02(11a)	NCUSGI	=+/s	blomma+s	blommas	3	(1+1*)
03(11a)	NCUSND	=+/n	blomma+n	blomman	101	(1)
04(11a)	NCUSGD	=+/n/s	blomma+n+s	blommans	4	(1)
05(11a)	NCUPNI	%sv+/or	blomm+or	blommor	2483	(1)
06(11a)	NCUPGI	%sv+/or/s	blomm+or+s	blommors	0	(1)
07(11a)	NCUPND	%sv+/or/na	blomm+or+na	blommorna	403	(1)
08(11a)	NCUPGD	%sv+/or/na/s	blomm+or+na+s	blommornas	3	(1)

Normalfördelning för substantiv: sg/pl: c. 75/25, bestämd form betydligt ovanligare än obestämd.

Adjektiv: sg genitiv(AQPUSGI/AQPNSGI) har normalt sett frekvensen 0, avvikelse kan tyda på substantivering

Transitiva verb, pres.ind./perf.part.utr:(V0IPA/AF0USNI) c 90/10, avvikelse kan tyda på lexikalisering

ff= antal superlemman i fullformstabellen som uppvisar denna form treställigt + tvåställigt, *= intern homografi för något av dessa

Figure 3: Frequency information shown while editing inflectional information for a dictionary entry

The example in Figure 3 shows how the word forms generated by the inflectional code *11a*, when applied to the noun *blomma* ('flower'), are presented to the editor together with frequency information from the corpus *FLASHBSAMH* (a popular internet discussion group). These figures can be compared to the hint below the form-frequency table regarding 'normal' distribution for singular and plural forms (roughly 75% and 25%, respectively) and definite and indefinite forms (the definite forms being much less frequent).

The existence of homographic word forms can obscure this kind of comparison, so the rightmost column shows the number of homographic word forms for each form. In this case there is indeed a homograph to the singular indefinite form, the verb *blomma* ('to bloom'), which means the singular noun form *blomma* ('flower') is in fact even less frequent than shown in the table.

While editing a dictionary entry, the editor has an integrated view of 1) the updated entry, 2) the published version of SO, 3) the word forms given in the latest version of the Swedish Academy Glossary, and 4) an overview of word form frequencies for a number of corpora of contemporary Swedish. See Figure 4:

