Between Grammars and Dictionaries: a Swedish Constructicon

Emma Sköldberg, Linnéa Bäckström, Lars Borin, Markus Forsberg, Benjamin Lyngfelt, Leif-Jöran Olsson, Julia Prentice, Rudolf Rydstedt, Sofia Tingsell, Jonatan Uppström

Department of Swedish, University of Gothenburg,
PO Box 200, SE-405 30 Gothenburg, Sweden
E-mail: {emma.skoeldberg; linnea.backstrom; lars.borin; markus.forsberg; benjamin.lyngfelt; leif-joran.olsson; julia.prentice; rudolf.rydstedt; sofia.tingsell; jonatan.uppstrom}@svenska.gu.se

Abstract

This paper introduces the Swedish Constructicon (SweCxn), a database of Swedish constructions currently under development. We also present a small study of the treatment of constructions in Swedish (paper) dictionaries, thus illustrating the need for a constructionist approach, and discuss three different methods used to identify potential constructions for inclusion in the Constructicon. SweCxn is a freely available electronic resource, with a particular focus on semi-general linguistic patterns of the type that are difficult to account for from a purely lexicographic or grammatical perspective, and which therefore have tended to be neglected in both dictionaries and grammars. Far from being a small set of borderline cases, such constructions are both numerous and common. They are also quite problematic for second language acquisition as well as LT applications. Accordingly, various kinds of multi-word units have received more attention in recent years, not least from a lexicographic perspective. The coverage, however, is only partial, and the productivity of many constructions is hard to capture from a lexical viewpoint. To identify constructions for SweCxn, we use a combination of methods, such as working from existing construction descriptions for Swedish and other languages, applying LT tools to discover recurring patterns in texts, and extrapolating constructional information from dictionaries.

Keywords: lexicography, construction, constructicon, Swedish, FrameNet, language technology

1. Introduction

Linguistic patterns that are too specific to be treated as general rules and too general to be tied to specific words are peripheral from both a grammatical and a lexicographic point of view. Hence, such patterns, which may be called constructions (cx), have (traditionally) tended to be neglected in grammars as well as dictionaries. Some typical Swedish examples are conventionalized time expressions like "[minuttal] i/över [timtal]" '[minutes] to/past [hour]' and semi-prefab phrases such as "i ADJEKTIV-aste laget" 'in ADJECTIVE-superlative the-measure'. The latter cx basically means 'too much' of the quality expressed by the adjective: *i hetaste laget* 'too hot for comfort', *i minsta laget* 'a bit on the small side' and *i senaste laget* 'at the last moment'.

These examples are partially schematic multi-word units, i.e. structures where at least one component is lexically fixed and at least one represents a morpho-syntactic category. Accounting for such constructions is a main priority for the Swedish Constructicon (SweCxn) currently under development. The resource is based on principles of Construction Grammar and developed as an addition to the Swedish FrameNet (SweFN). It is integrated with other freely available resources in Språkbanken (the Swedish Language Bank) by linked lexical entries (Lyngfelt et al., 2012). In most respects, the Swedish Constructicon is modelled on its English counterpart in Berkeley, and, thus, mostly adhering to the FrameNet format (see Fillmore, 2008; Fillmore et al., 2012). The SweCxn project is highly cross-disciplinary, involving experts on (construction) grammar, language technology, lexicography, phraseology, second language research, and semantics at the Department of Swedish, University of Gothenburg.

In the next section, the notion of constructions will be discussed. In section 3 we present a minor study of the treatment of cx in Swedish paper-dictionaries. The Swedish Cxn is presented (briefly) in section 4, followed by a presentation of possible methods to find new cx in section 5. Finally, in section 6, there is an outlook, addressing matters of cross-linguistic applicability.

2. Constructions

The type of cx mentioned above is far from being a small set of borderline cases that can simply be disregarded. On the contrary, semi-productive, partially schematic multi-word units are both numerous and common, arguably "of a comparable order of magnitude to the number of words" (Jackendoff, 2007: 57). Furthermore, these kinds of patterns have been shown to be highly problematic, e.g. in relation to L2 acquisition (cf. Ekberg, 2004; Wray, 2008; Prentice & Sköldberg, 2011) and language technology (LT; see Sag et al., 2002).

For the last few decades, however, the study of constructions is on the rise, due to the development of Construction Grammar (CxG; Fillmore et al., 1988; Goldberg, 1995; Hoffmann & Trousdale, 2013, and others) and other cx-oriented models. Furthermore, cx have also been gaining increased attention from some lexicalist perspectives, e.g., Head-Driven Phrase Structure Grammar (HPSG; Pollard & Sag, 1994), especially through the CxG-HPSG hybrid Sign-Based Construction Grammar (SBCG; Boas & Sag, 2012). Still, these approaches have mostly been applied to specific cx, or groups of such. To date, there are few, if any, large-scale constructional accounts.

Within CxG, cx are typically defined as conventionalized pairings of form and meaning/function. This definition can be compared to what in other linguistic contexts are called *signs* (cf. Saussure) or *symbolic units* (cf. Langacker). Linguistic patterns of any level, or combination of levels, from the most general to the most

specific, may be considered cx. Hence, instead of a sharp distinction between lexicon and grammar with a problematic grey area, one can see language as a network of cx along a continuum from extremely specific, lexically filled and fixed items to very general syntactic patterns (Fillmore et al., 1988; Lyngfelt & Sköldberg, forthcoming).

