# Encoding of Compounds in Swedish FrameNet

**Karin Friberg Heppin**
Språkbanken
University of Gothenburg
karin.heppin@svenska.gu.se

**Miriam R L Petruck**
International Computer Science Institute
Berkeley, CA
miriamp@icsi.berkeley.edu

## Abstract

Constructing a lexical resource for Swedish, where compounding is highly productive, requires a well-structured policy of encoding. This paper presents the treatment and encoding of a certain class of compounds in Swedish FrameNet, and proposes a new approach for the automatic analysis of Swedish compounds, i.e. one that leverages existing FrameNet (Ruppenhofer et al., 2010) and Swedish FrameNet (Borin et al. 2010), as well as proven techniques for automatic semantic role labeling (Johansson et al., 2012).

## 1 Introduction

Like other Germanic languages (e.g. Dutch, German), compounding is a very productive word formation process in Swedish. Swedish FrameNet,[1] which is part of the larger Swedish FrameNet++ effort to create Swedish resources for language technology purposes, analyzes Swedish compositional compounds in a way that reflects the language's grammatical structure, records information about the internal structure of these compounds in Frame Semantic terms, and proposes using that information to automate the analysis.

## 2 Swedish FrameNet

Swedish FrameNet (SweFN), which began in 2011, is part of Swedish FrameNet++ (Borin et al., 2010), a larger project whose main goal is building a panchronic lexical macro-resource for use in Swedish language technology. Basing its work on the original FrameNet developed in Berkeley (BFN) ((ed.), 2003), SweFN is creating a lexical resource of at least 50,000 lexical units (LUs) with the express goal of automating as much of the process as possible.

Swedish FrameNet bases its contents on three resources: (1) BFN's frames, frame definitions and frame-to-frame relations, for efficiency and compatibility with other FrameNet-like resources; (2) lexical entries from the SALDO lexicon; and (3) example sentences from the KORP corpus collection (Ahlberg et al., 2013).

Building SweFN frames includes several steps. The SweFN researcher chooses a BFN frame with a Swedish analogue, and populates that frame with appropriate LUs. LU selection involves determining which of the BFN LUs have equivalents in Swedish, or searching SALDO for LUs in the frame. Using the KORP corpora, the researcher finds example sentences that illustrate the LU's meaning and annotates each sentence with the frame's FEs. SweFN draws all examples from corpus data; this paper also provides the larger context in which compounds occur.

SweFN LUs, be they single words or multiword expressions (MWEs), evoke frames, i.e. cognitive structuring devices constituting the basic building blocks of any framenet knowledge base. LUs are pairings of lemmas and frames, the latter schematic representations of events, objects, situations or states of affairs. Frame elements (FEs) identify the semantic roles of the participants of the scenario characterized in a frame, e.g. AGENT, THEME, or TIME. For each frame, example sentences illustrate the linguistic realization of LUs together with the frame's FEs for the Frame Semantic annotation of the sentence's constituents (Borin et al., 2013a; Borin et al., 2013b).

## 3 Multiword expressions in SALDO

As mentioned above, the SALDO lexicon serves as the primary resource for LUs in SweFN++ and consequently also for LUs in SweFN. SALDO contains almost 6,000 MWEs of three types, dis-

[1] Vetenskapsradet (contract 2010-6013) funded SweFN (http://spraakbanken.gu.se/eng/swefn), a free resource (CC-BY-SA 3.0, LGPL 3.0)

tinguished as follows (Borin et al., 2013a):

- **Continuous MWEs** corresponding to fixed and semi-fixed expressions[2], which may have internal inflection, but no intervening words, e.g. *enarmad bandit* (one-armed bandit) - 'slot machine'.
- **Discontinuous MWEs** corresponding to syntactically flexible expressions[2], which may have intervening words, such as particle or phrasal verbs, e.g. *ge ut* (give out) - 'publish'.
- **Constructions** partially schematic constructions or syntactic templates with one or more slots filled with items of specific types, those described in construction grammars, e.g. *ju X desto Y* - 'The Xer the Yer' (Fillmore et al., 1988).

