



NLP for Resource Building

Downloaded from: <https://research.chalmers.se>, 2026-03-17 00:45 UTC

Citation for the original published paper (version of record):

Johansson, R. (2021). NLP for Resource Building. The Swedish FrameNet++. Harmonization, integration, method development and practical language technology applications: 169-190.
<http://dx.doi.org/10.1075/nlp.14.07joh>

N.B. When citing this work, cite the original published paper.

NLP for resource building*

Richard Johansson | University of Gothenburg

 <https://doi.org/10.1075/nlp.14.07joh>

 Available under a CC BY-NC-ND 4.0 license.

Pages 169–190 of

The Swedish FrameNet++: Harmonization, integration, method development and practical language technology applications

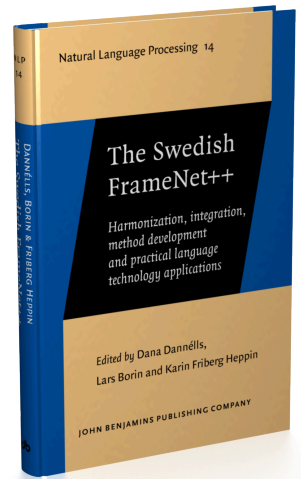
Edited by Dana Dannélls, Lars Borin and Karin Friberg Heppin

[*Natural Language Processing*, 14] 2021. xiv, 333 pp.

© John Benjamins Publishing Company

This electronic file may not be altered in any way. For any reuse of this material, beyond the permissions granted by the Open Access license, written permission should be obtained from the publishers or through the Copyright Clearance Center (for USA: www.copyright.com).

For further information, please contact rights@benjamins.nl or consult our website at benjamins.com/rights



NLP for resource building¹

Richard Johansson

University of Gothenburg

We evaluate several lexicon-based and corpus-based methods to automatically induce new lexical units for Swedish FrameNet, and we see that the best-performing setup uses a combination of both types of methods. A particular challenge for Swedish is the absence of a lexical resource such as WordNet; however, we show that the semantic network Saldo, which is organized according to lexicographical principles quite different from those of WordNet, is very useful for our purposes.

1. Introduction

Frame semantics has repeatedly been proposed as a practical, light-weight semantic representation suitable for lexicographic as well as natural language processing (NLP) purposes. A frame-semantic database such as FrameNet (Fillmore & Baker 2009) consists of two parts: first, a network of semantic frames, and secondly a lexicon that maps word senses to frames. In standard FrameNet terminology, the word senses of a frame are referred to as lexical units (LUs) evoking the frame. The frame network could to some extent be considered language-independent, while the lexicon is language-dependent (see also Chapter 8).

In order for frame semantics to be a viable alternative for implementing a practical NLP system, the lexicon must have a high coverage in the domain where the system is to be applied. It has been pointed out that the LU coverage is a major bottleneck for frame semantics that has a high negative impact on the quality of NLP systems using FrameNet (Palmer & Sporleder 2010). Burchardt et al. (2009) on the other hand claim that the difficulty of frame disambiguation is an even more severe problem than lexicon coverage for NLP applications.)

Addressing the resource bottleneck problem, several methods have been proposed to automatically suggest new candidates for inclusion. Johansson & Nugues

1. Parts of this chapter build on and elaborate content previously presented in Johansson (2014).

(2007b) approached this problem by applying a classifier on words based on the structure of their neighborhood in WordNet (Fellbaum 1998), and this led to a recall improvement in a frame-semantic analysis system (Johansson & Nugues 2007a). There have been several other ideas on how to find suitable lexical units. For instance, Pennacchiotti et al. (2008) used a combination of WordNet and geometric semantic representations automatically induced from corpora. Das & Smith (2011, 2012) applied a graph-based label propagation technique. Tonelli et al. (2013) used Wikipedia articles to propose new lexical units.

Swedish FrameNet (SweFN) is a resource under active development, with a primary focus on building a large lexicon with a good coverage (Friberg Heppin & Toporowska Gronostaj 2012). In order to make the lexicon-building process efficient, it may be useful to consider automatic or semi-automatic methods to integrate information derived from other lexicons or from corpora (Borin et al. 2010). The question is whether methods that have been applied to expand the FrameNet lexicon will also work for Swedish, as there are linguistic as well as resource availability considerations that make the Swedish situation different. On the linguistic side, Swedish has a slightly richer morphology than English; in particular, it uses compounding extensively,² and it is conceivable that this could lead to data sparsity that makes corpus-based methods infeasible. On the resource side, while most approaches for English have relied on WordNet, there is no widely available WordNet-like resource for Swedish. However, there is a large semantic network called Saldo that is organized by association rather than inheritance (Borin et al. 2013; see also Chapter 3), and it is an interesting question whether this makes the situation for frame-semantic lexicon expansion better or worse.

In this work, we evaluate and analyze a wide range of lexicon-based and corpus-based methods for automatic expansion of frame-semantic lexicons and see how well they work in SweFN. We see that Saldo works excellently for this purpose, and that the systems using Saldo outperform those using corpus-based methods. However, the best-performing setup uses a combination of both types of methods; in particular, it is useful to add corpus-derived information when handling frames where the set of LUs is not easily described in terms of the Saldo graph.

1.1 Frame semantics and frame-semantic lexicons

In the frame-semantic approach to lexical description, word meaning is specified by pairing the word's lemma with a *frame*: a structured representation of a type of situation, which defines a set of prototypical participants or properties of the

2. See the discussion of how compounds are treated in SweFN in Chapter 8.

scene, called *frame elements*. For instance, the verbs *eat* and *drink* evoke the frame `Ingestion`, which has frame elements such as `INGESTOR` (the one who eats or drinks) and `INGESTIBLES` (what is consumed). The frames are inter-connected in a network using relations such as inheritance or temporal order.

Lexical resources organized according to frame-semantic principles are usually called *framenets* after the first resource of this kind: the Berkeley FrameNet (BFN; Fillmore & Baker 2009). Although BFN was designed with English in mind, lexicographic projects for other languages have followed their approach (Boas 2009), sometimes also borrowing large portions of the BFN network structure. The motivation for this has been the conjecture that a system of frames and frame-to-frame relations in a given language would be structurally similar to its counterparts in other languages. This assumption has been used with some success to automatically construct frame-semantic resources in new languages (Johansson & Nugues 2006; Padó 2007). However, investigations of frame parallelism (Padó 2007) show that the frame structures in a pair of languages are not fully isomorphic, not even for closely related languages such as English and German.