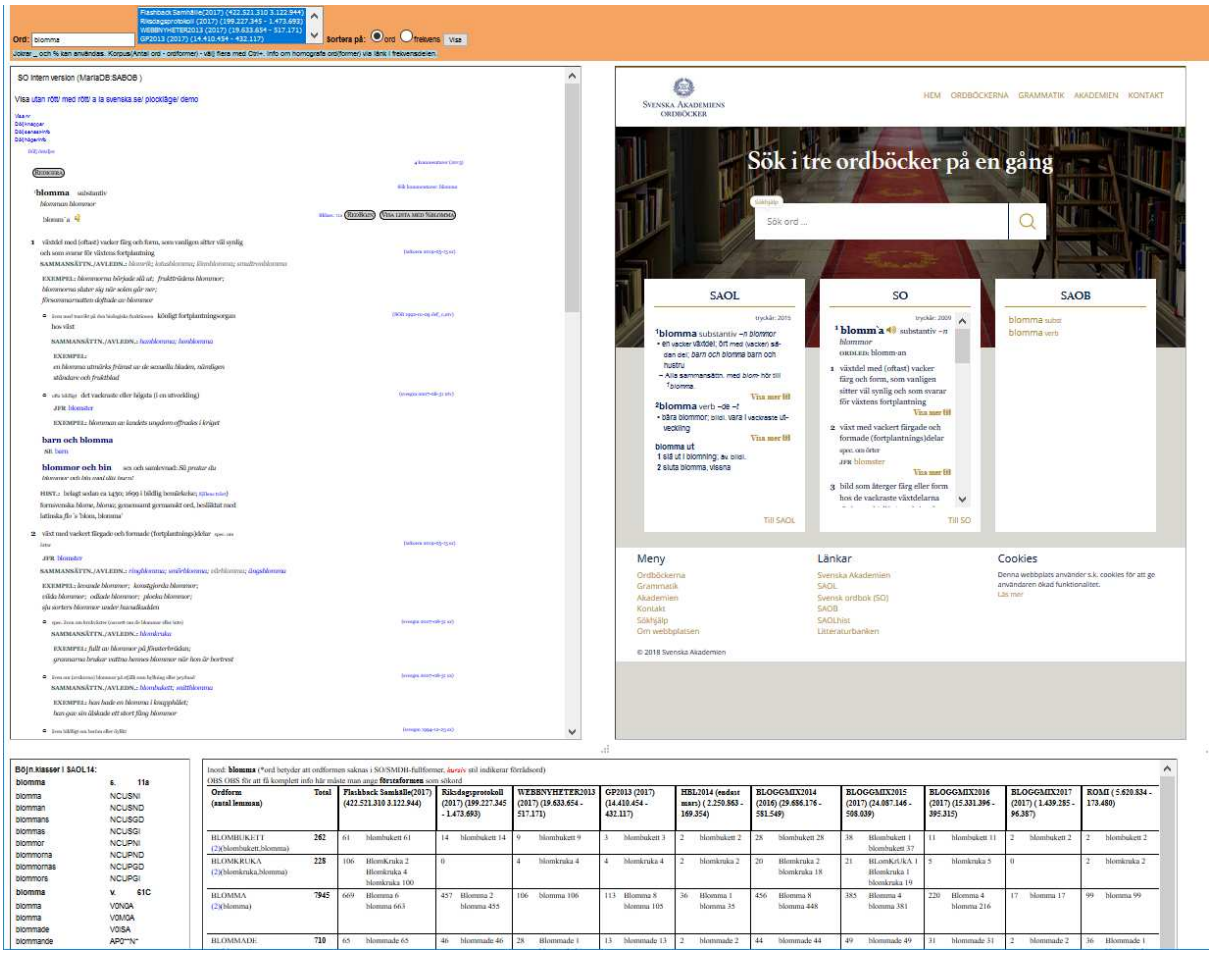
