Abstract

This paper presents the integrated Swedish resource network of Språkbanken in general, and its latest addition – a constructicon – in particular. The constructicon, which is still in its early stages, is a collection of (partially) schematic multi-word units, constructions, developed as an addition to the Swedish FrameNet (SweFN). SweFN and the constructicon are integrated with other parts of Språkbanken, both lexical resources and corpora, through the lexical resource SALDO. In most respects, the constructicon is modeled on its English counterpart in Berkeley, and, thus, following the FrameNet format. The most striking differences are the inclusion of so-called collostructional elements and the treatment of semantic roles, which are defined globally instead of locally as in FrameNet. Incorporating subprojects such as developing methods for automatic identification of constructions in authentic text on the one hand, and accounting for constructions problematic for L2 acquisition on the other, the approach is highly cross-disciplinary in nature, combining various theoretical linguistic perspectives on construction grammar with language technology, lexicography, and L2 research.

1 Introduction

Large-scale linguistic resources typically consist of a lexicon and/or a grammar, and so do the linguistic components in language technology (LT) applications. Lexical resources mainly account for words, whereas grammars focus on general linguistic rules. Arguably, this holds both for knowledge-driven and data-driven language processing, since what counts as a “word” (or “token”) is determined a priori in both cases. Consequently, patterns that are too general to be attributed to individual words but too specific to be considered general rules are peripheral from both perspectives and hence have tended to be neglected. Such constructions are not, however, peripheral to language, and neither are they a trivial phenomenon that can simply be disregarded. On the contrary, semi-productive, partially schematic multi-word units are highly problematic for language technology (Sag et al., 2002), L2 acquisition (Prentice and Sköldberg, 2011), and, given that idiosyncrasies are typically attributed to the lexicon, lexicography. They are also quite common (cf. e.g., Jackendoff 1997, 156). Accordingly, constructions have received more attention in recent years, but resources with large-scale coverage are still lacking.

In response to this situation, we are currently building a Swedish constructicon, a collection of (partially) schematic multi-word units, based on principles of Construction Grammar and developed as an addition to the Swedish FrameNet (SweFN). It will be integrated with other resources in Språkbanken by linked lexical entries. The constructicon project is a collaboration involving experts on (construction) grammar, language technology, lexicography, phraseology, second language research, and semantics.

The resource environment of Språkbanken is treated in section 2, and the work on integrating the resources in section 3. Constructions and
Construction Grammar are introduced in section 4 and the Swedish constructicon is presented in section 5, followed by an outlook in section 6.

2 Språkbanken

Språkbanken (the Swedish Language Bank)\(^1\) is a research and development unit at the University of Gothenburg, which was established with government funding already in 1975 as a national center for research on Swedish language resources, in particular corpora and lexical resources. The main focus of Språkbanken’s present-day activities is in the development and refinement of language resources and LT tools, and their application as research and teaching tools in various fields outside LT itself – several areas of linguistics: descriptive, typological, historical and genetic linguistics (e.g., Saxena and Borin 2011; Rama and Borin 2011), Swedish as a second language (e.g., Johansson Kokkinakis and Magnusson 2011; Voldina and Johansson Kokkinakis 2012), computer-assisted language learning, text complexity and lexical semantics (e.g., Borin 2012); other humanities disciplines: comparative literature (e.g., Borin and Kokkinakis 2010; Oelke et al. 2012) and history (e.g. Borin et al. 2011); and medicine and medical informatics (e.g., Kokkinakis 2012; Heppin 2011) – i.e., activities that can be broadly characterized as LT-based eScience.

Språkbanken’s LT research activities and in-house LT tools are characterized by a strong reliance on linguistic knowledge encoded in rich and varied lexical resources. The present focus is on the creation of a highly interlinked resource infrastructure informed by current work on LT resource standardization (e.g., in CLARIN, ISO TC37/SC4, and META-SHARE), as well as by work on linked open data (see, e.g., Chiarcos et al. 2012). This is the Swedish FrameNet++ project described in the next section.