SweFN (Friberg Heppin & Toporowska Gronostaj 2012; see also Chapter 2), is a lexical resource based on BFN. SweFN is a part of the SweFN++ project (Borin et al. 2010), a larger effort of linking and harmonizing several Swedish lexical resources (see Chapters 1 and 3). All lexical resources in SweFN++, including SweFN itself, are freely available for downloading. In April 2021, SweFN contained 1,195 frames populated by over 30,000 lexical units. It has been used as the basis for Swedish NLP systems such as a semantic role labeler (Johansson et al. 2012; see also Chapter 10 in this volume).

The structure of SweFN, including the frame and frame element nomenclature, is designed to follow BFN as much as possible; this makes it possible to use the English frame-to-frame relations as a way of sharing information between frames (Johansson 2012). One difference between SweFN and BFN is that the Swedish lexical units are defined with reference to a semantic lexicon, Saldo (see Section 2.1): instead of just saying that e.g. *run* evokes to `Self_motion`, SweFN lists a specific Saldo sense identifier, e.g. *springa..1*. Since all lexical resources in the SweFN++ project use Saldo as the backbone, this improves the interoperability between different resources.

2. Computational representation of the meaning of words

The field of computational lexical semantics deals with finding suitable ways for a computer to represent the meaning of words and word senses. In this work, we consider two orthogonal approaches of representing word meaning:

1. *The meaning of a word is defined by associating it to a concept in a semantic network.* In the simplest case, the network is just a list of concepts, but typically the concepts will be structured in a hierarchy ordered by relations such as *is-a*, *part-of* etc, so we can carry out logical inferences using the concepts they refer to: a mouse is a rodent, which is a mammal, etc.
2. *The meaning of a word is defined as a point or region in a geometric space.* While this view of meaning may be less intuitive to a lexicographer than network-based semantics, it is fairly popular in cognitive science (Gärdenfors 2000). This representation has some appealing properties, for instance that the notion of graded similarity becomes natural. To exemplify, we may say that the word *mouse* means something quite similar to the word *rat*.

In NLP, the first type of representation is typically associated with manually implemented lexical resources and the second with data-driven approaches, but this does not necessarily have to be the case.

2.1 The semantic network Saldo

As discussed in Chapter 3 of this volume, Saldo (Borin et al. 2013) is the most comprehensive open lexical resource for Swedish. In November 2013, it contained 125,227 entries organized into a single semantic network. Compared to WordNet (Fellbaum 1998), there are similarities as well as considerable differences. Both resources are large, manually constructed semantic networks intended to describe the language in general rather than any specific domain. However, while both resources are hierarchical, the main lexical-semantic relation of Saldo is the *association* relation based on centrality, while in WordNet the hierarchy is taxonomic. In Saldo, when we go up in the hierarchy we move from specialized vocabulary to the most central vocabulary of the language (e.g. ‘move’, ‘want’, ‘who’); in WordNet we move from specific to abstract (e.g. ‘entity’). Every entry in Saldo corresponds to a specific sense of a word, and the lexicon consists of word senses only. There is no correspondence to the notion of synonym set as in WordNet. The sense distinctions in Saldo are more coarse-grained than in WordNet, which reflects a difference between the Swedish and the Anglo-Saxon traditions of lexicographical methodology.

In Saldo, each entry except a special root is connected to other entries, its *semantic descriptors*. One of the semantic descriptors is called the *primary* descriptor, and this is the entry which better than any other entry fulfills two requirements: (1) it is a semantic neighbor of the entry to be described and (2) it is more central than it. That two words are semantic neighbors means that there is a direct semantic relationship between them, for instance synonymy, hyponymy, antonymy, meronymy, or argument–predicate relationship; in practice most primary descriptors

are either synonyms or hyperonyms. Centrality is determined by means of several criteria. The most important criterion is frequency: a frequent word is more central than an infrequent word. Other criteria include stylistic value (a stylistically unmarked word is more central), derivation (a derived form is less central than its base form), and semantic criteria (a hyperonym is more central than a hyponym).

To exemplify, here are a few instances of entries in Saldo and their descriptors.

Entry	Primary	Secondary
<i>bröd</i> ‘bread’	<i>mat</i> ‘food’	<i>mjöl</i> ‘flour’
<i>äta</i> ‘eat’	<i>leva</i> ‘to live’	
<i>kollision</i> ‘collision’	<i>kollidera</i> ‘to collide’	
<i>cykel</i> ‘bicycle’	<i>åka</i> ‘to go’	<i>hjul</i> ‘wheel’

2.2 Semantic representations induced from corpora

In NLP, the idea of representing word meaning geometrically is most closely associated with the distributional approach: the meaning of a word is reflected in the set of contexts where it appears, which is an idea with a long tradition in linguistics (Harris 1954). The distributional method allows us to automatically create meaning representations from corpora, which can be implemented either as geometric vectors or as a set of classes (clusters). Both types of automatically induced representations can be used to improve NLP systems significantly (Turian et al. 2010).

All corpus-based semantic representations used in this work were created using approximately 1 billion words downloaded from Språkbanken, the Swedish language bank. This is a mixed collection that comes from many different corpora: news, fiction, academic text, social media, Wikipedia, etc. The corpora are distributed in a format where the text has been tokenized and lemmatized, and compounds not listed in Saldo have been segmented. The representations described below were computed for lemmatized forms, not word forms.

2.2.1 Word representations from a class-based n -gram model

As a baseline, we evaluated one of the simplest corpus-based representations: automatically dividing the vocabulary into a small number of clusters. The Brown algorithm computes m hierarchically organized clusters of words, where m is a user-defined parameter, to maximize the likelihood of a corpus in a class-based n -gram language model (Brown et al. 1992). We computed word clusters using Percy Liang’s implementation of the Brown algorithm.³

3. <https://github.com/percyliang/brown-cluster>

Brown clusters have been used successfully in many NLP systems (Turian et al. 2010): instead of using a lexicalized feature representing a word directly, one can use a feature representing the cluster, which allows a generalization beyond the lexical level. However, the fact that the clusters are very coarse-grained, and that they are computed using a very small context in an n -gram model, leads to a grouping mainly according to syntactic criteria. This is not an issue e.g. for parsing (Koo et al. 2008), but may be problematic for semantic tasks such as LU induction. The popularity and simplicity of Brown clusters still makes it valuable to consider this representation as a baseline.