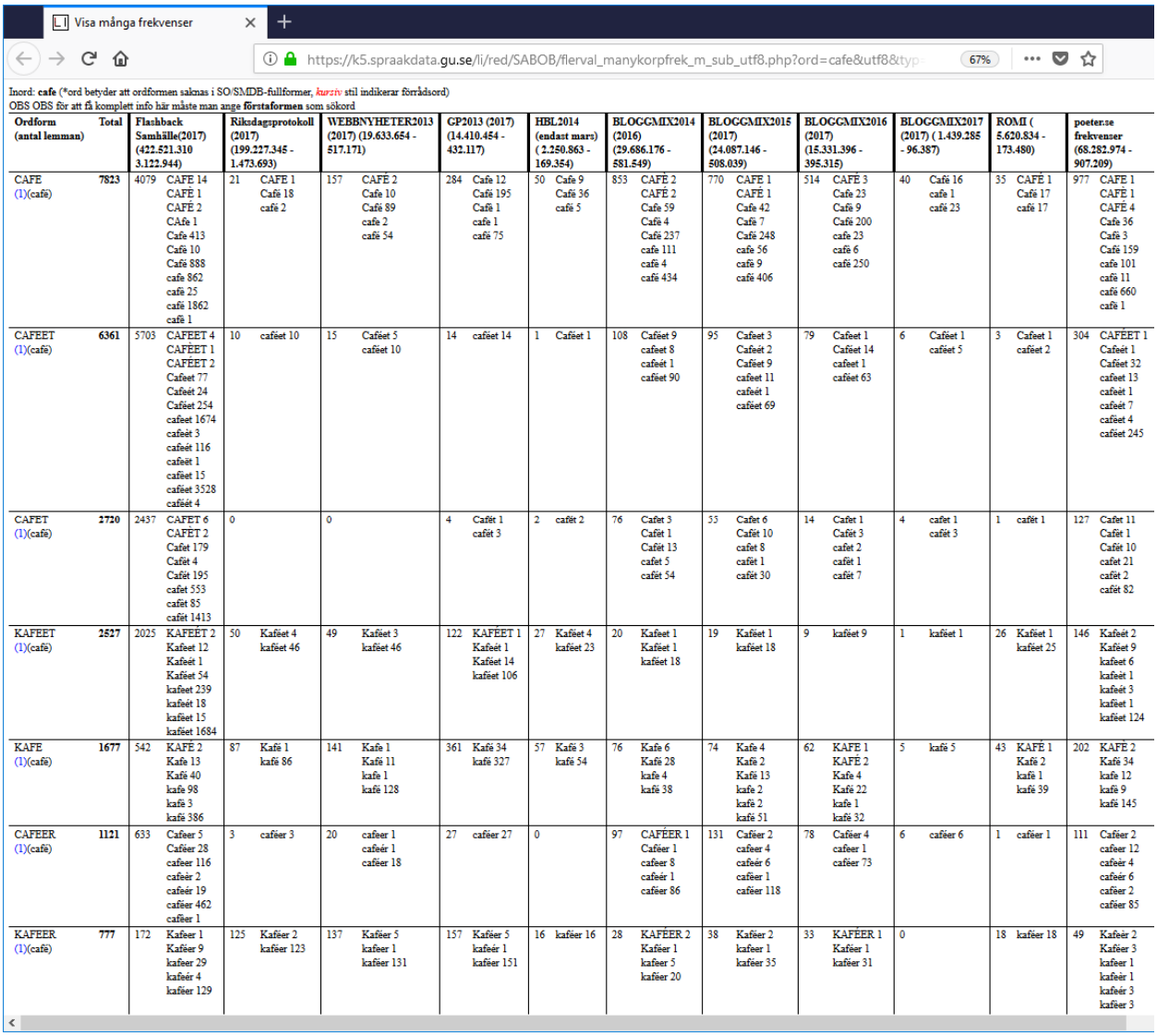


