The Past Meets the Present in Swedish FrameNet++

Lars Borin, Dana Danélls, Markus Forsberg, Dimitrios Kokkinakis and Maria Toporowska Gronostaj
Språkbanken, Department of Swedish Language, University of Gothenburg, Sweden

The paper is about a recently initiated pilot project which aims at the development of a Swedish framenet as an integral part of a larger lexical resource, hence the name "Swedish FrameNet++" (SweFN++). The SweFN++ project has four main goals: (1) to 'revitalize' a number of existing lexical resources and integrate them into a multi-faceted lexical resource for language technology (LT) applications, in the process enriching the individual resources using semi-automatic methods; (2) to construct a Swedish framenet (SweFN) and make it part of the integrated resource; (3) to develop a methodology and workflow which makes maximal use of LT and other tools in order to minimize the human effort needed to build the resource; and (4) to release the resource under an open content license.
The above goals are also of great significance for lexicological research and computational lexicography, as a SweFN will lend relevant support in bringing to light semantic relations implicit in word meanings. The theoretical assumptions elaborated by the Berkeley FrameNet make up the backbone of the SweFN resource, which will pay special attention to compounds and multi-words expressions when used as target lexical units or frame elements. In this article, we present an inventory of free electronic resources with a focus on their role in the semi-automatic acquisition and population of Swedish frames. After a brief overview of Swedish resources, we reflect on attempts to recycling and linking lexical data in a semi-automatic manner and report on our work in progress, which can be followed at http://spraakbanken.gu.se/swefn/eng/ .

1. Background and motivation

Access to multi-layered lexical, grammatical and semantic information representing text content is a prerequisite for lexicological and linguistic research, as well as for many LT applications. Information about the types of lexical frames of the words of the language, the frame elements of each such frame type described in terms of their semantic roles (semantic valency) and their syntactic manifestations (syntactic valency), are arguably necessary components of a full-fledged modern computational lexical resource. The earliest and best-known such resource is without doubt the Berkeley FrameNet (Ruppenhoffer et al. 2006; Fillmore 2008). Compiling dictionaries as well as text understanding and generation of natural language by computers are some applications which can benefit from the information provided by a framenet.

Currently FrameNet-like resources exist for a few languages,[1] including some domain-specific and multilingual initiatives (Boas 2009; Uematsu et al. 2009; Venturi et al. 2009), but are unavailable for most languages, including Swedish, except for some pilot studies exploring the semi-automatic acquisition of Swedish frames (Johansson & Nugues 2006; Borin et al. 2007).

At the University of Gothenburg, we are now embarking on a project to build a Swedish FrameNet-like resource. It is intended to be a free, full-scale, multi-functional resource covering morphological, syntactic and semantic description of 50,000 lexical units, with information accessible to both human users and LT systems. To make the work on this project cost and time effective, we intend to reuse freely available digital resources and software. A novel feature of this project is that the Swedish FrameNet will be an integral part of a larger many-faceted lexical resource. Hence the name *Swedish FrameNet++* (SweFN++). This larger resource will besides information on the modern Swedish encompass lexical data on 19th century Swedish, as well as eventually on Old Swedish (1225–1526). For

[1] See http://framenet.icsi.berkeley.edu.

more information on the ongoing work on the Old Swedish strand see Borin & Forsberg (2009a).

In this paper, the focus is on the resources and methodology aiming at the acquisition and population of the frames with lexical units for modern Swedish. A short presentation of the resources (section 2) is followed by an overview of methods aiming at acquisition of frames and targeting lexical units evoking particular frames (section 3). We report on our work in progress (section 4) and also address the issue of dealing with compounds and multiword expressions (section 5). We round off with conclusions (section 6).

2. Lexical resources for recycling

As a result of many years work on Swedish linguistic resources and Swedish lexicography, there are at our disposal a number of freely available digital linguistic resources of various kinds – including both data and processing resources – with various stages of coverage, and in various formats.[2] When now planning the construction of a Swedish FrameNet, thinking green should be the order of the day, i.e., recycling these resources should be a priority (cf. Green et al. 2004). If by this the resulting resource can become something more than a FrameNet – as we know it will – so much the better. This is the 'plus-plus' aspect of SweFN++. Below we describe briefly some of the existing lexical resources for the modern Swedish with a focus on their contribution to SweFN++.