3 Swedish FrameNet++

The goal of the Swedish FrameNet++ project (Borin et al., 2010a) is to create a large integrated lexical resource for Swedish – so far lacking – to be used as a basic infrastructural component in Swedish LT research and in the development of LT applications for Swedish. The specific objectives of the project are

- integrating a number of existing free lexical resources into a unified lexical resource network;
- creating a full-scale Swedish FrameNet with 50,000 lexical units;
- developing methodologies and workflows which make maximal use of LT and other tools in order to minimize the human effort needed in the work.

3.1 The lexical resource network

The lexical resource network has one primary lexical resource, a pivot, to which all other resources are linked. This is SALDO\(^2\) (Borin and Forsberg, 2009), a large (123K entries and 1.8M wordforms), freely available morphological and lexical-semantic lexicon for modern Swedish. It has been selected as the pivot partly because of its size and quality, but also because its form and sense units have been assigned persistent identifiers (PIDs) to which the lexical information in other resources are linked.

The standard scenario for a new resource to be integrated into the network is to (partially) link its entries to the sense PIDs of SALDO. This typically has the effect that the ambiguity of a resource becomes explicit: the bulk of the resources associate lexical information to PoS-tagged baseforms, information not always valid for all senses of that baseform. This is natural since most of the resources have initially been created for human consumption, and a human can usually deal with this kind of underspecification without problem. Some of these ambiguities can be resolved automatically – especially if information from several resources are combined – but in the end, manual work is required for complete disambiguation.

The network also includes historical lexical resources (Borin et al., 2010b; Borin and Forsberg, 2011), where the starting point is four digitized paper dictionaries: one 19th century dictionary, and three Old Swedish dictionaries. To make these dictionaries usable in a language technology setting, they need morphological information, a work that has been begun in the CON-
PLISIT project for 19th century Swedish (Borin et al., 2011) and in a pilot project for Old Swedish (Borin and Forsberg, 2008).

Linking SALDO to the historical resources is naturally a much more complex task than linking it to the modern resources, especially when moving further back in time. The hope is that a successful (but possibly partial) linking introduces the possibility to project the modern lexical-semantic relations onto the historical resources, so that, e.g., a wordnet-like resource for Old Swedish becomes available for use.

3.2 Swedish FrameNet

The Swedish FrameNet builds on the Berkeley FrameNet (Baker et al., 1998; Ruppenhofer et al., 2010), using its frame inventory and accompanying semantic roles. The current size of the Swedish FrameNet is 632 frames, 22,548 lexical units (+1,582 suggested lexical units), and 3,662 annotated examples, which may be compared with the size of Berkeley FrameNet with 1,159 frames, 12,601 lexical units, and 193,862 annotated examples. Some new frames have been created, but none of them are motivated by differences between English and Swedish, and they could just as well have been present in the Berkeley FrameNet.

3.3 Methodology development

The methodological work is conducted within the lexical infrastructure of Språkbanken, described by Borin et al. (2012b). Some of the features of the infrastructure are: daily publication of the resources, both through search interfaces and for downloading; a strong connection to the corpus infrastructure (Borin et al., 2012c); formal test protocols; statistics; and change history.

An important methodological task is the development of automatic methods for locating good corpus examples for the Swedish FrameNet. The task has been explored by, e.g., Kilgarriff et al. (2008) and Didakowski et al. (2012), but the notion of what constitutes a good example is still under active research. An important step has been taken by Borin et al. (2012a), where a tool has been developed that enables ranking of corpus examples based not only on a (tentative) measure of goodness but also on diversity (ideally, the examples should cover the full usage range of a linguistic item).

4 Constructions

Language consists to a quite large extent of semi-general linguistic patterns, neither general rules of grammar nor lexically specific idiosyncracies. Such patterns may be called constructions (cx). Peripheral from the view-point of grammar as well as lexicon, they have a long history of being neglected, despite being both numerous and common. 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, 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 and Sag 1994), especially through the CxG-HPSG hybrid Sign-Based Construction Grammar (SBCG; Sag 2010; Boas and Sag to appear). 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.

Cx are typically defined as conventionalized pairings of form and meaning/function. Hence, linguistic patterns of any level, or combination of levels, from the most general to the most specific, may be considered cx and therefore relevant for a constructicon. Since our goal is a constructicon of wide applicability we do not wish to exclude any types of cx beforehand; it may well turn out that many relevant cx are of types that have been previously overlooked. Indeed, one of the expected benefits of this project is coverage of cx not yet accounted for. On the other hand, we do not have infinite resources. We will therefore focus on semi-general cx in the borderland between grammar and lexicon, since this is where better empirical coverage is most sorely needed.