Figure 4: Editors' view of entry being edited, published dictionary entry, and word form frequencies in different corpora

The frequency-distribution view can also be used to check the relative frequency of different spellings. Cf. for example the word *kafé*/*café* ('café') in SO, which is more often spelled *café* in our corpus of contemporary Swedish (Figure 5). The former spelling variant *café* has now been upgraded to headword status, while the former headword, *kafé*, is considered a spelling variant.



Visa många frekvenser

https://k5.spraakdata.gu.se/li/red/SABOB/flerval_manykorpfrek_m_sub_utf8.php?ord=cafe&utf8&typ… 67%

Inord: **cafe** (*ord betyder att ordformen saknas i SO/SMDB-fullformer, *kursiv* stil indikerar förrådsord)
OBS OBS för att få komplett info här måste man ange **förstaformen** som sökord

Ordform (antal lemman)	Total	Flashback Samhälle(2017) (422.521.310 3.122.944)	Riksdagsprotokoll (2017) (199.227.345 - 1.473.693)	WEBBNYHETER2013 (2017) (19.633.654 - 517.171)	GP2013 (2017) (14.410.454 - 432.117)	HBL2014 (endast mars) (2.250.863 - 169.354)	BLOGGMIX2014 (2016) (29.686.176 - 581.549)	BLOGGMIX2015 (2017) (24.087.146 - 508.039)	BLOGGMIX2016 (2017) (15.331.396 - 395.315)	BLOGGMIX2017 (2017) (1.439.285 - 96.387)	ROMI (5.620.834 - 173.480)	poeter.se frekvenser (68.282.974 - 907.209)
CAFE (1)(café)	7823	4079 CAFE 14 CAFÉ 1 CAFÉ 2 CAfe 1 Cafe 413 Café 10 Café 888 cafe 862 café 25 café 1862 café 1	21 CAFE 1 Café 18 café 2	157 CAFÉ 2 Cafe 10 Café 89 cafe 2 café 54	284 Cafe 12 Café 195 Café 1 cafe 1 café 75	50 Cafe 9 Café 36 café 5	853 CAFÉ 2 CAFÉ 2 Cafe 59 Cafe 4 Café 237 cafe 111 café 4 café 434	770 CAFE 1 CAFÉ 1 Cafe 42 Cafe 7 Café 248 cafe 56 café 9 café 406	514 CAFÉ 3 Cafe 23 Café 9 Café 200 cafe 23 café 6 café 250	40 Café 16 cafe 1 café 23	35 CAFÉ 1 Café 17 café 17	977 CAFE 1 CAFÉ 1 CAFÉ 4 Cafe 36 Café 3 Café 159 cafe 101 café 11 café 660 café 1
CAFEET (1)(café)	6361	5703 CAFEET 4 CAFÉET 1 CAFÉET 2 Cafeet 77 Cafeét 24 Café̋t 254 cafeet 1674 cafeèt 3 cafeét 116 cafeët 1 cafeet 15 cafeet 3528 cafeét 4	10 cafeet 10	15 Cafeet 5 cafeet 10	14 cafeet 14	1 Cafeet 1	108 Cafeet 9 cafeet 8 cafeét 1 cafeet 90	95 Cafeet 3 Cafeét 2 Cafeet 9 cafeet 11 cafeét 1 cafeet 69	79 Cafeet 1 Cafeet 14 cafeet 1 cafeet 63	6 Cafeet 1 cafeet 5	3 Cafeet 1 cafeet 2	304 CAFEET 1 Cafeet 1 Cafeet 32 cafeet 13 cafeët 1 cafeét 7 cafeet 4 cafeet 245
CAFET (1)(café)	2720	2437 CAFET 6 CAFÉT 2 Cafet 179 Café̋t 4 Café̋t 195 cafet 553 cafet 85 cafet 1413	0	0	4 Café̋t 1 cafet 3	2 cafet 2	76 Cafet 3 Café̋t 1 Café̋t 13 cafet 5 café̋t 54	53 Cafet 6 Café̋t 10 cafet 8 café̋t 1 café̋t 30	14 Cafet 1 Café̋t 3 cafet 2 café̋t 1 café̋t 7	4 cafet 1 café̋t 3	1 café̋t 1	127 Cafet 11 Café̋t 1 Café̋t 10 cafet 21 café̋t 2 café̋t 82
KAFEET (1)(café)	2527	2025 KAFEÉT 2 Kafeet 12 Kafeét 1 Kafeet 54 kafeet 239 kafeét 18 kafeet 15 kafeet 1684	50 Kafeet 4 kafeet 46	49 Kafeet 3 kafeet 46	122 KAFÉET 1 Kafeét 1 Kafeét 14 kafeet 106	27 Kafeet 4 kafeet 23	20 Kafeet 1 Kafeét 1 kafeet 18	19 Kafeet 1 kafeet 18	9 kafeet 9	1 kafeet 1	26 Kafeet 1 kafeet 25	146 Kafeet 2 Kafeet 9 kafeet 6 kafeèt 1 kafeèt 3 kafeet 1 kafeet 124
KAFE (1)(café)	1677	542 KAFÉ 2 Kafe 13 Kafé 40 kafe 98 kafé 3 kafé 386	87 Kafé 1 kafé 86	141 Kafe 1 Kafé 11 kafe 1 kafé 128	361 Kafé 34 kafé 327	57 Kafé 3 kafé 54	76 Kafe 6 Kafé 28 kafe 4 kafé 38	74 Kafe 4 Kafé 2 Kafé 13 kafe 2 kafé 2 kafé 51	62 KAFE 1 KAFÉ 2 Kafe 4 Kafé 22 kafe 1 kafé 32	5 kafé 5	43 KAFÉ 1 Kafé 2 kafé 1 kafé 39	202 KAFÉ 2 Kafé 34 kafe 12 kafé 9 kafé 145
CAFEER (1)(café)	1121	633 Cafeer 5 Cafeér 28 cafeer 116 cafeèr 2 cafeér 19 cafeer 462 cafeer 1	3 cafeer 3	20 cafeer 1 cafeér 1 cafeer 18	27 cafeer 27	0	97 CAFÉER 1 Cafeer 1 cafeer 8 cafeér 1 cafeer 86	131 Cafeer 2 cafeer 4 cafeér 6 cafeer 1 cafeer 118	78 Cafeer 4 cafeer 1 cafeer 73	6 cafeer 6	1 cafeer 1	111 Cafeer 2 cafeer 12 cafeèr 4 cafeér 6 cafeer 2 cafeer 85
KAFEER (1)(café)	777	172 Kafeer 1 Kafeer 9 kafeer 29 kafeér 4 kafeer 129	125 Kafeer 2 kafeer 123	137 Kafeer 5 kafeer 1 kafeer 131	157 Kafeer 5 kafeér 1 kafeer 151	16 kafeer 16	28 KAFÉER 2 Kafeer 1 kafeer 5 kafeer 20	38 Kafeer 2 kafeer 1 kafeer 35	33 KAFÉER 1 Kafeer 1 kafeer 31	0	18 kafeer 18	49 Kafeer 2 Kafeer 3 kafeer 1 kafeèr 1 kafeér 3 kafeer 3

Figure 5: Frequency distribution of *kafé* and *café* (and inflections of the two spelling variants)

4.2 Back-office SQL tool

The 'back-office' SQL query tool provides access to stored procedures which are used for comparing word form distribution in selected corpora. These procedures are used for identifying words the presentation of which might need to be reviewed and updated based on the actual use. For example, the syntactic examples given in the dictionary should reflect the actual use.