2.2.2 *Geometric word representations from co-occurrences*

The easiest way to create a geometric word representation is to implement the distributional idea directly: for each word, we create a vector where each dimension corresponds to a feature describing a context where the word has appeared. Typically, such a feature corresponds to another word with which the word has co-occurred, but in principle we can define arbitrary contextual features, for instance the syntactic context.

Random indexing (RI) is a method to drastically reduce the dimensionality of high-dimensional vectors while to a large extent still preserving their properties such as similarity (Kanerva et al. 2000). This is useful to reduce the memory consumption and processing time of the programs that use the vectors, because even if we use only the simplest feature representations (contextual words only), we will end up with a vector space with a very high dimensionality. RI works by mapping each dimension in the original high-dimensional space to a small number of dimensions in the reduced space. To implement RI, it is practical and efficient to use hash functions when carrying out the dimensionality reduction (Veldal 2011); this idea allows us to use a very large number of features.

In this work, we evaluated two types of vector spaces built from co-occurrence statistics:

- The *positional* (RI-Pos) space: the context of a focus word is represented using two words before and two after, taking position into account.
- The *contextual bag-of-words* (RI-CBoW) space: the context of a focus word is represented using five words before and five after, disregarding their position.

2.2.3 *Geometric representations from contextual classifiers*

As an alternative to co-occurrence vectors, geometric word representations can also be derived indirectly, as a by-product when training classifiers that predict the context of a focus word. While these representations have traditionally been built using quite complex machine learning methods (e.g. Turian et al. 2010), it was recently

shown by Mikolov, Chen, et al. (2013) that such representations can be created using much simpler and computationally more efficient methods. Interestingly, these representations seem to be able to capture a number of syntactic and semantic relations between words (Mikolov, Yih, et al. 2013), and it will therefore be interesting to see whether they are more effective than the commonly used representations based on co-occurrence.

Mikolov, Chen, et al. (2013) proposed two classification models from which the representations are derived, and we evaluate both of them in this work.

- *Contextual bag of words* (CC-CBoW): given a context (five words before and five after), the classifier predicts the focus word.
- *Skip-gram* (CC-SG): given a focus word, the classifier predicts the words around it (again, five before and five after).

We used the `word2vec` tool to build the vectors.⁴

3. From word meaning to frame meaning

Now that we have discussed a number of ways to represent the meaning of single words (or word senses), we turn to the question of how to deal with a set of words – the set of lexical units evoking a frame – and how we can design automatic tests for membership of that set. We consider two different approaches: (1) testing for membership by measuring how similar (or dissimilar) the potential new LU is to the set of existing LUs; (2) using the LUs to train a statistical classifier that can be used to test for membership of new LUs.

3.1 Methods based on distance and similarity measures

Pennacchiotti et al. (2008) describe a successful lexicon expansion system based on a combination of WordNet-based and corpus-based similarity measures. We investigated such measures for Swedish, using Saldo instead of WordNet. Since SweFN LUs correspond to Saldo senses, our task is easier since we know the Saldo sense both of the potential LU and the LUs already in the frame.

There are many ways to measure similarity or dissimilarity between two concepts in an ontology (Blanchard et al. 2005). In this work, we used one of the most simple and intuitive measures: computing the shortest path between two Saldo senses s_1 and s_2 (Rada et al. 1989):

4. <https://code.google.com/p/word2vec>

$$\text{Rada-dist}(s_1, s_2) = \min_{p \in \text{paths}(s_1, s_2)} \text{length}(p)$$

We computed the path length in the subgraph of Saldo consisting of primary descriptor edges only. This subgraph is a tree, which makes it very efficient to compute the path between s_1 and s_2 by finding their lowest common ancestor. Using the full Saldo graph including secondary descriptors makes it computationally harder to compute the shortest path, and preliminary investigations showed lower accuracies than using the primary descriptors only.

Following Pennacchiotti et al. (2008), when we consider a Saldo sense s for a frame F , we use the minimal distance (or maximal similarity) to measure the distance between a potential LU s and the existing LUs evoking F .

$$\text{min-Rada-dist}(s, F) = \min_{s_f \in F} \text{Rada-dist}(s, s_f)$$

Geometric distance and similarity measures are an alternative to ontology-based measures (Mohammad & Hirst 2012). The most widely used geometric measure is the cosine similarity; again following Pennacchiotti et al. (2008), we represent a frame F geometrically by computing its centroid c_F . This is done by averaging the vectors v_l representing the lemmas l defined with respect to F according to one of the representations described in Section 2.2.

$$c_F = \frac{1}{|F|} \sum_{l \in F} v_l$$

We can then measure the similarity between a new lemma l and a frame F by computing the cosine similarity between v_l and c_F :

$$\text{cos-sim}(l|F) = \frac{v_l \cdot c_F}{|v_l| |c_F|}$$

For both types of similarity or distance measures, we need to introduce a threshold b_F if we want to make a hard decision on whether to assign a potential LU s to a frame F , rather than just ranking all LUs by suitability for F . In that case, we say that s is assigned to F if $\text{min-Rada-dist}(s, F) < b_F$ or if $\text{cos-sim}_F(s, F) > b_F$. For each frame F , we set b_F by maximizing the harmonic mean of the precision and recall in a 2-fold cross-validation over the training set.

3.2 Classification-based methods

Ontology-based or corpus-based similarity measures have been used successfully for lexical unit induction, but there may be other more flexible approaches. For instance, Johansson & Nugues (2007a) addressed this task as a classification problem:

using standard machine learning tools, a classifier is trained for each frame and then applied to potential LUs. If this classifier outputs a numerical score, it basically corresponds to a similarity function tailored for that frame, as opposed to a generic similarity function. Another advantage of using a classifier is that it is easy to add arbitrary features to the model without having to decide how they should affect the similarity score.