2.1. Resources for modern Swedish at Språkbanken

SALDO is the core lexicon of SweFN++ to which all other information is to be linked, as it provides morphological and lexical-semantic information on 73,000 entries (senses expressed by single words or multiword units).[3] The lexicon is an updated version of *The Swedish Associative Thesaurus* (Lönngren, 1989), remade into a fully digital resource and enhanced by Borin & Forsberg (2009b). SALDO is an unusual semantic network in which lexemes of all word classes are arranged into a hierarchical structure according to the principle of centrality, capturing semantic closeness between two lexemes. Each lexical unit is given an obligatory main descriptor, its *mother*, which can be complemented by an optional determinative descriptor, its *father*. In the examples below, shown in Table 1, the mother precedes the plus sign and the father follows it, whenever it is present.

bröd 'bread':	mat + mjöl	'food + flour'
brud 'bride':	gifta sig + hon	'get married + she'
bröllop 'wedding':	gifta sig	'get married'
gifta sig 'get married':	par	'pair'

Table 1. Lexical information taken from SALDO.

The mother descriptor is usually: (i) semantically closer to the entry word; (ii) semantically and/or morphologically less complex than the entry word; (iii) more frequent; (iv) stylistically more unmarked; and (v) acquired earlier in the first and second language acquisition. Father descriptors are used mainly to differentiate lexical units (senses) having the same mother. They are assigned to about 50% of the entries. Thus, information about the entry's mother

[2] It is important to our goals that the resources be freely available and modifiable, i.e., under an Open Source or Open Content license, since we plan to make the resulting resource available under this kind of license.

[3] See http://spraakbanken.gu.se/sal/eng/.

and father provides necessary clues for disambiguation of polysemous entries. It also offers some means for semantic classification of lexical units as those which share specific mother and father relations are more closely related to each other, as compared to those having different father descriptors. Thus, it is plausible to assume that such related lexical units can populate the same semantic frames as target units or as manifestations of frame elements. Consequently, this semantic information can be one of many factors conducive to populating the frames.

Due to the core status of SALDO in SweFN++, its lemmas and lexical units are the items to which information from other electronic resources, contributors to the SweFN++, is linked, e.g. the SIMPLE and PAROLE lexicons as well as the People's Synonym Dictionary. These resources are described below.

The SIMPLE and PAROLE lexicons for Swedish are lexical resources aimed at LT applications, results of the EU projects PAROLE (1996–1998) and SIMPLE (1998–2000) (Lenci et al. 2000). The Swedish SIMPLE lexicon contains 8,500 semantic units being characterised with respect to semantic type, domain and selection restrictions. All the items are also linked to the Swedish PAROLE lexicon, which contains 29,000 syntactic units representing syntactic valency information. The question is whether this meta-information indicating semantic and syntactic valency can support selection of the relevant frame(s) out of the repository of the FrameNet frames. Let's have a look at the verb *gifta sig* 'get married' and the information provided in the Swedish SIMPLE lexicon:

Semantic type:	Cooperative_activity
Domain	General
Selection restrictions:	Arg 0=Human Verb Arg1=Human

Using the repository of English frame elements, on the one hand, and the SIMPLE data on the other hand, a lexicographer can map the semantic type of arguments onto the semantic role Partner, as humans involved in a cooperative activity can be perceived as partners. In FrameNet, the Partner role is dealt with in terms of the core elements Partner_1, Partner_2 and Partners and occurs in the frames: Collaboration, Forming_relationships and Personal_relationship. Selecting a correct frame automatically can in many cases be far from trivial, even if both the semantic and syntactic criteria specified in the SIMPLE and PAROLE lexicons are taken into consideration. In this particular case, the information on reflexivity in combination with the information on the semantic type of the verb, cooperative activity, can support exclusion of the frame Personal_relationship, describing a static relation rather than an act. The Collaboration frame can be precluded because it implies occurrence of the frame element Undertaking, being absent in the frame Forming_relationships. Since these three frames manifest the semantic relations by means of different syntactic valency patterns the identification of an appropriate frame can be performed automatically, whenever the full set of core elements is represented in a text.

The Semantic Database (SDB) is a valency description of a number of verbs using a set of about 40 general semantic roles (Järborg 2001), and linking them to example instances in a balanced corpus (about 200,000 occurrences). One goal of the project will be to find effective ways of linking FEs to SDB roles.

2.2. Other free lexical resources

The People's Synonym Dictionary (Synlex) (Kann & Rosell 2006) is a collaborative effort where a large number of people have rated the synonymy of a word pair (randomly chosen from a large set of synonym candidates) on a scale from 0 to 5. The downloadable version contains all word pairs with an average rating in the interval 3 to 5,[4] almost 40,000 Swedish synonym pairs. This resource can make the work on populating the Swedish frames more efficient, as it provides partially preprocessed data for further semantic analysis aiming at identification of their frame elements and frame assigning. In the course of the pilot project, a subset of 8,552 monosemous items have been linked to SALDO.