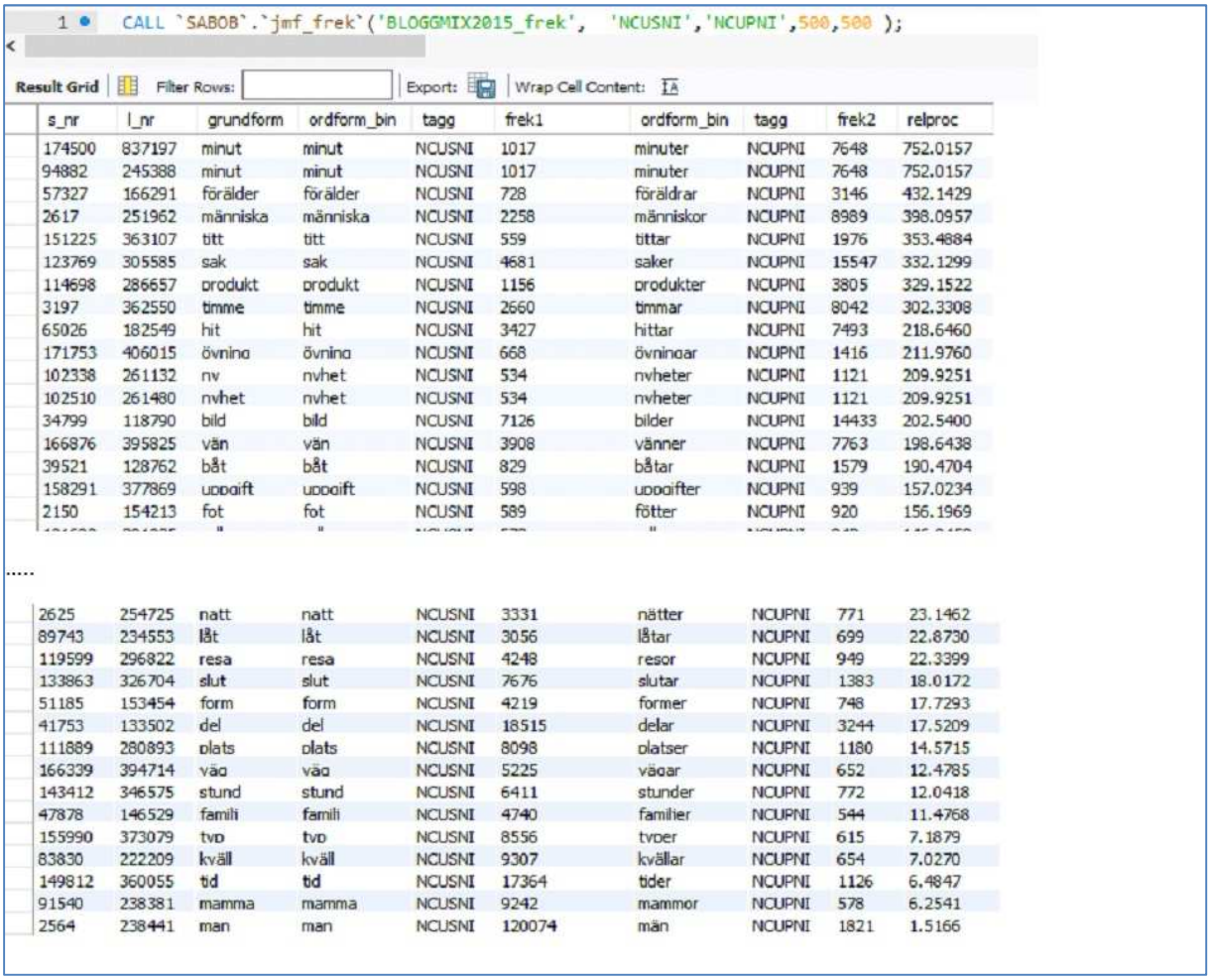
A call to a stored procedure can look like this:

CALL jmf_frek ('BLOGGMIX2015_frek', 'NCUSNI','NCUPNI',500,500)

Here, *BLOGGMIX2015_frek* is the corpus used to extract word-form frequencies, *NCUSNI* and *NCUPNI* are the word-form tags to be compared (in this case *indefinite*

singular vs. *indefinite plural* for neuter nouns) and the last two figures set the threshold for words to be considered, the minimum frequency for each of the two word forms.