Goldberg (2006) points out that one can often identify someone as a non-native speaker of a given language:

> [...] because much of the phrasing used and combination of lexical choices are non-conventional, even if fully grammatical. It is in fact often the case that one particular formulation is much more conventional than another, even though both conform to the general grammatical patterns in a language
>
> (Goldberg, 2006: 54; cf. Pawley & Syder, 1983).

Goldberg exemplifies the above with conventionalized time expressions that are language specific and have to be acquired through input like other lexical items, since a learner who has never met them before has no means to build them from scratch based on his knowledge of the L2-system (Goldberg, 2006: 54f.). A Swedish example of such language-specific properties concerns what Fillmore (2008) calls *day-level temporal units*. Although time adverbials are usually expressed as PPs, in Swedish as in many other languages, this is not the case if the time is a date: *Hon åker (*på) 7 maj* 'She will leave on May 7th', as opposed to *Hon åker på måndag* 'She will leave on Monday'. In L2 Swedish, incorrect inclusion of the preposition is not uncommon: **Jag är född på 2 mars* 'I was born on March 2nd' (cf. Fillmore, 2008).

So far, the project has to a large extent focused on partially schematic cx, where at least one of the component parts is lexically fixed. Such cx are somewhat similar to fixed multi-word expressions and are fairly close to the lexical end of the cx continuum (cf. Lyngfelt & Forsberg, 2012).

Of general theoretical interest are argument structure cx, which concern matters of transitivity, voice, and event structure, and are at the heart of discussions on the relationship between grammar and lexicon. Argument structure is usually assumed to be determined by lexical valence, but there are good reasons to assume that syntactic constructions also play a role (Goldberg, 1995).

Consider, for instance, the (Swedish) *reflexive resultative* cx (Jansson, 2006; Lyngfelt, 2007), as in *äta sig mätt* 'eat oneself full', *springa sig varm* 'run oneself warm', and *byta sig ledig* 'swap oneself free' (cf. Hanks, 2008). Its basic structure is Verb Reflexive Result, where the result is typically expressed by an AP, and its meaning can be described as 'achieve result by V-ing'. (Hence, an expression like *känna sig trött* 'feel tired' is not an instance of this cx, since it does not mean 'get tired by feeling'.) This pattern is applicable to both transitive and intransitive verbs, even when it conflicts with the verb's lexical valence patterns. Notably, the reflexive object does not correspond to a typical object role; for example, the *sig* in *äta sig*

mätt does not denote what is eaten. Such cx raise theoretically interesting questions regarding to what extent argument structure is lexically or constructionally determined.

3. Constructions in Swedish Dictionaries

To what extent are these and other constructions accounted for in dictionaries? Studies of the treatment of constructions in Swedish, or Nordic, dictionaries are few in number. However, Farø & Lorentzen (2009) have shown that the coverage of partially schematic cx is not satisfactory in Danish dictionaries. Many dictionaries tend to favor colorful fixed phrases, e.g. idioms, at the expense of more anonymous cx with variable component slots. This is a problem, as many such cx are arguably more relevant for language learners than, for example, the idioms which by comparison are used quite rarely (Farø & Lorentzen, 2009). The authors also observe that the dictionaries have problems in reproducing the productivity of these structures.

We studied to what extent, and how, about ten partially schematic cx already included in the SweCxn are treated in dictionaries of contemporary Swedish. More precisely, we examined the following four comprehensive monolingual paper dictionaries:

- *Natur och Kulturs Stora Svenska ordbok* (2006)
- *Svensk ordbok utgiven av Svenska Akademien* (2009)
- *Bonniers svenska ordbok* (2010)
- *Svenskt språkbruk. Ordbok över konstruktioner och fraser* (2003).

The three books mentioned first are general dictionaries and the fourth is a phraseological dictionary. Our study supports the results by Farø & Lorentzen (2009; cf. Lyngfelt & Sköldberg, forthcoming). A typical example is the treatment of the already mentioned time expression [minuttal] i/över [timtal] '[minutes] to/past [hour]. In one of the general dictionaries, *Natur och Kulturs Stora svenska ordbok*, you find the cx in two places: in the articles *i* 'to' and *över* 'past'. However, the other general dictionaries only account for one of the corresponding time expressions, the one with *i* 'to'. Surprisingly, in *Svenskt språkbruk*, the phraseological dictionary, this frequently used conventionalized expression is not mentioned at all. There might, of course, be many underlying causes behind this scanty and inconsistent treatment of this particular cx, but one plausible explanation is that the only lexically fixed components (*i* and *över*) are highly frequent prepositions. Hence, the cx can be hard to discern in corpora. In addition, the cx typically occurs in speech and not in newspaper texts (on which Swedish dictionaries are primarily based). Moreover, in many texts, time information is usually given in another way: you write, e.g. *06.15* or *18.15* instead of *kvart över sex* 'a quarter past six'. But even if this cx were accounted for in a more adequate way in the dictionaries, from a user's point of view the cx would still be difficult to find in a paper dictionary, as preposition articles like these

are extensive and hard to grasp. In that sense, an e-dictionary with more search options evidently has many advantages.

The other example mentioned in the introduction, "i ADJEKTIV-aste laget" 'in ADJECTIVE-superlative the-measure', has two lexical parts, the preposition *i* and the noun *lag*. In all the dictionaries the cx is treated in the noun entry. Due to space limitations, in the following we present only one of these entries, the one from *Bonniers svenska ordbok* (2010):

(1) [1]l**ag**[3] (i många uttryck) *i längsta* (*största*, *minsta*, *kortaste*, etc.) *laget* nästan för lång osv. [...]
'[1]l**ag**[3] (in many expressions) *rather a bit long* (*big*, *little*, *short*, etc.) almost too long and so on [...]'