SALDO treats solid compounds, i.e. single orthographic words, just as it treats single-word items, and does not normally define their formal structure explicitly. In most cases, Swedish compounds are written without a space between its constituents, as in *korvgubbe* (sausage+man) - 'hot dog seller'. However, different possible forms yield different meanings. The adjective + noun NP *varm korv* literally means 'hot sausage' (in the temperature sense); the continuous MWE *varm korv* means 'hot dog'; and the solid compound *varmkorv* refers to not necessarily prepared sausage for making hot dogs. As LUs in a SweFN frame, the solid compounds, when compositional or partially transparent, have constituents which manifest FEs or other LUs in the same frame. The next section discusses these compounds and their annotation in SweFN.

## 4 MWEs and compounds as LUs

SALDO's continuous MWEs, discontinuous MWEs, and solid compounds are candidates for SweFN LUs, much like simplex words. Solid endocentric compounds, which identify a more specific instance of the compound's head, constitute a large group of words in Swedish. SweFN provides an analysis for these, even though BFN does not (Friberg Heppin and Toporowska Gronostaj, 2012). In frames where solid endocentric compounds are LUs, SweFN records the pattern FE+LU, where the compound's modifier is a FE of the given frame and the head is another LU in the same frame. Thus, among others, `Observable_body_parts` has ATTACHMENT, and DESCRIPTOR, and POSSESSOR FEs. SweFN records the analysis shown below with segmentation points between compound parts marked with '|'.

- ATTACHMENT+LU stortå|nagel (big+toe+nail) - 'big toe nail', pekfinger|nagel (point+finger+nail) - 'index finger nail'
- DESCRIPTOR+LU ring|finger - 'ring finger', pek|finger (point+finger) - 'index finger', stor|tå 'big toe'
- POSSESSOR+LU häst|hov 'horse hoof'

Generally, compounds with more than two constituents consist of at least one constituent that is itself a compound. SweFN treats such compounds in the same way as it treats other compounds. For example, stortå|nagel (big+toe+nail) - 'big toe nail' instantiates ATTACHMENT+LU , where stortå (big+toe) - 'big toe' itself is analyzed as DESCRIPTOR+LU.

SweFN analyzes example sentences that include compounds of different types with FE and LU annotation tags. The next section describes this encoding.

## 5 Encoding of compounds

Ruppenhofer et al. (2010) describes two ways that BFN treats noun compounds. Conventionalized two-part words are treated as single LUs with no internal analysis, e.g., *firing squad*, *sugar daddy*, and *wine bottle*. When a frame-evoking compound has a modifier that happens to be a noun or relational adjective e.g., *restoration costs*, *military supremacy*, the compound's head is annotated as a LU of the frame in question and the modifier instantiates a FE of the same frame. Ruppenhofer et al. (2010) point out that the division between the two groups is not always clear.

SweFN relies on degree of compositionality to determine the extent to which compound analysis is encoded in a frame's example sentences, not the compound's degree of lexicalization. Thus far, the analysis has been produced manually. Section 6 presents a proposal for the automatic Frame Semantic analysis of compounds.

[2] As described by Sag et al. (2001)

### 5.1 Non-compositional compounds

Typically, non-compositional compounds are lexicalized. Otherwise, understanding them is not possible, since the sense of the compound is not apparent from its parts. SALDO lists lexicalized non-compositional compounds as individual entries, like simplex words. Taking its lead from SALDO, and because constituents of non-compositional compounds do not instantiate FEs, SweFN does not analyze such compounds further, as in (1), where *hästhov* (horse+hoof) - 'coltsfoot' (`Plants`) is annotated only as a whole.

(1) och [hästhovarna]$_{\text{LU}}$ lyser som solar
and coltsfeet+DEF shine like suns
*...and the coltsfeet are shining like suns.*

### 5.2 Compositional compounds

SALDO also treats solid compositional compounds as simplex words. In contrast, SweFN treats compositional compounds as LUs, analyzing them as FE+LU, as described above in section 4. Furthermore, SweFN annotates compositional compounds in example sentences both as wholes and with respect to their constituent parts, as in (2).

(2) ...klappret från [snabba]$_{\text{Descriptor}}$
...clatter+DEF from fast
[[häst]$_{\text{Possessor}}$[hovar]$_{\text{LU}}$]$_{\text{LU}}$
horse+hooves
*...the clatter from fast horse hooves.*

Rather than serving as a modifier, the first element of some compounds is the semantic head of that compound. In such cases, the syntactic head can be semantically transparent, as in *bakterietyp* (bacteria+type) - 'type of bacteria' and *kaffesort* (coffee+kind) - 'kind of coffee', or show the speaker's attitude toward the entity identified in the semantic head of the compound as in *gubbslem* (old man+mucus) - 'dirty old man' or *hästkrake* (horse+wretch) - 'wretched horse'. For this type of compound, the modifier and the whole compound are considered LUs in the given frame, as illustrated in (3); the syntactic head of the compound does not realize any frame element in the frame.