This call returns a table of words ordered by the relative percentage of the frequencies of the two word forms in the corpus (Figure 6). In this case, the top row holds the pair *minut/minuter* ('minute/minutes') with 1,017 occurrences of the singular indefinite vs. 7,648 occurrences of the plural indefinite, giving a relative percentage of 752%. The last row holds the pair *man/män* ('man/men') with a relative frequency of 1.5%. A quick check shows that the singular form *man* is homographic with other very frequent words (e.g. the generic pronoun *man* 'one'), which means our frequency information is not useful as a source of information regarding this word. The second last row displays *mamma/mammor* ('mum/mums'), having a relative frequency of 6.25% for the plural.



```
1 •  CALL `SABOB`.`jmf_frek`('BLOGGMIX2015_frek',  'NCUSNI','NCUPNI',500,500 );
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

s_nr	l_nr	grundform	ordform_bin	tagg	frek1	ordform_bin	tagg	frek2	relproc
174500	837197	minut	minut	NCUSNI	1017	minuter	NCUPNI	7648	752.0157
94882	245388	minut	minut	NCUSNI	1017	minuter	NCUPNI	7648	752.0157
57327	166291	förälder	förälder	NCUSNI	728	föräldrar	NCUPNI	3146	432.1429
2617	251962	människa	människa	NCUSNI	2258	människor	NCUPNI	8989	398.0957
151225	363107	titt	titt	NCUSNI	559	tittar	NCUPNI	1976	353.4884
123769	305585	sak	sak	NCUSNI	4681	saker	NCUPNI	15547	332.1299
114698	286657	produkt	produkt	NCUSNI	1156	produkter	NCUPNI	3805	329.1522
3197	362550	timme	timme	NCUSNI	2660	timmar	NCUPNI	8042	302.3308
65026	182549	hit	hit	NCUSNI	3427	hittar	NCUPNI	7493	218.6460
171753	406015	övning	övning	NCUSNI	668	övningar	NCUPNI	1416	211.9760
102338	261132	nv	nvhet	NCUSNI	534	nvheter	NCUPNI	1121	209.9251
102510	261480	nvhet	nvhet	NCUSNI	534	nvheter	NCUPNI	1121	209.9251
34799	118790	bild	bild	NCUSNI	7126	bilder	NCUPNI	14433	202.5400
166876	395825	vän	vän	NCUSNI	3908	vänner	NCUPNI	7763	198.6438
39521	128762	båt	båt	NCUSNI	829	båtar	NCUPNI	1579	190.4704
158291	377869	uppgift	uppgift	NCUSNI	598	uppgifter	NCUPNI	939	157.0234
2150	154213	fot	fot	NCUSNI	589	fötter	NCUPNI	920	156.1969
.....									
2625	254725	natt	natt	NCUSNI	3331	nätter	NCUPNI	771	23.1462
89743	234553	låt	låt	NCUSNI	3056	låtar	NCUPNI	699	22.8730
119599	296822	resa	resa	NCUSNI	4248	resor	NCUPNI	949	22.3399
133863	326704	slut	slut	NCUSNI	7676	slutar	NCUPNI	1383	18.0172
51185	153454	form	form	NCUSNI	4219	former	NCUPNI	748	17.7293
41753	133502	del	del	NCUSNI	18515	delar	NCUPNI	3244	17.5209
111889	280893	plats	plats	NCUSNI	8098	platser	NCUPNI	1180	14.5715
166339	394714	väg	väg	NCUSNI	5225	vägar	NCUPNI	652	12.4785
143412	346575	stund	stund	NCUSNI	6411	stunder	NCUPNI	772	12.0418
47878	146529	famili	famili	NCUSNI	4740	familier	NCUPNI	544	11.4768
155990	373079	typ	typ	NCUSNI	8556	typer	NCUPNI	615	7.1879
83830	222209	kväll	kväll	NCUSNI	9307	kvällar	NCUPNI	654	7.0270
149812	360055	tid	tid	NCUSNI	17364	tider	NCUPNI	1126	6.4847
91540	238381	mamma	mamma	NCUSNI	9242	mammor	NCUPNI	578	6.2541
2564	238441	man	man	NCUSNI	120074	män	NCUPNI	1821	1.5166

Figure 6: Stored procedure for comparing word forms.

The words in the top and bottom of the table stand out, and this could be important information for the dictionary user. For the word *minut*, the comment 'mostly in the plural form' could be added in the entry, and cases where the singular form is used might need to be analysed. Do examples like 'Give me a minute!' and 'it took him 15

minutes' fit under the same definition? For the word *mamma*, the overwhelming use of the singular indefinite bare form in this corpus is probably an indication of this word form often functioning as a name rather than an ordinary noun. *Mamma var här i går* ('mother was here yesterday') does not mean that an indefinite or a generic mother was here.

SQL queries are used for establishing the "normal" distribution of word-form frequencies for each word class, eventually resulting in informational hints to the editors in the editorial interface (Figure 3 above). Finding the normal distribution is done by excluding homographic word forms from accumulated queries and the result is validated through comparison with the frequency information for some typical words (Figure 2 above).