In the case of "i ADJEKTIV-aste laget", all lexicographers have tried to account for the productivity of the cx, but in different ways. In (1), the fact that this sense of *lag* appears in many expressions is commented. Similar comments, or other markers indicating the same thing, are also found in the other dictionaries. Furthermore, four different adjectives are given, i.e., *längsta* 'longest', *största* 'biggest', *minsta* 'smallest' and *kortaste* 'shortest' followed by an "etc." indicating that these adjectives can be replaced by others.

The word combinations in the dictionary examples are without doubt recurrent in the corpora at Språkbanken (of more than 1 billion tokens). Many also appear in the other dictionaries. Still, they are not totally representative of authentic language as all the adjectives refer to size. Other recurrent adjectives in the corpora are, e.g. *dyr* 'expensive', *tidig* 'early', *sen* 'late' and *tuff* 'tough' which are all missing in the dictionaries. However, in a traditional dictionary it is very difficult to give exhaustive information in this respect, as the productivity cannot be captured on a lexical basis.

Finally, only the first example in the dictionary article above, i.e. *i längsta laget* 'rather a bit long', is explicitly explained. The idea is that the user can figure out the meaning of the other variants by analogy. One of the intended user groups of this particular dictionary (L1-speakers) might be able to understand this information. However, for L2-learners on all levels, it can be a hard nut to crack.

To conclude, our study of the treatment of partially schematic cx in dictionaries of Swedish is limited, but it supports the results presented by Farø & Lorentzen (2009). Even if cx with specific lexical parts, such as "i ADJEKTIV-aste laget", to some extent are described in the dictionaries, many of them are missing. This also applies to dictionaries which normally account for many phraseological units, at least idioms. And even if all the dictionaries try to bring out the productivity of the cx, they cannot totally catch this characteristic feature due to the fact that they have lexical items as a starting point.

As shown by these examples, cx often combine features from several linguistic levels. They may be characterized by prosodic, morphological, lexical, syntactic, semantic, pragmatic features, in different combinations. How can such patterns be accounted for? Should cx of this type be described in dictionaries at all? Or do they "belong" to the grammars/grammarians? The questions bring to the fore an observation made by e.g. Hannesdóttir & Ralph (2010), who discuss the fact that lexicography and grammar description to a great extent are different activities. Patterns with both lexical and grammatical properties cannot be described in an adequate way as long as lexicography and grammar are kept strictly apart. Consequently, according to the authors, lexicographers and dictionary writers should cooperate more and jointly ensure that what is lacking in the one resource is covered in the other. Naturally an increased cooperation would be beneficial in many respects. However, such a solution is still based on a binary distinction between lexicon and grammar. Each linguistic phenomenon must be attributed to the one or the other – or perhaps both. In this paper we present a different approach, where the grammatical and lexical features are combined in the same description.

4. The Swedish Constructicon

The Swedish Constructicon (SweCxn) is a usage based database, where all cx descriptions are grounded in annotated corpus examples. At present, it consists of about 100 cx, still basically a pilot constructicon, but it is growing continually. Eventually, SweCxn is meant to be primarily a large-scale resource for linguistic research and language technology applications. In a longer perspective, the SweCxn should also be applicable in educational settings, not least for learners of Swedish as a second language. Today, the focus is on collecting the most essential linguistic information about a large number of cx often ignored by traditional lexicography and grammar, but the system is designed to be able to handle any kind of cx as the term is understood in Construction Grammar, including ordinary words, parts of speech, etc.

A typical example of cx currently in SweCxn is the so called *reflexiv resultativ* 'reflexive resultative' (cf. section 2 above), where the use of a reflexive pronoun adds a valency bound adjective phrase expressing the result of the action, as in *Han sprang sig varm* 'He ran himself warm' and *Kornet och havren får frysa sig mogen* 'The barley and the oats may freeze themselves ripe'. From a structural point of view, the cx consists of a verb, a reflexive pronoun and an adjective phrase. Seen as a whole, it is a multi-word verb with the reflexive pronoun as a fixed, construction-evoking element. The verb, the reflexive pronoun and the adjective phrase are parts of the cx itself, the subject is also important but it is not a part of the construction proper. The adjective phrase expresses the Result while the subject and the reflexive pronoun may be an Agent or a Theme (according to the system of semantic roles employed in SweCxn). This information is captured in the following way in the entry for *reflexiv resultativ* in SweCxn:

Name: *reflexiv_resultativ*
Category: vbm ('multi-word verb')
Structure: vb refl AP
Construction evoking element: refl
Internal construction elements:
role: name=Activity cat=vb
role: cx=refl name=Actor
role: cx=refl name=Theme
role: name=Result cat=NP
External construction elements:
role: name=Actor cat=NP
role: name=Theme cat=NP

The set of labels used for the category and the structure is quite large, since different cx require different granularity. An element may belong to a very general phrase type like XP or NP but also specific lexical items (possibly in a certain inflectional form), with NPdef etc., in between. Truly fixed elements are noted as construction-evoking elements, but it is also useful to list **common words** and word combinations merely typical for a cx (cf. *collostructural elements*, Stefanowitsch & Gries, 2003). The list for *reflexiv_resultativ* is {*äta* 'eat': *mätt* 'full'}, {*supa* 'drink': *full* 'drunk'}, {*skrika* 'scream': *hes* 'hoarse'} and *springa* 'run'.