There are also some cx types that we are particularly interested in. These include cx of relevance for L2 acquisition, e.g., date expressions, which can display surprising complexities and idiosyncrasies (Karttunen et al., 1996). Although time adverbials are usually expressed as PPs in Swedish, this is not the case if the time is a date: *Hon åker (*på) 7 maj ‘She will leave on May 7th’, as op-
posed 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’ (Prentice, 2011).

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 varn ‘run oneself warm’, and byta sig ledig ‘swap oneself free’. Its basic structure is Verb Reflexive Result, where the result is typically expressed by an AP, and its meaning roughly ‘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 inherent valence restrictions. Notably, in the case of transitive verbs, the reflexive object does not correspond to the object role typically associated with the verb; 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.

From a structural perspective, a cx type of high priority are so-called partially schematic idioms (cf. Fried to appear; Lyngfelt and Forsberg 2012), i.e., cx where some parts are fixed and some parts are variable. Typical examples are conventionalized time expressions like [min- uttal] ilöver [timtal] ‘[minutes] to/past [hour]’ and semi-prefab phrases such as i ADJECTIVE-aste laget ‘of ADJECTIVE-superlative 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’, i senaste laget ‘at the last moment’. These cx are somewhat similar to fixed multi-word expressions and are fairly close to the lexical end of the cx continuum. They should be easier to identify automatically than fully schematic cx, and are therefore a natural initial target for the development of LT tools. Also, these cx are the ones closest at hand for integration into lexical resources.

Parallel to Construction Grammar, Fillmore (1982) and associates have also developed Frame Semantics, which in turn constitutes the base for the FrameNet resource (Baker et al., 1998; Ruppenhofer et al., 2010). Frame Semantics and FrameNet treat meaning from a situation-based perspective and put more emphasis on semantic roles and other cx-related features than other lexical resources usually do. By its historical and theoretical connections to Construction Grammar, FrameNet is well suited for inclusion of constructional patterns. There is also a growing appreciation for the need to do so. Accordingly, an English constructicon is being developed as an addition to the Berkeley FrameNet (Fillmore, 2008; Fillmore et al., to appear). In a similar fashion, the Swedish constructicon will be an extension of SweFN (Lyngfelt and Forsberg, 2012). Furthermore, there are plans to add constructicons to the Japanese and the Brazilian Portuguese FrameNet.

5 The Swedish constructicon

The Swedish constructicon is still in its early stages of development, so far numbering only a few sample cx, but it is growing and getting more refined by the day. Its format is for the most part modeled on Berkeley’s English constructicon, and thus on FrameNet. The core units in a constructicon, however, are not frames but cx; and instead of frame elements, there are cx elements, i.e. syntactic constituents.

As in the Berkeley constructicon, the cx are presented with definitions in free text, schematic structural descriptions, definitions of cx elements, and annotated examples. We try to keep the analyses simple, to make the descriptions accessible and reduce the labor required. This goes against common practice in linguistic research, where depth and detail usually get higher priority than simplicity. In the words of Langacker (1991, 548), “the meaning of linguistic expressions cannot be characterized by means of short, dictionary type definitions”.

While Langacker is of course right about linguistic meaning being complex and multi-faceted,
## reflexiv_resultativ

<table>
<thead>
<tr>
<th>type</th>
<th>Cx</th>
</tr>
</thead>
<tbody>
<tr>
<td>category</td>
<td>vbm</td>
</tr>
</tbody>
</table>

**evokes** Causation_scenario

**definition** | ![Image](image.png)
|---|

**structure** | vb refi AP |
| cee | refi |

**coll** | {sta^1 : mätt^1} (supa^1 : nul^2) (skriva^1 : neg^1) sprenna^1 |

### internal construction elements
- role: name=Activity cat=vb
- role: cx=refi name=Actor
- role: cx=refi name=Theme
- role: name=Result cat=AP