The People's Dictionary is a bilingual resource covering Swedish and English,[5] free for downloading. It is initially based on the Lexin Swedish-English and English-Swedish dictionary, but it is being expanded and improved by a similar collective effort as Synlex. The proposed translational equivalents are rated as either: very bad, pretty bad, pretty good, very good, don't know and inappropriate/spam. The high-rated pairs provide data of interest for populating the frames in SweFN based on cross transfer of semantic information. In analogy to what has been said about the treatment of synonyms, even these need to be controlled whether they are compatible with description of the semantic valency characterising a particular frame.

The Intercontinental Dictionary Series (IDS) and Loanword Typology (LWT) lexical unit lists contain about 1800 concepts considered likely to have lexicalizations in a large number of languages (IDS is included in LWT).[6] [7] The freely available lists make good candidates for the SweFN++ core vocabulary and provide links to this vocabulary in many other languages.

Swedish Wiktionary contains 43,000 entries,[8] subdivided into senses with definitions. Definitions are rare in other free lexical resources, which makes Swedish Wiktionary interesting for our purposes.

The Lund University Frame List comes out of work by Johansson & Nugues (2006), who have made several experiments aiming at automatic creation of Swedish frame candidates by processing of parallel corpora. The lemma-frame list obtained with an automatic role labeller applied to a limited word-aligned bilingual corpus turns out to require a lot of manual revisions. Hence, we refrain from using it.

From what has been presented above follows that the available lexical resources are heterogeneous as to their content. They are also coded in different formats. They have been developed for different purposes by different groups of researchers, some by linguists, some by language technology researchers – possibly with little linguistic background or none at all – and yet others in Wikipedia-like collective efforts. Thus one of the main challenges for SweFN++ is to ensure content interoperability not only among the lexical resources but also between the available tools for text processing and lexical resources to be used by various

[4] See http://lexikon.nada.kth.se/synlex.html.

[5] See http://folkets-lexikon.csc.kth.se/folkets.en.html# (in English).

[6] See http://lingweb.eva.mpg.de/ids/.

[7] See http://world.livingsources.org/semanticfield/.

[8] See http://sv.wiktionary.org/.

pieces of software in a distributed processing environment consisting to a large part of web services. We will also need to formulate strategies for dealing with the uneven distribution of some types of information in the integrated resource (e.g. syntactic valency information is at present available for about one fourth of the entries). This is work that we have initiated quite independently of the SweFN++ plans, within the European infrastructure initiative CLARIN.[9] Developing a set of special strategies for semi-automatic linking of homographic lemmas and polysemous items across the lexical resources is a great challenge for this project.

3. Methodology for FrameNet development

Developing a resource like FrameNet for a language is a time and labour intensive task. Hence, we aim to find ways of conducting the work which will minimize human effort and focus it where it will be most useful. This involves both arranging the workflow with respect to automated processing and human work and devising effective tools with good interfaces for the latter. In a small pilot study, we have started the work of linking some of the existing lexical resources, mainly to get a rough estimate of how much manual work would be involved, and also work on the framenet itself, focusing on domain-specific vocabulary in medicine and art, and also some selected frames from a general domain. The outcomes are 38 skeleton frame descriptions for about 1,500 lexical units.[10]

In international framenet initiatives, the same two general methodological approaches can be observed as in international wordnet projects (Vossen et al. 2009): the extension approach, where the English resource is translated into the target language (e.g. the Spanish framenet), and the merging approach, where the resource is built from the ground up in the target language (e.g. the German SALSA project), and later often linked to the English resource. Both approaches have their pros and cons, which have been much discussed in the literature on wordnets. In the pilot study we have used mainly the extension approach, and this will be the main method also in the first phase of the project. Later, when the LT support tools for the project become available, we will switch to a more Swedish-centered approach reusing the data from merged lexical resources and corpus-based acquisition of frames.

3.1. Extension approach

A prerequisite for the extension approach based on lexical transfer of information from English to Swedish is access to a digital bilingual dictionary supplying Swedish equivalents to lexical units evoking frames. Free bilingual resources such as the People's Dictionary and the LWT list are potential sources of information which can support translation of the lexical units in the FrameNet into Swedish. However, the fact that a lexical unit can have two or more translational equivalents creates a bottleneck for automatic processing of data, as there is often a spectrum of shades of meaning, which are associated with differences in the semantic and syntactic valency structures.

This can be exemplified with the English verb *marry* which evokes the frame Forming_relationships and which has the following eight translational equivalents according the People's Dictionary: *förena* (figuratively), *gifta bort, gifta sig, ingå äktenskap, gifta sig med, äkta, viga, förena i äktenskap.* Out of these, the following five *gifta sig, ingå äktenskap, gifta sig med, äkta, förena i äktenskap* meet the criteria posed on the frame elements in the frame Forming_relationships, namely that "Partner_1 interacts with Partner_2 (also

[9] See http://www.clarin.eu.