Construction elements are defined by a list of feature value pairs. There is no set of features fitting all construction elements, so it is not meaningful to require all of them to have the same features defined. The format does not imply that all possible construction elements are instantiated, which is why the external element and the reflexive have two alternatives with different definitions.

Semantic roles are described in two ways resembling Goldberg's (1995) argument roles and participant roles. Argument roles are typically small sets of general roles useful for describing general semantic features whereas participant roles give a local description of the frames of specific (lexical) items. Agent is an argument role and Eater is a participant role. But the neat distinction between argument and participant roles becomes less clear when dealing with cx in the continuum between the purely grammatical and lexical. In practice, this means that the set of semantic roles needed for describing general features with sufficient precision becomes larger than what is needed for arguments in traditional syntax. The set of general roles employed in SweCxn consists of 33 primitive roles augmented by some modifications and a mechanism for combining roles, e.g. Agent-Source.

General roles are noted explicitly whenever appropriate, as shown above. They are also used as the default name for the construction element. Local, frame specific roles are assigned indirectly when a construction **evokes** a FrameNet frame, e.g. the entry for *reflexiv_resultativ* is declared to evoke the frame Causation_scenario.

The meaning of a construction and how the construction elements contribute to it are described in the **definition**, in the case of *reflexiv_resultativ*:

Definition: [Någon]$_{Actor}$ eller [något]$_{Theme}$ utför eller undergår [en aktion]$_{Activity}$ som leder (eller antas leda) till att [aktören]$_{Actor}$ / [temat]$_{Theme}$, uttryckt med reflexiv, uppnår ett [tillstånd]$_{Resultat}$.
'**Definition:** [Someone]$_{Actor}$ or [something]$_{Theme}$ performs or undergoes [an action]$_{Activity}$ which leads to (or is assumed to lead to) the [actor]$_{Actor}$ / the [theme]$_{Theme}$, expressed by a reflexive pronoun, reaching a [state]$_{Result}$'

The format for the definitions is inspired by ordinary dictionary type definitions but there are striking differences. One is that the definitions are annotated in almost the same way as the corpus examples included in the entry. The only difference is that the cx itself is delimited in the examples, as in:

[Kornet och havren]$_{Theme}$ får [[frysa]$_{Activity}$ [sig]$_{Theme}$ [mogen]$_{Result}$]$_{resultativ\_reflexiv}$
'[The barley and the oats]$_{Theme}$ may [[freeze]$_{Activity}$ [themselves]$_{Theme}$ [ripe]$_{Result}$]$_{resultativ\_reflexiv}$'

Another difference between definitions in dictionaries and in SweCxn is that one does not expect explicit information in a dictionary definition about how parts of the meaning are expressed, e.g. that the theme is expressed by a reflexive pronoun. But there are also deep similarities. One is that readability for humans gets a higher priority than tractability for computers. Another, not apparent from the definition of *reflexiv_resultativ*, is the use of dictionary type modifications as *typically*, *also* and *etc*. This makes it relatively easy to write definitions which are reasonably nuanced and easy to understand. The price is that further formalization will be required to make some information in the definitions useful for technical systems, but that is probably a price worth paying to facilitate the collection of the information in the first place.

But it is worth noting that SweCxn is a formally well defined system in most respects. All names of semantic roles, lexical units etc., are declared or defined either within SweCxn proper or imported in an orderly way from external resources, such as FrameNet or the lexical resource SALDO at Språkbanken. The cx are also ordered in an inheritance hierarchy so that more specialized cx, e.g. *jämförelse.likhet* [*comparison.equality*] and *jämförelse.olikhet* [*comparison.inequality*] inherit from the more general *jämförelse* [*comparison*] in order to increase consistency and maintainability.

5. Data and Methods

Since no comprehensive collection of cx descriptions has ever existed for Swedish, an important methodological question for the project is to discover those cx that have not been described before. To identify cx for SweCxn, we use a combination of methods, such as working from existing cx descriptions for Swedish and other

languages (section 5.1), applying LT tools to discover recurring patterns in texts (5.2), and extrapolating constructional information from dictionaries (5.3).

5.1 Digging where we stand

The natural starting point for SweCxn has been existing cx analyses, for Swedish and for other languages. These analyses include quite a few term papers by our own students, produced over the years in relation to CxG courses and earlier CxG projects. The typical CxG paper presents an in-depth analysis of a particular type of cx. From there, we can a) make simplified analyses to include in SweCxn, and b) trace other cx with related properties. On the basis of the initial, familiar cx, we have developed preliminary standards for SweCxn descriptions. It should be noted in this context that we always consult corpora before arriving at a SweCxn account, even when the cx in question has been described by others.

Cx descriptions for other languages provide a more indirect source of inspiration. Each cx in another language raises the question what more or less corresponding patterns exist in Swedish. However, cx are essentially language specific, and even when similar cx occur in different languages, they cannot be presumed to be identical. The SweCxn entries must always be based on Swedish data, but the foreign cx provide hypotheses to explore.