For each frame F , we created a linear scoring function score_F of the following form:

$$\text{score}_F(s) = w_F \cdot \varphi(s)$$

Here, w_F is a weight vector specific to F , and $\varphi(s)$ is a *feature representation* of the Saldo sense s . If $\text{score}_F(s)$ is high, it means that our model considers s likely to evoke F . To create the vector w_F , we computed the features $\varphi(s)$ of all LUs in F and then applied the LIBLINEAR software for linear support vector machines (Fan et al. 2008). Again, in a use case where a hard decision is needed rather than a ranking of potential LUs, we assign a threshold b_F so that s is assigned to F if $\text{score}_F > b_F$. This threshold is also computed using LIBLINEAR.

Using a linear scoring function is a very simple method and it is reasonable to ask whether a more complex approach would be more effective, for instance a nonlinear support vector machine such as that used by Johansson & Nugues (2007a). However, we have only considered linear classifiers in this work, since computationally they are several orders of magnitudes more efficient than their nonlinear counterparts.

The feature representation function $\varphi(s)$ consists of three types of features:

- features representing the context of s in Saldo;
- features based on corpus-induced semantic representations;
- the part-of-speech tag of s .

It is straightforward to add the corpus-based features and the part-of-speech tag to a feature vector, but it is not obvious how to represent the Saldo context. We now consider this question.

3.2.1 *Representing the meaning of a word using Saldo*

When using a semantic network, the meaning of a word sense is defined by how it is related to other word senses; in Saldo, the immediate neighborhood of a word sense s consists of a primary descriptor and possibly a set of secondary descriptors, and the meaning of s can be further analyzed by following primary and secondary edges in the Saldo graph.

When using Saldo to create features for a potential LU s , we first compute a neighborhood representation that represents the meaning of s as a subgraph of the Saldo graph. A neighborhood representation consists of a set of Saldo senses, and the identifiers of the senses then become features in the representation $\varphi(s)$ used in the classifier. The purpose of a neighborhood representation is that it should describe the LUs in a way that allows us to generalize: to capture the meaning of all the LUs in the frame.

In this work, we consider five different neighborhood representations of a sense s :

1. *Descriptors*. Our baseline consists of the basic units of Saldo: the primary descriptor and secondary descriptors, if any. This is based on the idea that related words may be derived from a common source, or share a hyperonym. To exemplify, we consider the word *jogga* ‘to jog,’ which has the primary and secondary descriptors *springa* ‘to run’ and *motion* ‘physical exercise,’ respectively.
2. *Local context*. It is intuitive that the local context in Saldo of a sense s consists of a semantic field of related words. Such a set of words could possibly allow generalization from word meaning to frame meaning. We define this set as all senses at most 1 step away from s in the Saldo graph. This includes s , its primary and secondary descriptors, and its descendants: the senses for which it is the primary or secondary descriptor. For *jogga*, this set consists of the word itself, its primary and secondary descriptors *springa* and *motion*, and its descendants: *joggare* ‘jogger,’ *joggnig* ‘jogging,’ *joggande* ‘jogging,’ and *joggingsko* ‘jogging shoe.’
3. *Second-order context*. Based on the same intuition as the local context, we also tried a larger context. This set consists of all senses at most 2 steps away from s .
4. *Primary descriptor chain*. Another idea for generalizing in Saldo could be to describe s by following the primary links up to the root node. This representation is similar to the “hypernym tree” used by Johansson & Nugues (2007a), although as mentioned above, Saldo primary descriptors are not necessarily hyperonyms. The primary descriptor chain of *jogga* consists of the words *jogga*, *springa*, *röra sig* ‘to move oneself,’ *röra* ‘to move.’
5. *Primary descriptor chain with secondary descriptors*. The secondary descriptors contain semantic information that could also potentially be useful for discriminating frame membership, so we evaluated a representation where we add the secondary descriptors of all senses in the primary descriptor chain. For *jogga*, we add two secondary descriptors: *motion* and *fort* ‘fast.’

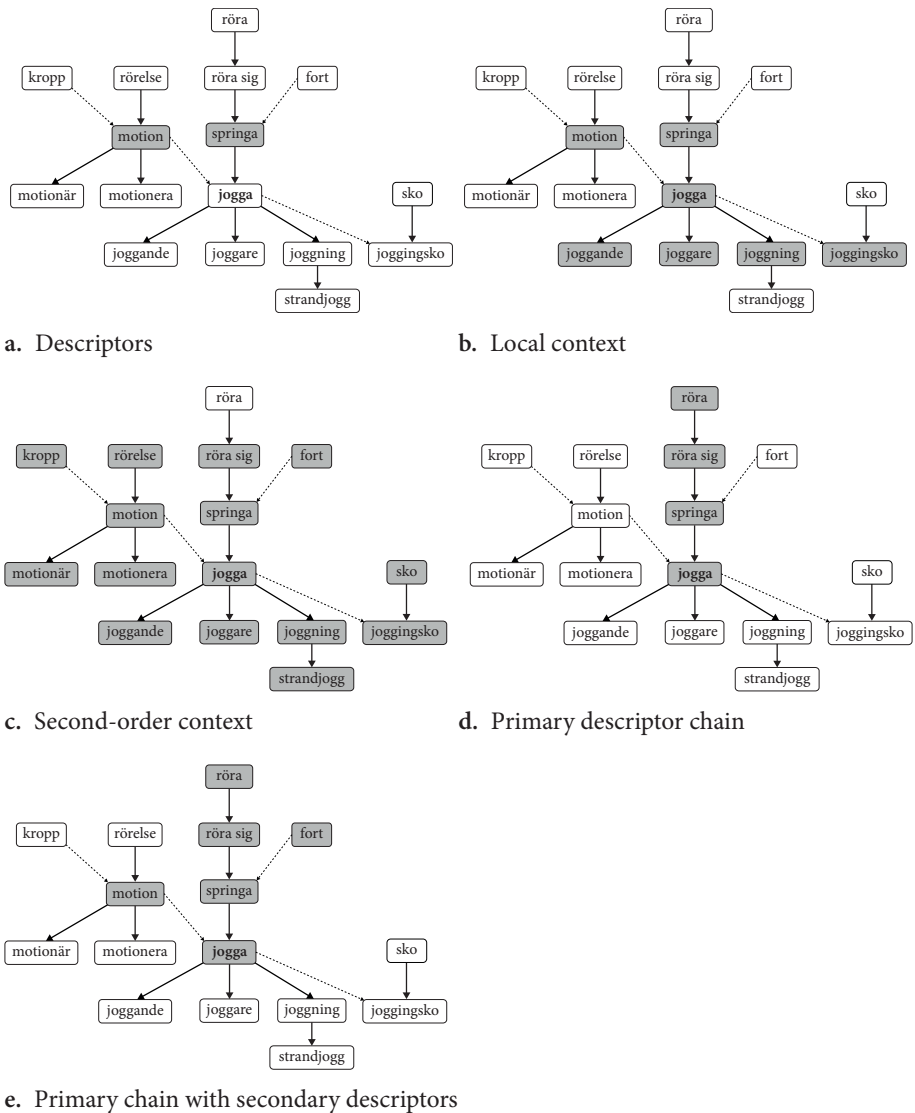
Figure 1 illustrates how the five representations are computed for *jogga*.