### external construction elements
- role: name=Actor cat=NP
- role: name=Theme cat=NP

### examples
- ![Example 1](image.png)
- ![Example 2](image.png)


**reference**
- http://hdl.handle.net/2077/19000

---

Figure 1: The reflexive-resultative construction

---

the approximations presented in dictionaries have after all turned out to be quite useful in many respects. Our expectation is that a corresponding level of complexity will work for a constructicon as well. More detailed analyses are both space and especially time consuming and therefore difficult to conduct on a large scale. Hence, simplicity is a main priority. Still, it is necessary to add some complexity compared to lexical definitions, since descriptions of syntactic cx also must contain constituent structure. Therefore, initially the core of the cx descriptions consists of a simple structural sketch and a free text definition of dictionary type. The intention is to refine and extend the description formalism incrementally as needed to reflect the complexity and variability of constructions as we come across them in our work, while still striving to keep it as simple as possible, not least in order for it to be usable in LT applications.

An example constructicon entry, the reflexive resultative cx (cf. section 4), as represented in the current preliminary interface, is shown in figure 1. Like in other FrameNet based models, semantic roles and other cx elements are explicitly included in the definitions and annotated in the accompanying examples. The treatment of the roles themselves, however, is somewhat different. In FrameNet, semantic roles are locally defined for each frame, which has led to 125 different defi-
nitions of Agent, for example. Instead, we define
roles etc. globally – generalizing where we can
and maintaining specific roles where we must, but
the same role label always has the same definition.
Accordingly, cx elements are represented as sets
of features, where each feature is a database entry
of its own. In addition to the linguistic value of a
consistent treatment of semantic roles, global role
definitions will be helpful to LT applications (cf.
Johansson 2012). The treatment of semantic roles
is therefore an important subproject, based on the
model in Rydstedt (2012).

As in the Berkeley constructicon, fixed cx ele-
ments are specifically indexed (cee, construction
evoking element). In addition, the Swedish con-
structicon also lists collostructional elements, i.e.,
words that are not fixed, but significantly frequent
in a certain cx (cf. Stefanowitsch and Gries 2003),
coll in figure 1. Such information is useful for
LT, and likely also for educational purposes. For
reasons of time, this will not be based on full-
fledged collostructional analyses; we will simply
note salient common elements.

To enable cross-linguistic compatibility, much
of the English terminology from FrameNet and
the Berkeley constructicon is maintained. How-
ever, for constructicons to be cross-linguistically
useful, additional information is required. In
FrameNet, the frames serve as a technical lingua
franca. Lexical units are language specific, but
by assuming the same frames across languages,
FrameNet resources for different languages may
still be connected. In a constructicon, on the other
hand, the central units are not frames but cx,
and cx are typically language specific. Therefore,
some form of common metalanguage is needed.

Initially, however, the constructicon is primar-
ily designed for Swedish users. Hence, cx names
and definitions are all in Swedish. This makes
things easier for us, but the main reason is that
the descriptions are eventually intended to be us-
able in an interface for non-linguists. An interna-
tional representation may be added later on and
should reasonably be developed in collaboration
with the other constructicon projects under way.
Awaiting that, we indicate the frame closest to the
meaning of a certain cx, whenever applicable, as
an approximation (cf. evokes in figure 1). A cx
with causative meaning, such as the reflexive re-
sultative, may thus be associated with a frame like
Causation_scenario.

The constructicon is usage-based, i.e., cx are
identified and characterized according to authen-
tic usage, as perceived from corpora. Such studies
will chiefly be conducted using Korp, the main
corpus tool of Språkbanken, where several cor-
pora of different types are integrated and search-
able by a common interface (Borin et al., 2012c).
Korp gives access to around 1 billion running
words (and growing), annotated for lexical unit,
part of speech, morphosyntactic category, tex-
tual properties etc. This annotation is a vital fea-
ture for this project, since a cx may be defined
by constraints on different levels: word, word-
form, part of speech, morphosyntactic category,
grammatical function, information structure etc.
(as illustrated by the examples in the preceding
paragraphs). The Korp interface can also present
statistic information about grammatical and lex-
ical contexts, as well as text type.