Of particular interest are cxn resources for other languages. There is a small cxn for English (Fillmore et al., 2012), and cxn projects are under way for Japanese (Ohara, 2012) and Brazilian Portuguese (Torrent et al., 2013). In this case, the cxn entries are not only a source of inspiration; we also wish to establish correspondences for future cross-linguistic cxn applications. Such applications, however, require compatible description formats for the cxn resources involved. We will return to this issue in the final section.

As a first step in this direction, we conducted an inventory of the entries in the English cxn at Berkeley (BCxn), investigating to what extent there are corresponding Swedish cx for each of them (Bäckström et al., 2013b). BCxn consists of 50 complete and 23 incomplete cx entries. Out of the 50 full cxn entries, we established 37 one-to-one correspondents. In five cases, one BCxn entry corresponds to two Swedish cx, whereas the remaining eight entries lack satisfactory matches in Swedish.

As might be expected, more general and abstract cx are typically among the closest cx equivalents, whereas more specific idioms tend to differ to a greater extent. Formal differences between corresponding cx typically concern grammatical markers for number, agreement, definiteness, etc., and relational expressions within the cx. For instance, consider the following pair of examples of corresponding Rate cx in English and Swedish:

(2) a. twice an hour

b. *två gånger i timmen*
'two times in hour-DEF'

As shown in (2), the denominator is headed by an article in English but by a preposition (*i* / 'in') in Swedish, and its complement is indefinite in English but definite in Swedish. (In addition, the word *twice* corresponds to a phrase, *två gånger* / 'two times', but this last difference is not a property of the respective rate cx per se.) Functionally, however, the two Rate cx are basically equivalent, although their distribution may differ somewhat. In summary, the comparison with BCxn both provided SweCxn with a set of cx entries and may serve as a first step towards multilingual constructicography.

5.2 Cx-candidates via corpora

One of the goals of SweCxn is to develop tools for automatic identification of constructions in authentic texts. This is a highly desirable research objective in itself, with potential uses in a number of LT applications. In addition, the same methods provide the project with a heuristic tool. By automatically extracting various kinds of regularities in texts, we may discover patterns that might otherwise have been overlooked. This especially concerns seemingly insignificant constructions that do not stand out against the context the way spectacular idioms do. The resulting findings are treated as cx candidates, a subset of which may be considered actual cx after manual evaluation (see Bäckström et al., 2013a).

The general setting for our experiment is the resource infrastructure of Språkbanken, a modular set of resources and tools in the form of web services for accessing, browsing, editing and automatically annotating resources. The two facets of the infrastructure most relevant for the present purposes are the corpus infrastructure Korp (Borin et al., 2012b) and the lexicon infrastructure Karp (Borin et al., 2012a).

The data source for the experiment is SUC 2.0, a balanced text corpus for Swedish consisting of 1.17M tokens that have been manually annotated with lemmas and MSDs (morpho-syntactic description). SUC was selected in order to avoid annotation errors confounding the experiment results, but the experiment can be (and has been) run on any of the more than hundred corpora of Språkbanken that have been automatically annotated with the same information.

The experiment is based on the work on StringNet (Tsao & Wible, 2009; Wible & Tsao, 2010, 2011), where the notion of *hybrid n-gram* plays a central role. A hybrid n-gram is a generalization of an n-gram where not only the word forms are included in the process, but also the information from the annotation layers. If we limit ourselves to lemmas and part-of-speech, which is the case for this experiment, then the 2-gram *Hur är* 'How is' would generate four cx candidates: *hur vara* 'how be', *hur VB* 'how VB', *HA vara* 'HA be', and *HA VB*.

Focusing on the discovery of partially schematic constructions, we discarded all candidates that are fully schematic or fully lexical, i.e., consisting of only PoS tags (e.g., *HA VB*) or lemmas (e.g., *hur vara* 'how be'). Moreover, we removed all hybrid n-grams containing punctuation marks and/or words marked as foreign. They are not necessarily uninteresting, but since they did introduce a lot of noise in the candidate list, we decided to remove them. For SUC 2.0 with 2-, 3- and 4-grams we ended up with 16M hybrid n-grams of which 8.8M were unique.

The next step was to rank all hybrid n-grams, which can be done with a wide range of association measures. We have followed StringNet in using point-wise mutual information (PMI). PMI has a known shortcoming in these kinds of experiments – it has a preference for the low-frequency items – which can be remedied by multiplying PMI with the absolute frequency. This does not solve another problem, however, which is boilerplate text, e.g., "For subscription enquiries e-mail:...". But with a small modification, instead of counting hybrid n-grams, we count UIF (unique instance frequency), which is the number of unique n-grams underlying the target hybrid n-gram, we can counteract that problem too.

There was still one more problem that needed to be solved: since the bulk of the hybrid n-grams are subsets of other hybrid n-grams, we first arrived at a ranking list with massive redundancy. This was solved, in the same spirit as StringNet's vertical/horizontal pruning (Tsao & Wible, 2009; Wible & Tsao, 2010), by removing all hybrid n-grams that were subsets of other hybrid n-grams with a higher PMI-UIF. A hybrid n-gram is considered a subset of another if it occurs as a subsequence that is either equal or consisting of non-conflicting items sharing the same part-of-speech; e.g., *vara*$_{\text{VB}}$ is considered equal to *VB*.