Figure 1. Representations of the Saldo context of the word *jogga* ‘to jog’. The shaded words are those included in the result. Solid arrows represent primary descriptor relations and dotted arrows secondary descriptor relations

4. Quantitative evaluation

To compare and analyze the different lexicon expansion methods, we carried out several quantitative evaluations. In these evaluations, we used the Saldo and SweFN versions of November, 2013. This version of SweFN contains 25,989 LUs and 896 frames. We randomly partitioned the verb, noun, adjective, and adverb LUs listed in SweFN into a training set (67%) and a test set (33%). We then removed all frames for which we had fewer than five training LUs, so in the end we had 458 frames in total. The most frequently occurring frames in this set are `Animals` with 520 training instances, `Food` with 496, and `People_by_origin` with 373. There are 39 frames containing exactly five LUs.

There are a number of methodological complications when carrying out quantitative evaluations using this set, which means that we should see the measurements presented in this section as upper bounds; however, we still believe that they are useful for *comparison* of different methods, and similar experimental setups have been used previously (Johansson & Nugues 2007a; Pennacchiotti et al. 2008). These issues arise from the fact that we use a part of the lexicon as a test set: such a set will contain many frequent lemmas, so an evaluation is not necessarily indicative of the performance on unseen lemmas. In reality, the low-hanging fruits will already have been picked by the lexicographers. A related consideration is how we define positive and negative instances. For instance, about 1% of the Saldo lemmas in SweFN are connected to the `Removing` frame, but is this true of the lexicon in general? Can we even be certain that a listed Saldo lemma should not be associated with more frames than it currently is?

4.1 Evaluation metrics

When deciding how to evaluate our methods, we considered two use cases:

1. *Lemma-based classification scenario*: given a Saldo lemma that we haven't seen before, determine the SweFN frames associated with it, if any. In this scenario we must make a hard decision, so for each frame we need to compute a threshold.
2. *Frame-based ranking scenario*: given a SweFN frame and a list of Saldo lemmas, rank all lemmas according to how suitable they are for that frame. This approach may be the most appropriate in a semiautomatic setting involving lexicographers. No thresholds are used in this scenario.

Our methods can be evaluated for both use cases with standard metrics (Manning et al. 2008). The first scenario is a multilabel classification task, so we can apply commonly used classification evaluation metrics such as precision (P) and recall (R)

$$P = \frac{N_{CF}}{N_{GF}} \quad R = \frac{N_C}{N_F}$$

where N_F is the number of LUs evoking the frame F , N_{GF} the number of LUs we have automatically proposed for F , and N_{CF} the number of correctly identified LUs evoking F . Intuitively, if we guess very carefully we get a high precision, and if we guess more aggressively we get a high recall. Both measures have a maximum value of 1.0. The precision and recall values are conventionally presented together with the F -measure, the harmonic mean of the two values.

The second scenario is more similar to information retrieval problems, and it is therefore better to evaluate it using a metric for evaluation of ranking systems (e.g. search engines). There are many such metrics, and we selected the average precision (AP)

$$AP = \frac{\sum_{k=1}^N P_k \cdot I_k}{N_F}$$

where N is the total number of lemmas in the wordlist, P_k the precision if we assign the k top-ranked lemmas to F , and I_k an indicator that is 1 if the lemma at position k evokes F and 0 otherwise. The more the true LUs are concentrated to the top of the ranked output, the higher the AP will be. The maximum value of AP is 1.0, which occurs when all the true LUs are ranked above all the false LUs.

When evaluating the performance of the lexicon expansion methods for all frames, we aggregated over the lexical units (*micro-average*) in the first scenario and over the frames (*macro-average*) in the second. All tables present precision, recall, and mean average precision (MAP) figures.

4.2 Which way is the best to make use of the Saldo lexicon?

Table 1 shows the performance of the LU induction systems using Saldo: classifiers using the five context representations from Section 3.2.1, and the system based on distance in the Saldo graph. It is clear that the best context representations are those that generalize by moving up in the hierarchy, rather than those constructing a semantic field using the Saldo context. The simple primary descriptor chain achieves the highest precision and recall values; adding secondary descriptors improved the performance for some frames with a large number of LUs, such as `Animals`, `Food`, and `Medical_disorders`, but did not have a positive effect on the overall performance.

Table 1. Precision, recall, and mean average precision values for the lexicon-based methods

Method	Precision	Recall	MAP
Descriptors	0.670	0.555	0.455
Local context	0.769	0.490	0.580
Second-order context	0.636	0.627	0.580
Primary descriptor chain	0.789	0.686	0.685
Primary chain with sec. descr.	0.774	0.681	0.685
min Rada distance	0.236	0.767	0.565

Except for the very simple context representation using descriptors only, the classifiers outperform the distance-based system. It is possible that we could improve the performance by considering more complex distance measures (Blanchard et al. 2005), for instance by reweighting graph edges by their distance from the root, but we can also note that the classifiers work as frame-specific weighted similarity measures with automatically learned weights, and it seems unlikely that the end result would be much different.

4.3 Which corpus-based semantic representations are most effective?

We evaluated the classifier-based and similarity-based systems using corpus-based semantic representations presented in Section 2.2, and Table 2 shows the results. The vector and cluster models were trained on the lemmatized dataset described in Section 2.2. All vector representations used 1,024 dimensions, and we had 1,024 Brown clusters.