[10] See http://spraakbanken.gu.se/swefn/resurser/swefn-db-stat.html.

collectively expressable as Partners) to change their social relationship". According to this definition, Partner_2 is a kind of a co-agent which makes it possible to conflate the Partner_1 and Partner_2 roles into one, namely Partners. The remaining three verbs, *förena* 'unite, combine', *gifta bort* 'marry sb to sb', *viga* 'marry bride and bridegroom', imply roles which do not conform to those in the quoted definition, as the underlying semantic relations are those of an Agent and Patient respectively, if expressed in terms of a more abstract role repository. Thus, the automatic processing of strings of equivalents is far from simple and might require a lot of pre-processing of texts on both the syntactic and semantic levels.

One of the minor disadvantages of this method is that it is biased toward creating a Swedish copy of the English FrameNet, rather than developing an original resource capturing the nature of the Swedish language. But as already hinted at, it will be balanced with the merging approach, as the development of methods for semi-automatic direct acquisition of frames is one of the main goals in the SweFN++ project.

3.2. Merging approach

The merging approach aims at direct acquisition of frames from Swedish corpora that are semantically and syntactically annotated (Borin et al. 2007). From the literature it is clear that linguistic annotation yields better data than working directly with raw text, and that higher-level annotation, e.g. functional syntax, is better than less sophisticated processing, e.g. part-of-speech tagging. We will use existing freely available automatic annotation tools, such as the SALDO morphology (Borin et al. 2008), the Swedish named entity recognition system of Kokkinakis (2004), the Hunpos tagger (Halacsy et al. 2007) and Maltparser (Nivre et al. 2007) in order to achieve functionality along the lines of (the commercial systems) Sketch Engine (Kilgarriff et al. 2004) or Deepdict (Bick 2009).

Even if an extensively annotated corpus of general language is not available at this stage of the project, we were able to experiment with syntactic and semantic tagging of Swedish texts. The parser that we applied to the annotated data is based on finite-state cascades (Kokkinakis & Johansson Kokkinakis 1999). It is aware of shallow semantic annotations from both medical thesauruses and a generic named entity recognition component (Kokkinakis 2004). The experiments conducted so far have indicated that this annotation can pinpoint relevant frame elements and frames (see Fig 1). For example, the semantic annotation tags, like DISEASE, SYMPTOM, ANATOMY, CHEMICAL, TECHNIQUE occurring in example sentences with the verb *lindra* (meaning 'relieve', 'mitigate', 'soothe', 'palliate') can be relatively easily mapped onto frame elements such as Affliction, Body-part, Medication, Treatment occurring in the Cure frame The matching can be conducted manually to assure a high quality of the mapping. However, the mapping of the label PERSON onto frame elements Healer and Patient is less straightforward, as access to either deeper semantic information or contextual preferences imposed by the frame evoking unit is required for identification of those particular semantic roles. It can also be noted that several general non-core element e.g. Time, Place correlate with the tag set used for the annotation.

Blodgruppsdieten påstås ha en sammansättning som förhindrar eller lindrar <SYMPTOM>symtomen</> vid allvarliga <DISEASE>sjukdomar</> som <DISEASE>hjärt- och kärlsjukdomar</> , <DISEASE>cancer</> , <DISEASE>diabetes</> , <DISEASE>astma</> och <DISEASE>allergier</> samt <DISEASE>infektionssjukdomar</>

<TECHNIQUE>Steloperation</> av <ANATOMY>fotleden</> lindrar <SYMPTOM>smärta</> väl men medför en del komplikationer .

<TIME>Idag</> behandlas <DISEASE>restless legs</> framför allt med <CHEMICAL>L-dopa</> , vilket lindrar <SYMPTOM>symtomen</> .

I en nyligen genomförd studie på <ORG>Karolinska sjukhuset</> fann vi att <CHEMICAL>Emla</> inte heller lindrar <SYMPTOM>smärtan</> vid hälstick på <PERSON>nyfödda fullgångna barn</> i samband med PKU-prov [8] .

Figure 1. An excerpt from the *Läkartidningen* (Journal of the Swedish Medical Society) corpus with semantic named entity annotations. (ORG is an acronym for organisation)

The tests on the medical corpora reaffirm our assumption that richly annotated text can be helpful in the acquisition of frames, as the set of frame elements associated with the frame evoking lexical units can narrow the set of relevant frames and in the best case indicate a proper frame. Semantic pre-processing of text can also result in improvement of accuracy for syntactic relation extraction (e.g. subject, object), which is relevant to the development of a FrameNet, and in particular to the recovering of frequent semantic syntactic valency patterns. This information can hopefully contribute to the decrease of coordination and structural ambiguity errors in the process of syntactic parsing. It is our conviction that mutual feedback can further enrich both the named entity system and the content of SweFN.