5. Suggestions and discussion

Here, we provide examples of "candidates for change" found using our morphological frequency tool and typical considerations that arise when studying actual words and how they are presented in the dictionaries.

5.1 Word forms with a slightly different meaning than the base form: Revise article structure or content?

When the word *blomma* 'flower' was identified as having a non-standard distribution of word form frequencies we reviewed the examples given in the dictionary entries in SO and Lexin. In both dictionaries the very common phrase *plocka blommor* ('pick flowers') is given as an example for the first sense of the word, which was surprising as the dictionaries have ordered the senses differently. The outcome was a decision to move not only this, but several syntactic examples from the first 'colourful reproductive part of a plant' to the second, 'plant with flowers' sense in the coming edition of SO.

5.2 Frequency of inflected forms varying with orthography of the headword: Change lemma form?

As for the word *skanner/scanner* with optional plurals *-rar/-s* discussed in Section 3.2, the more frequent form *scanner* will be the headword in the coming, updated version of SO. Our investigation suggests the plural *scanners* should be given as the preferred plural form for *scanner*, while *skannrar* would be the preferred plural for *skanner*, but the formal decision still has to be made.

5.3 Very frequent derivations

As for the word *ambulerande*, the present participle and *nomen actionis* of the verb *ambulera*, discussed in Section 3.3, there is already a note "(ofta pres. part.)", 'often (used as) present participle' in the existing dictionary entry (see Figure 7). We also suggest adding a syntactic example illustrating this usage, as we cannot expect all

dictionary users to be familiar with the implications of the grammatical note.

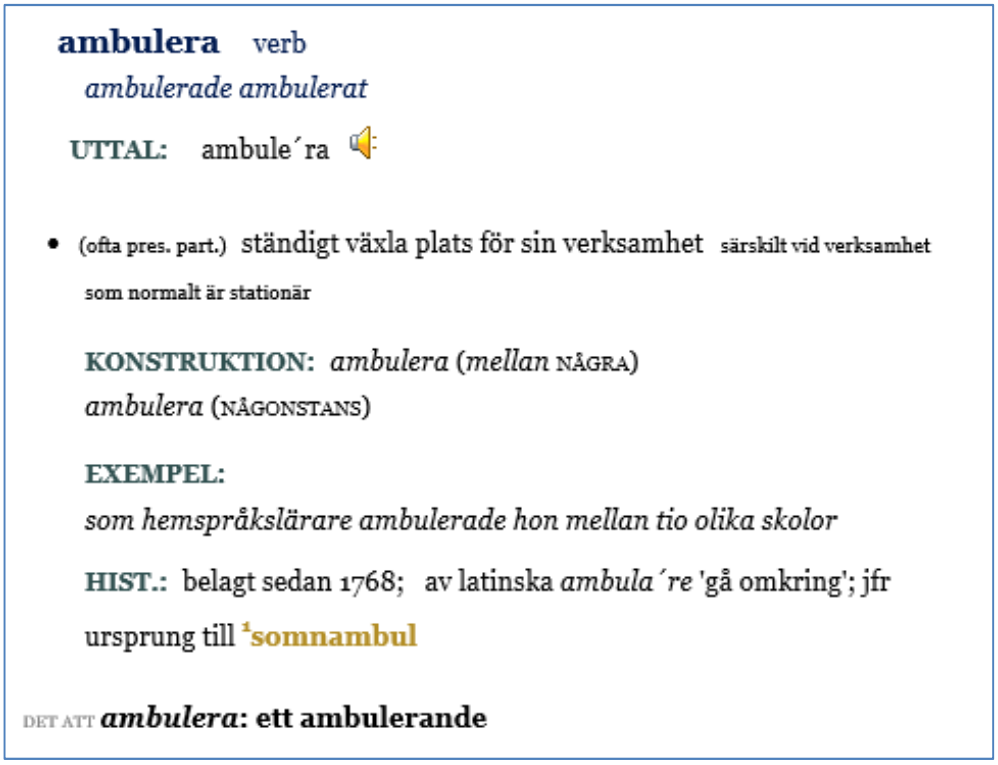


Figure 7: The verb *ambulera* lacking an example for the pres. participle *ambulerande*

5.4 Word forms in phrases which special syntactic functions

Our investigation suggests that an explicit grammatical note regarding the (almost exclusively) adverbial usage of the word *undantagslös* should be added to the dictionary, in conformance with how other similarly behaving words are presented.

5.5 Word form signalling multi-word expression

Certain words, regularly restricted to particular inflections, are almost exclusively associated with special constructions, for example the plural-only noun *döddagar* (lit. *dying-days*) in the prepositional phrase *till döddagar* 'to my dying day/to the end of time' or the fossilized indefinite singular *korvspad* 'sausage stock' in the adjective phrase *klart som korvspad* 'plain as a pikestaff' (see Sköldberg, 2007; Sköldberg, 2008). For these words it is, of course, essential that the special constructions they are associated with appear in the dictionary entry. Other words might occur frequently in an inflected form in common collocations that could be identified by corpora searches initiated based on word form frequency distribution anomalies. As yet, we have not had the time to do such systematic corpora searches.

5.6 Final discussion

The tools described in this paper have only been available to us for a limited time and we are still in a process of learning how to best take advantage of the new possibilities at hand. Moreover, limited personnel resources have not allowed for a thorough investigation of all words with an exceptional frequency distribution of word forms, but already looking at a few of these words has proved to us that the word-form relative frequency information gives a valuable additional aspect of knowledge to the lexicographer, providing a means to add quality to dictionary entries.