Table 2. Precision, recall, and mean average precision values for the corpus-based methods

Method	Precision	Recall	MAP
Brown	0.291	0.125	0.051
RI-Pos	0.658	0.209	0.203
RI-CBoW	0.682	0.235	0.242
CC-CBoW	0.622	0.260	0.286
CC-SG	0.641	0.299	0.322
cos-sim CC-CBoW	0.158	0.332	0.271
cos-sim CC-SG	0.366	0.154	0.225

What is most striking here is that all systems considered here perform worse than all Saldo-based systems; a similar result was found by Pennacchiotti et al. (2008). Note that the Saldo-based systems have an advantage in an evaluation taking senses into account: Saldo classifiers operate directly on senses, while the vectors

or clusters express no sense information. For instance, among the 100 Saldo senses ranked most highly for the frame `Animals`, there were eight misclassifications out of which six were due to sense ambiguity.

The best-performing systems are classifiers using representations derived from the contextual classifiers, in particular the skip-gram model. The vectors computed with random indexing are less useful, and the Brown clusters are clearly much too coarse-grained for this task. We still include them in this evaluation since they have been used successfully in a wide range of NLP tasks.

As we saw previously for the lexicon-based systems, classifiers outperform the systems using general-purpose measures, in this case the cosine similarity. The MAP values of the similarity-based systems are comparable to those of the classifiers, but the precision and recall values are worse. We did not include the vectors computed using random indexing in this evaluation since they performed very poorly in preliminary investigations.

Figure 2 shows how the MAP values are affected by the number of dimensions of the vector representations. For completeness, we include the classifiers using Brown clusters in this plot as well; in this case, the horizontal axis corresponds to the number of clusters. In general, a higher dimensionality of the representation resulted in a higher performance, but at least for the best representation (CC-SG) there is hardly any improvement when going from 512 to 1024 dimensions. The RI classifiers could probably be improved slightly, but it should be noted that increasing the dimensionality is computationally costly.

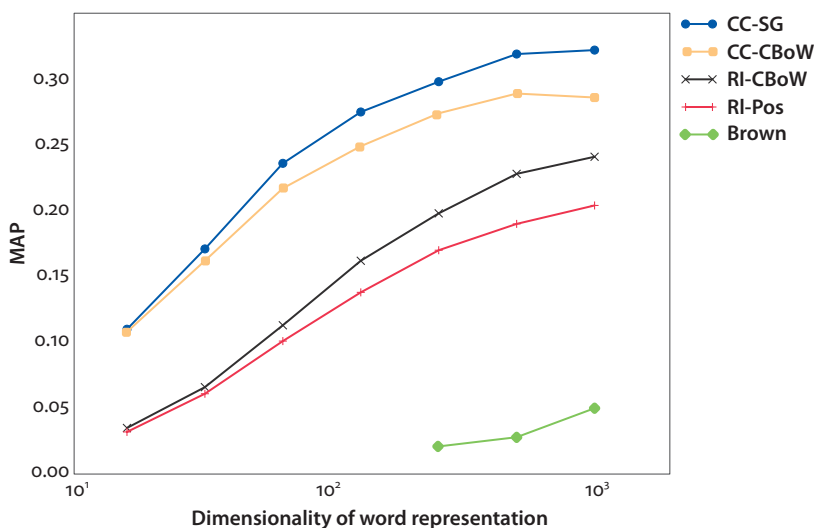


Figure 2. Mean average precision as a function of word representation dimensionality

4.4 Combining lexicon-based and corpus-based classifiers

We finally combined the best feature sets using corpus-based and lexicon-based features (CC-SG and primary descriptor chain, respectively), and we also added a word class (part-of-speech tag) feature. Table 3 shows the performance figures of the combined classifiers on the test set. We observe that the combination of word class, Saldo context representation, and corpus-based semantic representation gives us by far the most effective system out of all we evaluated, so it seems that the representations derived from lexicons and from corpora are complementary to some extent.

Table 3. Precision, recall, and mean average precision values for the combined methods

Method	Precision	Recall	MAP
PoS + PDC	0.811	0.714	0.715
PoS + CC-SG	0.612	0.312	0.316
PoS + PDC + CC-SG	0.863	0.715	0.750

Interestingly, the Saldo-based classifiers gain a lot from adding the word class feature, while the performance of the vector-based classifiers does not improve. This is probably because Saldo is a single graph with no separation between word classes, while the vector representations are probably already well separated by word class.

4.5 For which frames are our methods successful?

There is significant variation among the frames in how successful we are in detecting new LUs. One explanation could be that the number of training LUs varies between the frames, affecting the quality of the classifier. Also, it is possible that the performance is affected by intrinsic properties of the frame: for instance, it might be possible that strongly noun-dominated frames corresponding to physical entities (e.g. Food) would be easy to handle. To study this question, we categorized the frames into four types based on the dominant part-of-speech tag of the LUs. We said that a frame is verb-dominated if at least 25% of its LUs were verbs; otherwise, we selected the most common tag. The reason for the special treatment of verbs is that event-related frames contain verbs as well as their derivations (e.g. nominalizations and agent nouns).

Figure 3 shows how the AP for a frame is affected by the number of training LUs in that frame; the frame types are also shown. As can be seen, frames for which we have much training data generally have better AP, but the correlation is actually not very strong (Pearson $r = 0.2$). We noted that there is much variation of the AP

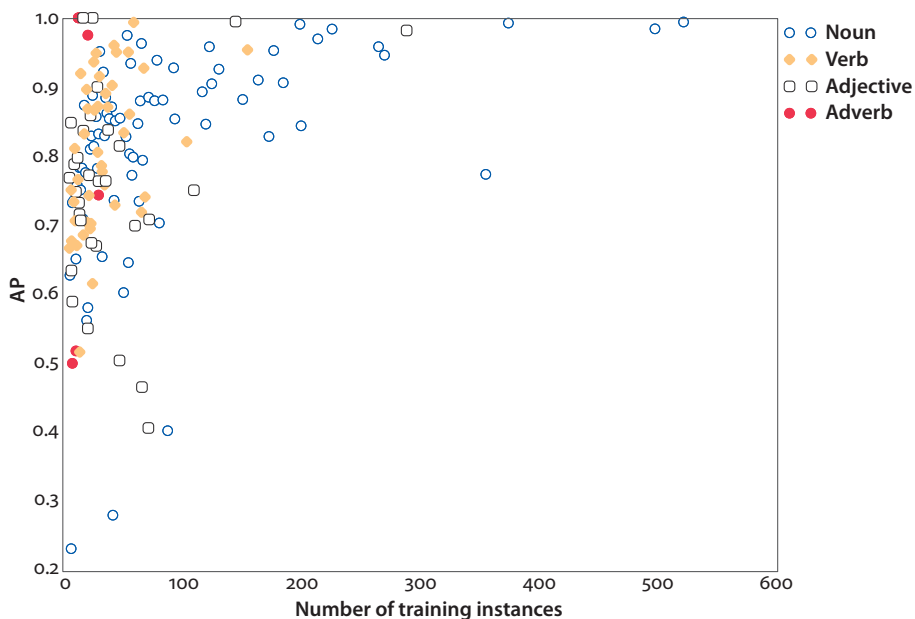


Figure 3. Average precision values for different frames in the LU detection task. Frames are compared with respect to the number of training instances and are divided by the dominant part-of-speech category