4. Work in progress and results

4.1. Methodological perspective

SweFN++ will differ from the Berkeley FrameNet in one crucial aspect: It is from the beginning planned as an LT resource, which has implications for how it is created and maintained. For example, to be able to use the information about frame elements there is no room for human intuition, everything must be consistently formalized.

One of the main questions in this context is how to find a good balance between computational methods and manual work. Computational methods always produce errors for non-trivial linguistic descriptions but are completely consistent. Linguists are able to deal with the irregular and more difficult linguistic description problems, but humans are not as good as computers at being completely consistent, and also much more expensive than computer processing. Thus, the focus is on developing a workflow where automatic processing is used to produce, group and rank data and corpus examples for manual inspection. A part of the task is to exactly define what is meant by a relevant data item or a good corpus example (Kilgarriff et al. 2008).

At the same time, it has been important to us to get the work started, rather than wait for the 'best' solution to emerge first. Hence, methodological development is explicitly iterative in the SweFN++ project, so that even though we knew that the technical solutions used in the pilot study would not be ideal, they allowed us to get started quickly, and to get a better understanding of some of the methodological issues involved.

One valuable insight from the pilot study concerned the estimated manual effort involved in linking the existing lexical resources. If we plot the number of senses against the number of citation forms in SALDO, we get a Zipfian-like distribution: Only about 4,300 out of the

67,300 citation forms have more than one sense, and the average number of senses per citation form is slightly above two among the 4,300. The practical consequence of this, which seems to hold for the other lexical resources as well, is that most of the content mapping can be automated, leaving a manageable residue for manual or semi-automatic resolution. This will generate some errors, but our working hypothesis is that the result will be good enough for practical LT applications. Importantly, the information encoded with great investments of time and money in these resources will not be lost.

Consequently, in linking the existing lexical resources we work according to a model where an initial automatic linking is made, by identifying citation forms (lemmas) with SALDO items and adding the corresponding SALDO sense identifiers. Since most of the resources are large, exhaustive manual inspection of the linkages is precluded, however. For instance, we have made an initial linking of Synlex and SALDO, for Synlex pairs where both members correspond to only one SALDO sense each. These were gathered into WordNet-style synsets using different synonymy rating thresholds.[11] With all but the highest degree of synonymy, it invariably turns out that some synsets become very large. In part this may be due to the people's notion of synonymy being 'looser' than that of lexicographers, but in part it is because senses in SALDO tend to be fairly coarse-grained. In such cases, manual inspection has resulted in senses being added to SALDO. An interesting methodological issue is whether such anomalous synsets could be detected by some more linguistically informed means than their size, e.g. using the topology of SALDO seen as a lexical-semantic network. Another possibility is to use automatic methods known to yield good results on judging synonymy on the basis of text occurrences, e.g., random indexing (Sahlgren 2006; Rosell et al. 2009), and in particular the linguistically informed approach of Pado & Lapata (2007). This will be investigated in the project.

4.2. Lexicographic perspective

The theoretical framework of the Berkeley FrameNet together with its rich spectrum of lexical information on the frames and the semantic and syntactic valency for about 10,000 lexical units representing the frames has been a point of departure for creation of SweFN. In the initial phase of the project, we focus on the description of semantic valency information. In our work we follow the Berkeley frames specifications concerning: (i) the name of frame; (ii) its definition pointing out semantic relations between the set of core elements inclusive their definitions; as well as (iii) the specification of non-core elements and their definitions. We also take advantage of the meta-information provided on the types of semantic relations between the frames. As the mentioned information is considered as default in the process of our frame encoding, and as it overlaps with that explicitly specified in the Berkeley FrameNet, it is not repeated and explicitly specified in the SweFN database interface. However, the SweFN the frame information is expanded with data on: (i) domains; (ii) links between frames and the notation of semantic types from the SIMPLE lexicon; and (iii) information on instantiated compounding (see section 5.1 for more information). Inclusion of the domain information opens for creation of sub-framenets for special vocabularies, e.g. art and medicine in contrast to the general language domain. Viewing the frames and their lexical units from the semantic type perspective by projecting the semantic type classification elaborated in the SIMPLE project facilitates merging the data from the two resources and populating the frames with lexical units, (e.g. all the frames dealing with commerce in the SweFN are linked to the Transaction type in the SIMPLE classification of semantic units.)