Some sample candidates are given in Figure 1. The hybrid n-grams are linked to the Korp interface to enable inspection of their instances in the corpus. We also see the most frequent instance, followed by the absolute frequency, relative frequency, and the PMI-UIF.

vara$_{\text{VB}}$ ute$_{\text{AB}}$ och$_{\text{KN}}$ VB	*är ute och letar (3)*	15	0.93	52.24
vara$_{\text{VB}}$ JJ för$_{\text{PP}}$ att$_{\text{IE}}$	*är viktiga för att (2)*	26	1.61	52.83
stänga$_{\text{VB}}$ av$_{\text{PL}}$ NN	*stängt av motorn (1)*	11	0.68	52.25

Figure 1. Some example hybrid n-grams from SUC 2.0 ranked by PMI-UIF

The candidate lists are accessible from here: <http://spraakbanken.gu.se/eng/resource/konstruktikon/candidates>. Here you will find other materials as well that have been annotated automatically using the Korp pipeline (see 5.3 below).

The construction candidate list makes it possible to go through a large amount of examples quickly, since every hybrid n-gram is directly linked to the instances in the

corpus. However, it was a difficult task to draw the line between relevant and non-relevant constructions and this is still an ongoing matter of discussion in the project group. Of the 2500 items included in the list, 50 constructions were decided to be relevant construction candidates according to our criteria, i.e., that they are partially schematic and productive multiword units that are "too general to be attributed to individual words but too specific to be considered general rules" (Lyngfelt et al., 2012).

The final list of 50 relevant constructions was extracted in several steps. First, one project member went through the whole list extracting a list of 143 interesting candidates (approximately a day's work). This list was then, in consultation with the other members of the project group, gradually reduced and the final result of this process was, as mentioned above, 50 cx that were found relevant for entries in the SweCxn. As the main goal was to discover cx that are difficult to find with other methods, the result of 50 is not the whole story: a cx candidate can also inspire descriptions of other similar cx, which is a question of the researchers' capacity for creative thinking at a given moment in time.

5.3 Cx-candidates from general dictionaries

Currently we are also exploring the possibility of finding relevant cx-candidates within the articles in Swedish definition dictionaries. First of all we are interested in partly schematic patterns not so emphasized but rather indicated by comments like "in many expressions" (cf. section 3 above). Of course, this kind of usage marker is more easily found in e-dictionaries. Unfortunately, there are very few modern electronic definition dictionaries of high quality for Swedish. As a matter of fact, the existing ones are just e-versions of older paper dictionaries, which now have been subject to extensive revisions. Unfortunately, these revised versions are not published electronically (cf. i.e. NEO from 1995–1996 online with the printed SO from 2009; see below).

However, in the SweCxn project we have access to the whole database of the two-volume paper dictionary of Swedish published by the Swedish Academy (2009; henceforward SO). The dictionary, comprising about 65,000 lemmas, is the most comprehensive monolingual dictionary of contemporary Swedish that there is. By advanced search options in the database, we can extract information on different kinds of relatively anonymous word combinations indicated in the microstructure.

For example, the marker "i uttryck" 'in expression(s)' is used about 700 times within the SO articles. One cx observed by this method is "[X efter X]" '[X after X]', i.e. a certain lexical item appears just before and after the preposition *efter* 'after'. SO have tried to capture the cx as a subordinate sense of the word *efter* ('after') in the following way:

(3) **efter** prep. (...) **[**äv. i uttr. för upprepning] *dag efter dag*; *mil efter mil*; (...)
'**after** prep. (...) [also in expressions of repetition] *day after day*; *mile after mile*; (...)'

In the dictionary only two examples are given, including the nouns *dag* ('day') and *mil* ('mile'). Furthermore, the information on the semantic and pragmatic characteristics of the cx is very scanty. However, by searching in the corpora of Språkbanken, you get more data on this structure. In the texts the cx is used in a frequent and productive way. The repeated word may be a noun (as in the dictionary examples), but it can also be a numeral (*en*, *ett*):

(4) *... hon dricker glas efter glas*
'... she drinks glass after glass'

(5) *I brev efter brev utbytte de tankar om kriget*
'In letter after letter they were exchanging their thoughts about the war'

(6) *De kom allesammans, en efter en*
'They all came, one by one'

(7) *Också träden försvann, ett efter ett*
'All the trees disappeared, one by one'

Many of the hits (here from a corpus of modern novels) can be paraphrased by 'many X in succession', emphasizing the repetition. As indicated by the examples, the cx also infers some kind of process. If the repeated word is a noun referring to time, the cx also expresses extension in time and some kind of continuity. This is the case with *dag efter dag* 'day after day' in SO. Other typical examples from the corpora are *kväll efter kväll* 'evening after evening', *natt efter natt* 'night after night' and *år efter år* 'year after year'.

In other words, well hidden in the SO-articles you find several partially schematic patterns – like "[X efter X]" '[X after X]' – that could be emphasized and accounted for in a more exhaustive way. In SweCxn this problem can be solved. In that sense, the SweCxn can serve an important purpose towards a more detailed description of different kinds of Swedish word combinations.

In the project we also have access to the about 90,000 editorial examples found in the SO articles. One important function of the examples is, of course, to clarify the meaning(s) of the lemmas in the dictionary. But they also reveal typical usage of the lemmas by specifying constructions and collocations (Svensén, 2009: 285). The examples have been tokenized, lemmatized and PoS-tagged and constitute a corpus of its own in Språkbanken. Using the method described in section 5.2 above, we have also extracted SweCxn candidates from that corpus. On the list one can find, for example, the structure [*var*$_{DT}$ RO NN] which is typically realized in the following ways in the corpus:

(8) *var tjugonde minut* 'every twenty minutes'

(9) *var tredje timme* 'every three hours'

(10) *vart fjärde år* 'every four years'.