for any given frame size. For instance, we measured an AP of 1.0 for 19 of the 39 frames with five training LUs. The frames with the largest number of LUs are mostly noun-dominated frames, and we can see that our system works very well for those frames. However, there does not seem to be a significant difference between the four frame types with respect to the performance as a function of the number of LUs.

Table 4 shows the five frames of each part-of-speech category for which the number of training LUs was highest. As we can see, we are in general successful in frames with a large number of LUs, although there are also less populated frames where our methods work well. In this table, we compare the performance of lexicon-based classifiers with a part-of-speech tag feature (P+L) to classifiers that also include a corpus-based vector representation (P+L+V). We see that in most frames, we get a modest improvement by adding the vectors.

The outlier that we notice in Figure 3 and Table 4 is *People_by_vocation*, a noun-dominated frame containing words such as *lärare* ‘teacher’ and *kock* ‘cook,’ which has a much lower AP than other noun-dominated frames of that size. In particular, the Saldo-based classifier performs very poorly here, and we believe the reason is that the primary descriptor of a vocation noun (and agent nouns in general) is the verb from which it was derived: for instance *lärare* is connected

Table 4. Average precision in the LU detection task for the five largest frames in each part-of-speech category**a. Noun**

Frame	Size	P+L	P+L+V
Animals	520	0.986	0.994
Food	496	0.983	0.984
People_by_origin	373	0.988	0.992
People_by_vocation	354	0.447	0.774
Medical_disorders	269	0.893	0.946

b. Verb

Frame	Size	P+L	P+L+V
Removing	153	0.959	0.954
Make_noise	103	0.772	0.819
Judgment_comm.	68	0.751	0.741
Self_motion	67	0.894	0.928
Experiencer_obj	65	0.665	0.719

c. Adjective

Frame	Size	P+L	P+L+V
Origin	288	0.977	0.982
Color	144	0.994	0.995
Social_int._eval.	109	0.737	0.752
Mental_property	71	0.676	0.708
Emotion_directed	70	0.316	0.407

d. Adverb

Frame	Size	P+L	P+L+V
Frequency	29	0.697	0.744
Time_vector	20	0.975	0.975
Direction	12	0.856	1.000
Degree	10	0.514	0.518
Sufficiency	7	0.500	0.500

to *lära* ‘to teach.’ Because of this, generalization in terms of Saldo is hard since vocation words have little in common and are spread out over the Saldo network. However, the geometric representations seem to capture vocation words better, and the combined system gives us a much improved AP.

4.6 Use by lexicographers

The best-performing system was used to generate suggestions for 7,578 verbs listed in Saldo that were not covered by the SweFN lexicon. 214 of these verbs have been added to SweFN by lexicographers, which gives us a small but more realistic test set. For evaluation purposes, we removed 30 verbs that were assigned to newly created frames.

It can be expected that the LUs that have been added at a late stage of the lexicographic process either occur more rarely in corpora or are harder to classify for lexicographers. These reasons will make the automatic classification harder, so unsurprisingly the results are less impressive when we evaluate in this setting. Of the 184 verbs, the automatic system made a suggestion in 105 cases, out of which 72 were correct: this gives us a precision and recall of 0.686 and 0.391, respectively.

Disregarding the classification threshold and selecting the top-scoring frame suggestion gave an accuracy of 0.527, while the accuracy of the top-five list was 0.717. This result compares well with that reported by Pennacchiotti et al. (2008), who had a top-frame accuracy of 0.25 and a top-ten-frames accuracy of 0.69 for a set of 24 new LUs.

5. Conclusion

We have investigated a number of methods to automatically suggest new lexical units to lexicographers working in the SweFN project. Our systems are applied to word senses in Saldo, a semantic network that is the largest open lexical resource for Swedish. The best-performing system is now used by the lexicographers. We evaluated systems using the Saldo lexicon, using corpus-based representations, and a combination of both types of methods. The best-performing systems were developed using machine learning classification methods, which outperformed methods based on similarity or distance measures such as those used by Pennacchiotti et al. (2008). Classifiers using features describing the neighborhood in Saldo are very effective, which shows that this resource is a viable alternative to WordNet for this task, despite the very significant differences in the underlying theoretical frameworks. The semantic representations from corpora are less effective than the lexicon for this task, but still valuable since they can be applied to words not yet listed in the lexical resource. The combined system outperforms all other systems by a wide margin, which suggests that the corpus-induced semantic representations encode some information that is not available in our lexical resource. In particular, in frames that are hard to describe in terms of Saldo (for instance `People_by_vocation`) the combined system performs very well.

Since the LUs in SweFN are defined in terms of Saldo, our task differs slightly from previous work on adding WordNet senses to FrameNet, which had to deal with sense ambiguity. This probably makes the task easier for lexicon-based classifiers and harder for classifiers using semantic representations derived from corpora. While running the risk of comparing apples and oranges, it is clear that our systems perform well compared to previous approaches. For instance, Pennacchiotti et al. (2008) reported an accuracy of 0.52 for their best system, while our best recall was 0.715. Furthermore, our system performs very well compared to Pennacchiotti et al. (2008) in the more realistic use case of suggesting new lexical units to lexicographers.

Funding

This work was funded by the Swedish Research Council under grant 2013–04944, *Distributional methods to represent the meaning of frames and constructions*, and grant 2012–05738, *Towards a knowledge-based culturomics*.