[11] See http://spraakbanken.gu.se/swefn/eng/, under "Lexical Resources > SWESAURUS".

An example providing an overview of the content in the Swedish frames is given in Figure 2. The labels in the left column specify the type of content in the column to the right and they stand for: name of the frame, domain, semantic type (referring to English FrameNet//SIMPLE), list of core elements, list of non-core elements, examples, list of instantiated compound types, examples illustrating the listed compound types, list of lexical units linked to SALDO, list of new lexical units for inclusion in SALDO, comment field, created by, date of creation and date of modification. The frame elements and their corresponding text instantiations in the example field are in matching colors. The examples listed come from the internet, but in the future will be taken from the annotated corpora being made available at Språkbanken.

ram	**Cure**
domän	Med
semantisk typ	0//Cause_change_of_state
kärnelement	**Affliction**, **Body_part**, **Healer**, **Medication**, **Patient**, **Treatment**
periferielement	**Degree**, **Duration**, **Manner**, **Motivation**, **Place**, **Purpose**, **Time**, **Type**
exempel	**Salvan lindrar** även **besvär** som **skavsår**, **sprickor på fingertopparna**, **stickor i fingrarna** samt **skärsår**. **Man** kan **behandla cancer med flera olika metoder**. **Läkaren opererade höger öga** i stället för vänster. **ST-läkaren behandlade henne med höga doser kortison**. **Salvan läker skrubbsår och brännsår**. **Genterapi botade dödssjuka i cancer**. **Transplantation** kan **ha botat hiv-smittad**. **Ljusterapi lindrar** och förebygger **nedstämdhet**, ökar din energinivå och stärker ditt inre lugn.
sms	Type+Treatment, Body_part+Treatment, Medication+Treatment
sms-exempel	Type+LU_EX_ljus.behandling, röntgen.behandling, strål.behandling, värme.behandling Body_part+LU_EX_hjärn.operation, hjärt.operation Medication+LU_EX_kortison.behandling
saldo	**vb:** avvänja..1 behandla..2 bota..1 hela..1 kurera..1 lindra..1 läka..1 läka..2 medicinera..1 operera..1 rehabilitera..1 vårda..1 återanpassa..1 **nn:** behandling..2 dialys..1 diatermi..1 hjärnoperation..1 hjärtoperation..1 huskur..1 läkning..1 knejpkur..1 kortvågsbehandling..1 kur..1 lindring..1 ljusbehandling..1 lobotomi..1 medicinering..1 operation..1 radikaloperation..1 rutinoperation..1 rehabilitering..1 resektion..1 röntgenbehandling..1 skrapning..1 stomi..1 strålbehandling..1 värmebehandling..1 värmeterapi..1 återanpassning..1
saldo (nya)	
kommentar	Obs. behandling förekommer här som både Treatment och det rambärande lexemet, LU.; En sammanflätning (conflation) kan förekomma i objektpositionen mellan Patient och Affliction med vissa verb i den här semantiska gruppen, tex bota hiv-smittad/aids. (Se The Book, s.25 cure epileptic/epilepcy); (I Eng Cure-ramen tolkas 'affliction' som the injuries, disease, pain.);
skapad av	MTG
skapad	2010-01-30
modifierad	2010-02-15

Figure 2. The Cure frame in SweFN from
http://spraakbanken.gu.se/swe/forskning/swefn/utvecklingsversion#Cure

Out of the 38 frames elaborated up to now (January 2010) in the SweFN pilot study, the above Cure frame is one of 34 frames matching its sister frames in the Berkeley FrameNet. The new ones in SweFN are Falling_ill, Health_status, Medical_disorders and People_by_disease. The Health_status and Medical_disorders frames are elaborations of the Berkeley FN frame Medical_conditions.

It should be noted that all the data created in this pilot project are made available for immediate inspection by project participants and subjected to a range of control programs checking the encoding. The linguistic data, technical reports and error reports are made publicly available for inspection on the project homepage (updated automatically twice a day).[12] We believe that open access to our work will be of benefit for other researchers and the wide public. We hope that the openness will promote wider collaboration and early feedback.

[12] See http://spraakbanken.gu.se/swefn/eng.

5. Specific issues to be addressed: compounds and multiword expressions

In SweFN++, we will pay special attention to the cases where lexical units and text words do not coincide. Firstly, we will need to deal with productive compounding (where compounds are written as one orthographic word, a characteristic of Swedish and many other languages, but not English), e.g. the implicit semantic relations underlying compounds. Secondly, multiword lexemes (multiwords) are an area of increasing interest in the LT community.