In other words, the method reveals another highly productive cx, which also is a challenge to language learners. First of all, as hinted by the examples, the noun can be composed by any time expression. Secondly, the cx includes a variable ordinal number. Thirdly, the pronoun *var* 'every', constituting the only lexically-filled component of the cx, has to agree in gender with the noun. And, once again, the cx is an ordeal to lexicographers; it is hard to place and render adequately in the dictionary as the only lexically-filled component is the unstressed pronoun.

6. Outlook

SweCxn is a resource under development, initially designed to suit the needs of linguistic research and LT application. In a longer perspective, it is meant to also support (second) language pedagogy and eventually be presented in a format adapted to a wider audience. Furthermore, in collaboration with the cxn projects of other languages, we are working towards cross-linguistic applicability.

The latter endeavor is probably best characterized as multilingual constructicography. It differs from lexicography in that a cxn must also account for the formal structure of a cx and its constituents. What is expressed by syntax or morphology is highly relevant, whether a certain construction element is an NP or a PP, whether NPs are definite or indefinite, if any particular agreement patterns apply, etc. Such features are language-specific, but must be represented in a way in which the relevant information may be linked across languages.

Since all existing cxn resources are developed in relation to a FrameNet of that language, it is desirable to make the two types of resource compatible from a cross-linguistic perspective as well. In FrameNet, which is essentially a lexicographic resource, all cross-linguistic relations are established through the frames. These are semantic units, which have been fairly successfully applied to different languages, since language-specific idiosyncrasies are instead attributed to the lexical units instantiating the frames in each language (cf., however, Pado, 2007; Friberg Heppin & Toporowska Gronostaj, 2012).

For cx with a meaning roughly equivalent to a frame, the same strategy is a viable option, provided that information about cx internal structure is added; but not all cx correspond to frames. Alternatively, as mentioned in section 5.1 above, some cx might be treated as direct equivalents in different languages, but clearly not all of them: especially not when languages less similar than English and Swedish are taken into account. Hence, a cross-linguistically applicable format for cx descriptions is required. Devising such a format will be a challenge for future constructicon development.

Awaiting that, each cxn should be nonetheless useful as a monolingual resource. SweCxn is still small, compared to a comprehensive dictionary, but it already contains a substantial number of linguistic patterns that would be hard to account for from a lexical viewpoint. Some of these cx are of course relevant for lexicography as well – to the extent that they are lexically entrenched. Their productivity, however, is beyond any resource restricted to lexical entries. An appealing future development would be to integrate the constructicon with an e-dictionary, where the possible entries are no longer limited to lexical units. In such a resource, one could navigate from grammatical constructions to the lexical entries that instantiate them and vice versa. Ideally, a user should only have to enter an expression, and the e-resource would be able to identify the constructional pattern to which it corresponds.

7. Acknowledgements

The research presented here was supported by the Swedish Research Council (grant agreement 2010-6013), by the Bank of Sweden Tercentenary Foundation (grant agreement P12-0076:1), by the University of Gothenburg through its support of the Centre for Language Technology and of Språkbanken, and by Swedish Academy Fellowships for Benjamin Lyngfelt and Emma Sköldberg, sponsored by the Knut and Alice Wallenberg Foundation.

8. References

Bäckström, L., Borin, L., Forsberg, M., Lyngfelt, B., Prentice, J. & Sköldberg, E. (2013a). Automatic identification of construction candidates for a Swedish constructicon. *Proceedings of the workshop on lexical semantic resources for NLP at NODALIDA 2013*. Linköping Electronic Conference Proceedings 88, pp. 2-11.

Bäckström, L., Lyngfelt, B. & Sköldberg, E. (2013b). Constructions in constrast. Approaching Swedish correspondents to the entries in the Berkeley FrameNet Constructicon. *International FrameNet Workshop 2013* (IFNW-13), Berkeley, CA.

Boas, H. C. & Sag, I. A. (eds.) (2012). *Sign-Based Construction Grammar*. Stanford: CSLI Publications.

Bonniers svenska ordbok (2010). (10 ed.) Stockholm: Bonniers.

Borin, L., Forsberg, M., Olsson, L.-J., & Uppström, J. (2012a). The open lexical infrastructure of Språkbanken. In *Proceedings of LREC 2012*. Istanbul. ELRA, pp. 3598-3602.

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

Ekberg, L. (2004). Grammatik och lexikon i svenska som andraspråk på nästan infödd nivå. I: Hyltenstam, K. & Lindberg, I. (eds.), *Svenska som andraspråk* –

i forskning, undervisning och samhälle. Lund: Studentlitteratur, pp. 259-276.

Farø, K. & Lorentzen, H. (2009). De oversete og mishandlede ordforbindelser – hvilke, hvor og hvorfor? *LexicoNordica* 16, pp. 75-101.

Fillmore, C. (2008). Border Conflicts: FrameNet Meets Construction Grammar. In Bernal, E. & DeCesaris, J. (eds.) *Proceedings of the XIII EURALEX International Congress*. Barcelona: Universitat Pompeu Fabra, pp. 49-68.

Fillmore, C., Kay, P. & O'Connor, M. C. (1988). Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone. *Language* 64, pp. 501-538.

Fillmore, C., Lee-Goldman, R. & Rhomieux, R. (2012). The FrameNet Constructicon I: Boas, H. & Sag, I. (eds.) *Sign-Based Construction Grammar*. Stanford: CSLI, pp. 309-372.