(3) Han fick syn på en [gammal]$_{\text{Age}}$
He got sight of an old
[vit]$_{\text{Persistent\_characteristics}}$ [[häst]$_{\text{LU}}$krake]$_{\text{LU}}$
white horse+wretch
*He caught sight of an old wretched white horse.*

### 5.3 Partially transparent compounds

For some non-compositional compounds, one constituent clearly instantiates a FE or LU of the frame that the compound evokes, but the other is opaque, as in *ryggskott* (back+shot) - 'lumbago' from `Medical_disorders`. The modifier *rygg* - 'back' is the body part location of the disorder; the head *skott* - 'shot' is interpreted as something that appears suddenly, as a gunshot, but its meaning is not transparent. Example (4) shows that SweFN treats the compound as a LU, and the modifier as instantiating the FE BODY_PART; SweFN does not treat the head separately.

(4) [Han]$_{\text{Patient}}$ fick [[rygg]$_{\text{Body_Part}}$skott]$_{\text{LU}}$
He got back+shot
[under uppvärmningen]$_{\text{Time}}$ och
during up+warming+DEF and
tvingades vila
forced+PASS rest+INF
*He got lumbago during the warm-up and had to rest.*

Naming different types or species of a class of entities often results in groups of compounds whose heads are the name of the class, e.g. *blåbär* (blue+berry) - 'blueberry'; the compound names a type of berry. In these compounds, the modifier may not have an independent meaning, e.g. *körsbär* (?+berry) - 'cherry', where *körs* is a cran morph, i.e. it has no meaning outside of its use as a modifier in the compound. SweFN annotates the modifiers of these compounds with the FE TYPE, as in (5), since they have no meaning except to discern one type (cherry) of the LU in question (berry) from other types.

(5) Ska vi plocka [[körs]$_{\text{Type}}$[bär]$_{\text{LU}}$]$_{\text{LU}}$
Shall we pick cherries
*Do you want to pick cherries?*

### 5.4 Modifier as lexical unit

SweFN also chooses to analyze sentences (that illustrate the use of a LU) where a compound's modifier evokes the frame under consideration. For example, the compound *gasdetektor* - 'gas detector' is analyzed with respect to the `Devices` frame, given the head *detektor* - 'detector'. However, the modifier *gas* - 'gas' is analyzed with respect to `Substances`. Typically, SweFN does not choose sentences for annotation where only the modifier of a compound evokes the frame in question. Still, doing so is possible, as in (6).

(6) En vätesensor är en
A hydrogen+sensor is a
[gas]$_{LU}$detektor som visar
gas+detector which shows
närvaron av väte
presence+DEF of hydrogen
*A hydrogen sensor is a gas detector showing the presence of hydrogen.*

If analyzing a sentence where the LU under consideration is a modifier of a compound, SweFN does not annotate the compound's head. This practice reflects that of BFN (Ruppenhofer et al., 2010, 42).

> [W]e never annotate the head noun as a frame element of a frame that may be evoked by the non-head...While the non-head must figure in some frame evoked by the head, the reverse is not true in the same way....

## 6 Future Research

With a well-designed encoding for compounds, SweFN is positioned to develop ways to automate its heretofore manual annotation of compounds. Here, we sketch out plans to pursue the automatic Frame Semantic annotation of modifiers of compounds.

Johansson and Nugues (2006) demonstrated the effective use of FN annotation for automatic semantic role labeling (ASRL) of Swedish text to produce annotation (comparable to Padó and Lapata (2005)). More recently, Johansson et al. (2012) investigated the feasibility of using Swedish FrameNet annotation as a resource in constructing an automatic semantic role analyzer for Swedish. We suggest the possibility of using comparable techniques for the analysis of Swedish compound forms, also including FN data for developing and testing the efficacy of the algorithms.