5.1. Compounds in SweFN

In the course of the work on SweFN, it has become obvious that compounds deserve special attention, as they are an inherent feature of the Swedish language. They can be produced on the fly, express a number of implicit semantic relations which need to be made explicit for LT use, and their components need to have SALDO sense identifiers. Furthermore, an explicit semantic annotation of compounds is required for specification of the alternations in semantic and syntactic valency patterns evoked by compound target lexical units, e.g. *Läkaren undersökte barnet* 'the doctor examined the child' versus *Barnet läkarundersöktes* lit. 'the child was doctor-examined'.

A closer look at the compound types and their examples shows that noun compounds, both deverbal (e.g. car purchase) and others (e.g. water jug), abound in the data. In Table 2 we list some examples showing the core (Buyer, Goods) and non core elements building compounds with the verbal noun *köp* 'buy, purchase'.

Goods+LU	*markköp*	'land purchase'
Manner+LU	*skenköp*	'under the guise purchase'
Means+LU	*avbetalningsköp*	'hire-purchase'
Purpose+LU	*tröstköp*	'comfort shopping purchase'
Purpose_of_goods+LU	*sexköp*	'sex-buying'

Table 2. Compound types with examples taken from the Commerce_buy frame.

There are several issues concerning compounds which will be examined in the future work: (i) their potential role in automatic disambiguation of polysemous lexical units; (ii) their effect on changes in syntactic valency patterns of the compound lexical units; and (iii) the preferences for variation in types of semantic patterns with respect to a frame. The formalisation of these issues will hopefully improve several LT applications, not to mention the merging procedures in SweFN++.

5.2. Multiwords expressions in SweFN

One aim of the SweFN++ project is to provide a principled treatment of Swedish multiwords in the resulting resource. The methodological implications of this are in fact far-reaching. If we are to be able to use LT tools for automatic and computer-assisted acquisition of frames and frame elements, these tools must be able to find and propose multiword candidates. Thus, an important component of the project will be to successively (non-trivially) modify and refine existing LT tools in this direction. This will involve the entire processing chain from raw text to syntactically and semantically annotated sentences. That is, it will involve the processing stages of tokenization, part of speech tagging, morphological analysis/lemmatization, word sense assignment, and syntactic analysis. It will be a central methodological issue in the project how to accomplish this in a way that will not disrupt the workflow, and where new information can be integrated in a principled way.

6. Conclusions

The very initial phase of the SweFN++ project has already engendered some general methodological reflections on the recycling of available resources, and concerning acquisition of frames and some possible ways to populate the frames semi-automatically. As the work will progress, the attention will be focused on: (i) coordinating the meta-language used in different resources; (ii) ensuring the correctness of the available frame data by manual inspection and semi-automatic mining of the lexicons and text corpora; (iii) refining the set of procedures which aim at the acquisition of new frames from corpora; (iv) optimising the methods to populate the frames with regard to a type of the frame (e.g., artefact frames like Clothing require different processing from event oriented ones like Forming_relationships). We expect that in its final shape SweFN++ will be relatively free from shortcomings of the particular lexical resources due to their corrective recycling and the support from a well-balanced semantically and syntactically annotated corpus.

It is still not clear how we can utilize the Berkeley FrameNet frame definitions, the Swedish parser, and the additional semantic and syntactic information from SIMPLE/PAROLE in an effective way that will facilitate disambiguation related tasks which are relevant to computational natural language processing systems. This is one of the many challenges we hope to resolve in the course of the presented research project.

References

Bick, E. (2009). 'DeepDict – A graphical corpus-based dictionary of word relations'. In Kristiina Jokinen; Eckhard Bick. (eds.). *Proceedings of the 17th NODALIDA*. NEALT Proceedings Series, Vol. 4. 268-271.

Boas, H. C. (ed.; 2009). *Multilingual framenets in computational lexicography*. Berlin: Mouton de Gruyter.

Borin, L.; Forsberg, M. (2009a). 'Something old, something new: A computational morphological description of Old Swedish'. In *LREC 2008 workshop on language technology for cultural heritage data (LaTeCH 2008)*. Marrakech: ELRA. 9–16.

Borin, L.; Forsberg, M. (2009b). 'All in the family: A comparison of SALDO and WordNet'. In Kristiina Jokinen; Eckhard Bick. (eds.). *Proceedings of the 17th NODALIDA*.

Borin, L.; Forsberg, M.; Lönngren, L. (2008). 'The hunting of the BLARK - SALDO, a freely available lexical database for Swedish language technology'. In Joakim Nivre; Mats Dahllöf; Beáta Megyesi. (eds.). *Resourceful language technology. Festschrift in honor of Anna Sågvall Hein.* Acta Universitatis Upsaliensis: Studia Linguistica Upsaliensia 7. 21–32.