References

- Blanchard, Emmanuel, Mounira Harzallah, Henri Briand & Pascale Kuntz. 2005. A typology of ontology-based semantic measures. In *EMOI-INTEROP 2005: Proceedings*. Aachen: CEUR-WS.org.
- Boas, Hans C. (ed.). 2009. *Multilingual FrameNets in computational lexicography: Methods and applications*. Berlin: Mouton de Gruyter. <https://doi.org/10.1515/9783110212976>
- Borin, Lars, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj & Dimitrios Kokkinakis. 2010. The past meets the present in Swedish FrameNet++. In *Proceedings of EURALEX 2010*, 269–281. Ljouwert/ Leeuwarden: Fryske Akademy.
- Borin, Lars, Markus Forsberg & Lennart Lönngren. 2013. SALDO: A touch of yin to WordNet's yang. *Language Resources and Evaluation* 47(4): 1191–1211. <https://doi.org/10.1007/s10579-013-9233-4>
- Brown, Peter F., Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra & Jenifer C. Lai. 1992. Class-based n -gram models of natural language. *Computational Linguistics* 18(4): 467–479.
- Burchardt, Aljoscha, Marco Pennacchiotti, Stefan Thater & Manfred Pinkal. 2009. Assessing the impact of frame semantics on textual entailment. *Natural Language Engineering* 15: 527–550. <https://doi.org/10.1017/S1351324909990131>
- Das, Dipanjan & Noah A. Smith. 2011. Semi-supervised frame-semantic parsing for unknown predicates. In *Proceedings of ACL/HLT 2011*, 1435–1444. Portland: ACL.
- Das, Dipanjan & Noah A. Smith. 2012. Graph-based lexicon expansion with sparsity-inducing penalties. In *Proceedings of NAACL/HLT 2012*, 677–687. Montréal: ACL.
- Fan, Rong-En, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang & Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *Journal of Machine Learning Research* 9: 1871–1874.
- Fellbaum, Christiane (ed.). 1998. *WordNet: An electronic lexical database*. Cambridge: MIT Press. <https://doi.org/10.7551/mitpress/7287.001.0001>

- Fillmore, Charles J. & Collin Baker. 2009. A frames approach to semantic analysis. In Bernd Heine & Heiko Narrog (eds.), *The Oxford handbook of linguistic analysis*, 313–339. Oxford: Oxford University Press.
- Friberg Heppin, Karin & Maria Toporowska Gronostaj. 2012. The rocky road towards a Swedish FrameNet – creating SweFN. In *Proceedings of LREC 2012*, 256–261. Istanbul: ELRA.
- Gärdenfors, Peter. 2000. *Conceptual spaces: The geometry of thought*. Cambridge: Bradford Books. <https://doi.org/10.7551/mitpress/2076.001.0001>
- Harris, Zellig. 1954. Distributional structure. *Word* 10(23): 146–162. <https://doi.org/10.1080/00437956.1954.11659520>
- Johansson, Richard. 2012. Non-atomic classification to improve a semantic role labeler for a low-resource language. In *Proceedings of *SEM 2012*, 95–99. Montréal: ACL.
- Johansson, Richard. 2014. Automatic expansion of the Swedish FrameNet lexicon: Comparing and combining lexicon-based and corpus-based methods. *Constructions and Frames* 6(1): 91–112. <https://doi.org/10.1075/cf.6.1.06joh>
- Johansson, Richard, Karin Friberg Heppin & Dimitrios Kokkinakis. 2012. Semantic role labeling with the Swedish FrameNet. In *Proceedings of LREC 2012*, 3697–3700. Istanbul: ELRA.
- Johansson, Richard & Pierre Nugues. 2006. A FrameNet-based semantic role labeler for Swedish. In *Proceedings of Coling/ACL 2006*, 436–443. Sydney: ACL. <https://doi.org/10.3115/1273073.1273130>
- Johansson, Richard & Pierre Nugues. 2007a. LTH: Semantic structure extraction using nonprojective dependency trees. In *Proceedings of SemEval 2007*, 227–230. Prague: ACL. <https://doi.org/10.3115/1621474.1621522>
- Johansson, Richard & Pierre Nugues. 2007b. Using WordNet to extend FrameNet coverage. In *Proceedings of the Nodalida workshop FRAME 2007: Building frame semantics resources for Scandinavian and Baltic languages*, 27–30. Lund: Lund University.
- Kanerva, Pentti, Jan Kristoffersson & Anders Holst. 2000. Random indexing of text samples for latent semantic analysis. In *Proceedings of the 22nd annual conference of the Cognitive Science Society*, 103–106. London: Erlbaum.
- Koo, Terry, Xavier Carreras & Michael Collins. 2008. Simple semi-supervised dependency parsing. In *Proceedings of ACL/HLT 2008*, 595–603. Columbus: ACL.
- Manning, Christopher D., Prabhakar Raghavan & Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511809071>
- Mikolov, Tomáš, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of ICLR 2013: Workshop track*. Scottsdale.
- Mikolov, Tomáš, Wen-tau Yih & Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL/HLT 2013*, 746–751. Atlanta: ACL.
- Mohammad, Saif & Graeme Hirst. 2012. Distributional measures of semantic distance: a survey. *CoRR* abs/1203.1858.
- Padó, Sebastian. 2007. Cross-lingual annotation projection models for role-semantic information. Saarland University. (PhD thesis).
- Palmer, Alexis & Caroline Sporleder. 2010. Evaluating FrameNet-style semantic parsing: the role of coverage gaps in FrameNet. In *Proceedings of Coling 2010: Posters*, 928–936. Beijing: ACL.
- Pennacchiotti, Marco, Diego De Cao, Roberto Basili, Danilo Croce & Michael Roth. 2008. Automatic induction of FrameNet lexical units. In *Proceedings of EMNLP 2008*, 457–465. Honolulu: ACL. <https://doi.org/10.3115/1613715.1613773>

- Rada, R., H. Mili, E. Bicknell & M. Blettner. 1989. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics* 19: 17–30.
<https://doi.org/10.1109/21.24528>
- Tonelli, Sara, Claudio Giuliano & Kateryna Tymoshenko. 2013. Wikipedia-based WSD for multilingual frame annotation. *Artificial Intelligence* 194: 203–221.
<https://doi.org/10.1016/j.artint.2012.06.002>
- Turian, Joseph, Lev-Arie Ratinov & Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of ACL 2010*, 384–394. Uppsala: ACL.
- Velldal, Erik. 2011. Random indexing re-hashed. In *Proceedings of Nodalida 2011*, 224–229. Riga: NEALT.