Friberg Heppin, K. & Toporowska Gronostaj, M. (2012). The Rocky Road towards a Swedish FrameNet – Creating SweFN. In *Proceedings of LREC 2012*, Istanbul.

Goldberg, A. (1995). *Constructions. A Construction Grammar Approach to Argument Structure*. Chicago/ London: University of Chicago Press.

Goldberg, A. (2006). *Constructions at work. The Nature of Generalization in Language*. Oxford: Oxford University Press.

Hanks, P. (2008). The lexicographical legacy of John Sinclair. In *International Journal of Lexicography* 21(3), pp. 219-229.

Hannesdóttir, A. H. & Ralph, B. (2010). Explicit och implicit information i tvåspråkig lexikografi. I: Lönnroth, H. & Nikula, K. (eds.), *Nordiska studier i lexikografi* 10. Tammerfors, pp.150-163.

Hoffmann, T. & Trousdale, G. (eds.) (2013): *The Oxford Handbook of Construction Grammar*. Oxford: OUP.

Jackendoff, R. (2007). *Language, Consciousness, Culture: Essays on Mental Structure*. Cambridge: MA: MIT Press.

Jansson, H. (2006). *Har du ölat dig odödlig? En undersökning av resultativkonstruktioner i svenskan* (MISS 57). Inst. f. svenska språket, Göteborgs universitet.

Lyngfelt, B. (2007). Mellan polerna. Reflexiv- och deponens-konstruktioner i svenskan. *Språk och stil* NF 17, pp. 86-134.

Lyngfelt, B., Borin, L. Forsberg, M., Prentice, J., Rydstedt, R., Sköldberg, E. & Tingsell, S. (2012). Adding a Constructicon to the Swedish resource network of Språkbanken. *Proceedings of KONVENS 2012 (LexSem 2012 workshop)*, Wien, pp. 452-461.

<http://www.oegai.at/konvens2012/proceedings/66_lyngfelt12w/>.

Lyngfelt, B. & Forsberg, M. (2012). *Ett svenskt konstruktikon. Utgångspunkter och preliminära ramar*. (GU-ISS 2012-02) Inst. f. svenska språket, Göteborgs

universitet. <http://hdl.handle.net/2077/29198>.

Lyngfelt, B. & Sköldberg, E. (forthcoming). Lexikon och konstruktikon – ett konstruktionsgrammatiskt perspektiv på lexikografi. *LexicoNordica* 20.

Natur och Kulturs Stora Svenska Ordbok (2006). Stockholm: Natur och Kultur.

Ohara K. (2012). Japanese FrameNet: Toward constructicon building for Japanese. *Seventh International Conference on Construction Grammar* (ICCG-7), Seoul, Korea

NEO = *Nationalencyklopedins ordbok* (1995-96). Höganäs: Bra böcker.

Pado, S. (2007). Translational Equivalence and Cross-lingual Parallelism: The Case of FrameNet Frames. *Proceedings of the NODALIDA Workshop on Building Frame Semantics Resources for Scandinavian and Baltic Languages*. Tartu, Estonia.

Pawley, A. & Syder, F. H. (1983). Two puzzles for linguistic theory: nativelike selection & nativelike fluency. In Richards, J. & Smith, R. (eds.) *Language and communication*. London: Longman, pp. 191-221.

Pollard, C, & Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar* Chicago: University of Chicago Press and Stanford: CSLI Publications.

Prentice, J. & Sköldberg, E. (2011). Figurative word combinations in texts written by adolescents in multilingual school environments. In Källström, R. & Lindberg, I. (eds.) *Young urban Swedish. Variation and change in multilingual settings*. University of Gothenburg: Dept. of Swedish, pp. 195-217.

Sag, I., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multi-word expressions: A pain in the neck for NLP. I: Gelbukh, A. (ed.) *Computational Linguistics and Intelligent Text Processing* (Proceedings of CICLING-2002). Berlin: Springer, pp. 1-15.

SO = *Svensk ordbok utgiven av Svenska Akademien* (2009). Stockholm: Norstedts.

Språkbanken < http://spraakbanken.gu.se/>.

Stefanowitsch, A. & Gries S. (2003). Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8.2, pp. 209-43.

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

Svenskt språkbruk. Ordbok över konstruktioner och fraser (2003). Utarbetad av Svenska språknämnden. Stockholm: Norstedts Ordbok.

SweCxn = *Svenskt konstruktikon*.

<http://spraakbanken.gu.se/swe/resurs/konstruktikon>.

Torrent, T., Lage, L. Sampaio, T., Tavares, T. & Matos, E. (2013). Revisiting Border

Conflicts between FrameNet and Construction Grammar: annotation policies for the Brazilian Portuguese Constructicon. *International FrameNet Workshop 2013* (IFNW-13), Berkeley, CA.

Tsao, N.-L. & Wible, D. (2009). A method for unsupervised broad-coverage lexical error detection and correction. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*. Boulder. ACL, pp. 51-54.

Wray, A. (2008). *Formulaic language: pushing the boundaries.* Oxford: OUP.

Wible, D. & Tsao, N.-L. (2010). StringNet as a computational resource for discovering and investigating linguistic constructions. In *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics*. Los Angeles. ACL, pp. 25-31.

Wible, D. & Tsao, N.-L. (2011). The StringNet lexico-grammatical knowledgebase and its applications. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, Portland. ACL, pp. 128-130.