Borin, L.; Gronostaj, M. T.; Kokkinakis, D. (2007). 'Medical frames as target and tool'. In *Frame 2007: Building frame semantics resources for Scandinavian and Baltic languages*. University of Tartu. 11–18.

Fillmore, C. J. (2008). 'FrameNet meets Construction Grammar'. In *Proceedings of the XIII Euralex international congress*. Barcelona. 49–69.

Green, R.; Dorr, B. J.; Resnik, P. (2004). 'Inducing frame semantic verb classes from WordNet and LDOCE'. In *Proceedings of the 42nd ACL*. Barcelona: ACL. 375–382.

Halacsy, P.; Kornai, A.; Oravecz, Cs. (2007). 'Hunpos – an open source trigram tagger'. In *Proceedings of the 45th ACL, Demo and Poster Sessions*, Prague: ACL. 209–212.

Järborg, J. (2001). *Roller i Semantisk databas* (Research Reports from the Department of Swedish, No. GU-ISS-01-3). University of Gothenburg: Dept. of Swedish Language.

Johansson, R.; Nugues, P. (2005). 'Using parallel corpora for automatic transfer of FrameNet annotation'. In *Proceedings of the 1st ROMANCE FrameNet workshop*. 26–28.

Johansson, R.; Nugues, P. (2006). 'A FrameNet-based semantic role labeller for Swedish'. In *Proceedings of Coling/ACL 2006*. Sydney: ACL.

Kann V.; Rosell, M. (2006). 'Free construction of a free Swedish dictionary of synonyms'. In *Proceedings of the 15th NODALIDA*. Dept. of Linguistics, University of Joensuu. 105–110.

Kilgarriff, A.; Rychly, P.; Smrz, P.; Tugwell, D. (2004). 'The Sketch Engine'. In *Proceedings of the 11th Euralex International Congress*. Lorient, France. 105–116.

Kilgarriff, A.; Husák, M.; McAdam, K.; Rundell, M.; Rychlý, P. (2008). 'GDEX: Automatically finding good dictionary examples in a corpus'. In *Proceedings of the XIII Euralex international congress*. Barcelona.

Kokkinakis, D. (2004). 'Reducing the effect of name explosion'. In *Proceedings of the LREC Workshop: Beyond Named Entity Recognition, Semantic labelling for NLP tasks*. LREC 2004. Lisbon: ELRA.

Kokkinakis, D.; Johansson Kokkinakis, S. (1999). 'A cascaded finite-state parser for syntactic analysis of swedish'. In *Proceedings of the 9th EACL*. Bergen: ACL.

Lenci, A.; Bel, N.; Busa, F.; Calzolari, N.; Gola, E.; Monachini, M.; et al. (2000). 'SIMPLE: A general framework for the development of multilingual lexicons'. In: *Lexicography* 13(4). 249–263.

Lönngren, L. (1989). 'A Swedish associative thesaurus'. In *Euralex 1998 Proceedings*. Liège: University of Liège. 467–474.

Nivre, J.; Hall, J.; Nilsson, J.; Chanev, A.; Eryigit, G.; Kübler, S.; Marinov, S.; Marsi, E. (2007). 'MaltParser: A language-independent system for data-driven dependency parsing'. In *Natural Language Engineering* 13(2): 95–135.

Pado, S.; Lapata, M. (2005). 'Cross-linguistic projection of role-semantic information'. In *Proceedings of HLT/EMNLP 2005*. Vancouver: ACL. 859–866.

Ruppehoffer J.; Ellsworth M.; Petruck, M. R. L.; Johnson, Ch. R.; Scheffczyk, J. (2006). *FrameNet II: Extended theory and practice.* http://framenet.icsi.berkeley.edu/index.php?option=com_wrapper&Itemid=126 [access date: 16022010].

Uematsu, S.; Kim, J. D.; Tsujii, J. (2009). 'Bridging the gap between domain-oriented and linguistically-oriented semantics'. In *Proceedings of the BioNLP 2009 workshop*. Boulder, Colorado, USA: ACL. 162–170.

Venturi, G.; Lenci, A.; Montemagni, S.; Vecchi, E. M.; Sagri, M. T.; Tiscornia, D.; Agnoloni, T. (2009). 'Towards a FrameNet resource for the legal domain'. In *Proceedings of the IIIth Workshop on legal ontologies and artificial intelligence techniques (LOAIT '09).* Barcelona.

Vossen, P.; Fellbaum, Ch. (2009). 'Universals and idiosyncrasies in multilingual WordNets'. In H.C. Boas (ed.). *Multilingual FrameNets in Computational Lexicography. Methods and Applications*. Berlin: Mouton de Gruyter. 319–345.