Providing an optimal toolbox for lexicographers, and giving the right amount of useful information at the right time, is a challenging task. Overloading the editorial interface with too much information can be perceived as a hindrance to the creative work of writing, but no-one is happy with getting important information too late in the process, when already having moved on mentally to the next task. The process of reviewing and enhancing dictionary content, and the tools provided for supporting this process, must therefore ideally be developed in close co-operation between system developers and editors.

6. References

Atkins, S. B. T. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography.* Oxford: Oxford University Press.

Battenburg, J. (1992). *English Monolingual Learners' Dictionaries: A User-Oriented Study.* Tübingen: Niemeyer Press.

Berg, S. & Cederholm, Y. (2001). Att hålla på formerna. Om framväxten av Svensk morfologisk databas. In S. Allén (ed.) *Gäller stam, suffix och ord. Festskrift till Martin Gellerstam den 15 oktober 2001.* Göteborg: Meijerbergs institut för svensk etymologisk forskning, pp. 58–69.

Borin, L, Forsberg, M. & Roxendal, J. (2012). Korp – the corpus infrastructure of Språkbanken. *Proceedings of LREC 2012.* Istanbul: ELRA, pp. 474–478.

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes*, 10, pp. 425–455.

Dąbrowska, E. (2008). The effects of frequency and neighbourhood density on adult speakers' productivity with Polish case inflections: An empirical test of usage-based approaches to morphology. *Journal of Memory and Language*, 58, pp. 931–951.

Hannesdóttir, A. H. (1998). *Lexikografihistorisk spegel. Den enspråkiga svenska lexikografins utveckling ur den tvåspråkiga.* Göteborg: Meijerbergs institut för svensk etymologisk forskning.

Hay, J. (2001). Lexical frequency in morphology: is everything relative? *Linguistics*, 39(6), pp. 1041–1070.

Heuberger, R. (2018). Dictionaries to assist teaching and learning. In P. A. Fuertes-

Olivera (ed.) *The Routledge Handbook of Lexicography*. Milton Park, Abingdon, Oxon: Routledge, pp. 300–316.

Lew, R. (2012). How can we make electronic dictionaries more effective? In S. Granger & M. Paquot (eds.) *Electronic Lexicography*. Oxford: Oxford University Press, pp. 343–361.

Lexin introduction: Om lexinordböckerna. Accessed at: http://lexin.nada.kth.se/lexin/doc/Lexin.pdf. (3 June 2019)

Lexin: *Lexins svenska lexikon*. Accessed at: http://lexin.nada.kth.se/lexin/. (3 June 2019)

Schenberg, P. (1739). *Lexicon latino-svecanum*. Norcopiæ: C.F. Broocmanno.

Sköldberg, E. (2007). Birds of a feather and sheep's clothing: on unique constituents in Swedish. In M. Nenonen & S. Niemi (eds.) *Collocations and Idioms 1. Papers from the First Nordic Conference on Syntactic Freezes Joensuu, May 19-20, 2006*. Joensuu: Joensuun yliopisto, pp. 297–305.

Sköldberg, E. (2008). Från *vrångstrupen* till *fotabjället*. Om presentationen av unika konstituenter i Svensk ordbok utgiven av Svenska Akademien. In Á. Svavarsdóttir, G. Kvaran, G. Ingólfsson & J. H. Jónsson (eds.) *Nordiske Studier i Leksikografi 9. Rapport fra konference om leksikografi i Norden. Akureyri 22.–26. maj 2007*. Reykjavík: Nordisk forening for leksikografi, pp. 421–432.

Sköldberg, E. (2017). Innehållet i Svensk ordbok utgiven av Svenska Akademien – eller kampen mellan norm och bruk. In S. Bendegard, U. Melander Marttala & M. Westman (eds.) *Språk och norm. Rapport från ASLA:s symposium Uppsala universitet 21–22 april 2016*. Uppsala, ASLA: Svenska föreningen för tillämpad språkvetenskap, pp. 123–129.

Spegel, H. (1712). *Glossarium – Sveo-Gothicum Eller Swensk-Ordabook*. Lund: A. Habereger.

SO: *Svensk ordbok utgiven av Svenska Akademien*. (2009). Accessed at: https://svenska.se/so/. (3 June 2019)

Svensén, B. (2009). *A handbook of lexicography. The theory and practice of dictionary-making*. Cambridge: Cambridge University Press.

Trap-Jensen, L., Lorentzen, H. & Sørensen, N. H. (2014). An odd couple – corpus frequency and look-up frequency: What relationship? *Slovenščina 2.0*, 2(2), pp. 94–113.

Trueswell, J. C. (1996). The role of lexical frequency in syntactic ambiguity resolution. *Journal of Memory and Language*, 35, pp. 566–585.