This proposal involves the following: (1) manually add solid compounds from SALDO to appropriate frames based on the head of the compound; (2) use Kokkinakis's (2001) compound analysis technique to identify the component parts of the compound, by finding n-grams of characters which do not occur in simplex words; (3) exploit existing SweFN annotation for adjacent non-compounded words to develop an ASRL system to annotate modifiers of Swedish compounds and test the system; (4) exploit existing SweFN annotation of full sentences to determine if a system trained on that data would improve ASRL of modifiers in compounds; (5) using the same basic techniques for developing training data, determine if BFN data would benefit ASRL of modifiers, as Johansson and Nugues (2006) demonstrated for Swedish text in general.

Initially, the proposed plan for ASRL of modifiers of compounds addresses compounds with (only) two elements. In principle, the same approach can be expanded to annotate multiple modifiers of head nouns, i.e. compounds with more than two elements. These compounds consist at least one constituent that is itself a compound, i.e. the compounding process has occurred more than once as described in section 4.

As more language technology and NLP researchers develop FrameNet knowledge bases for languages other than English, the need for automatic processes to produce annotation that suits the grammatical requirements of the particular language will increase, as will the importance of using existing resources efficiently and effectively. The research proposed here offers the possibility of providing an automatic process that would be useful for the Frame Semantic analysis of Swedish in particular and for other compounding languages (e.g. Dutch, German). Additionally, the technique may prove useful for the processing of compounds more generally.

## 7 Conclusion

Given the highly productive nature of Swedish compounding, lexical resources such as Swedish FrameNet must attend to the representation and analysis of compounds. This paper has presented a new feature in SweFN, the explicit recording of the FE+LU pattern for the analysis of compositional compounds, and suggests a research plan to analyze Swedish compounds automatically.

## Acknowledgments

The authors thank the anonymous reviewers for their helpful comments. The Swedish Research Council (grant agreement 2010-6013) and the University of Gothenburg via its support of the Centre for Language Technology and Språkbanken have supported this work.

## References


Malin Ahlberg, Lars Borin, Markus Forsberg, Martin Hammarstedt, Leif Jöran Olsson, Olof Olsson, Johan Roxendal, and Jonatan Uppström. 2013. Korp and Karp – a bestiary of language resources: the research infrastructure of Språkbanken. In *Proceedings of the 19$^{th}$ Nordic Conference of Computational Linguistics, NODALIDA*.

Lars Borin, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj, and Dimitrios Kokkinakis. 2010. The past meets the present in Swedish Framenet++. In *Proceedings of the 14$^{th}$ EURALEX International Congress*.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013a. SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4).

Lars Borin, Markus Forsberg, and Benjamin Lyngfelt. 2013b. Close encounters of the fifth kind: Some linguistic and computational aspects of the Swedish FrameNet++ project. *Veredas: Frame Semantics and Its Technological Applications*, 17(1).

Thierry Fonetenelle (ed.). 2003. *International Journal of Lexicography*. Number 16.3: 231–385. Oxford University Press.

Charles J. Fillmore, Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, 64.

Karin Friberg Heppin and Maria Toporowska Gronostaj. 2012. The rocky road towards a Swedish FrameNet. In *Proceedings of the 8$^{th}$ Conference on International Language Resources and Evaluation*, Istanbul.

Richard Johansson and Pierre Nugues. 2006. A FrameNet-based semantic role labeler for Swedish. Sydney.

Richard Johansson, Karin Friberg Heppin, and Dimitrios Kokkinakis. 2012. Semantic role labeling with the swedish framenet,. In *Proceedings of the 8$^{th}$ Conference on International Language Resources and Evaluation (LREC-2012);*, Istanbul, Turkey.

Dimitrios Kokkinakis. 2001. *A framework for the aquisition of lexical knowledge; Description and applications*. Ph.D. thesis, Department of Swedish, University of Gothenburg.

Sebastian Padó and Mirella Lapata. 2005. Cross-lingual bootstrapping of semantic lexicons: The case of framenet. In *Proceedings of the American Association of Artificial Intelligence Conference*.

Josef Ruppenhofer, Michael Ellsworth, Miriam R.L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2010. *FrameNet II: Extended theory and practice*. <https://framenet2.icsi. berkeley.edu/ocs/r1.5/book.pdf>.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2001. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3$^{rd}$ International Conference on Intelligent Text Processingand Computational Linguistics (CICLing-2002*. Berlin: Springer.