Up until now, we have mainly relied on lin-
guistic methods for the identification of cx, but
we will also develop tools to identify cx auto-
matically. As a first step, we will explore meth-
ods for the identification of unknown cx, or rather
cx candidates. For this purpose, StringNet (Wible
and Tsao, 2010) is one of many possible research
directions. StringNet identifies recurring n-gram
patterns of two or more units, where every unit
is classified on three levels – word form, lemma,
and grammatical category – potentially revealing
patterns of lexical units and form classes in com-
bination. Narrowing down the search by combin-
ing the result of StringNet with methods for auto-
matic morphology induction (Hammarström and
Borin, 2011) and word segmentation (Hewlett and
Cohen, 2011), should make it possible to identify
likely cx candidates, which must then be judged
manually, but the heuristic work process should
be greatly simplified using these kinds of meth-
ods. Another possible research direction is the
type of methods used to locate multiword expres-
sions and terminology (see, e.g., Pecina 2010),
which need to be further developed to cater for
the identification of cx, where a position might
have schematic content rather than being a spe-
cific word. For the latter, the morphosyntactic and
syntactic information provided in the Korp anno-
tations will be used (cf. Baroni and Lenci 2010; Piitulainen 2010).

6 Outlook

Developing tools for automatic identification of cx is both a methodological approach and a highly relevant research objective in its own right. If we are able to automatically identify cx in authentic text, the ambiguity that has always plagued automatic syntactic analysis, where even relatively short sentences may have tenths or even hundreds of analyses, can be greatly reduced. Kokkinakis (2008) has shown that the identification of complex terminology and named entities simplifies a subsequent syntactic analysis considerably. Also, Attardi and Dell’Orletta (2008) and Gadde et al. (2010), and others, have shown how pre-identification of different types of local continuous syntactic units may improve a subsequent global dependency analysis. Our hypothesis is that cx can be used in the same way, and exploring this would be a valuable contribution to LT research. The cx primarily targeted in the project are largely language-specific, partly by virtue of containing lexical material. However, on a more abstract level, many of the classes of cx – and consequently the methods both for their discovery in corpora and for their use in LT applications – are expected to be cross-linguistically relevant. Hence our research on Swedish will be relevant to LT in general.

The constructicon is meant to be a large-scale, freely available electronic resource for linguistic purposes and language technology applications, in the first place. As already mentioned, it will be integrated in the SweFN network and, of course, benefit the network enriching it with cx. But the constructicon can also be regarded as a lexicographic resource per se, and of relevance for lexicography/lexicographers in general (cf. Hanks 2008). Cx have traditionally been neglected in dictionaries. Some cx can be found in the information given on valency, and many cx are indirectly presented in the usage examples (cf. Svensén 2009, 141ff., 188ff.). The coverage, however, is only partial, since the dictionaries tend to favor colorful fixed phrases 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, idioms, which by comparison are used quite rarely (Farø and Lorentzen, 2009). Furthermore, paper dictionaries are inherently limited, partly due to their size, partly due to their structure; they are mainly based on headwords/lemmas. Electronic dictionaries, on the other hand, offer new opportunities through alternative search paths and (more or less) unlimited amount of space. Hence, in a longer perspective, the constructicon can be further developed and adapted as an extension to a future, general language e-dictionary of Swedish.

The improved coverage of constructional patterns provided by the constructicon should also be a valuable contribution to the fields of second language research and second language learning. As mentioned above, it is a special priority to account for cx that are problematic for second and foreign language acquisition. Besides such cx in particular, the constructicon in general should be highly relevant for L2 research and teaching. Usually, L2-learners do not acquire cx to any larger extent, except for the most general types. On the contrary, even advanced L2 learners have to rely on grammatical rules in their language production – in contrast to native speakers, who use prefabricated cx-templates extensively. This results in unidiomatic L2 production. It also adds a cognitive strain on the L2 speaker, since combinatorial language production is more taxing for the processing memory (Ekberg, 2004, 272), which makes L2 production more difficult than it needs to be. Adding the aspect of cx to L2 teaching situations would facilitate L2 learning for advanced students as well as for those who find traditional grammar an obstacle.

In summary, the constructicon is not only a desirable and natural development of the FrameNet tradition, it is also potentially useful in a number of areas, such as language technology, lexicography and (L2) acquisition research and teaching. In addition to these practical uses, we hope that this work will lead to theoretically valuable insights about the relation between grammar and lexicon.

Acknowledgements

The research presented here was supported by the Swedish Research Council (the project Swedish
References


Julia Prentice and Emma Sköldberg. 2011. Figurative word combinations in texts written by adolescents in multilingual school environments. In Roger


