VETENSKAPSRÅDET
THE SWEDISH RESEARCH COUNCIL

Kansliets noteringar Kod	Dnr
2013-42850-102405-55	2013-4944

2013
Project Grant Junior Researchers

Area of science
Natural and Engineering Sciences
Announced grants
Research grants NT April 11, 2013
Total amount for which applied (kSEK)

2014	2015	2016	2017	2018
583	616	1059	1147	1255

APPLICANT

Name(Last name, First name)
Johansson, Richard

Date of birth
750709-2034

Gender
Male

Email address
richard.johansson@svenska.gu.se

Academic title
PhD (engi)

Position
Forskare

Phone
031-7864418

Doctoral degree awarded (yyyy-mm-dd)
2008-12-05

WORKING ADDRESS

University/corresponding, Department, Section/Unit, Address, etc.
Göteborgs universitet
Institutionen för svenska språket
Språkbanken
Box 200
40530 Göteborg, Sweden

ADMINISTRATING ORGANISATION

Administrating Organisation
Göteborgs universitet

DESCRIPTIVE DATA

Project title, Swedish (max 200 char)
Distributionella metoder för ramsemantik och konstruktionsgrammatik

Project title, English (max 200 char)
Distributional Methods to Represent the Meaning of Frames and Constructions

Abstract (max 1500 char)
The project aims to find automatic, corpus-based methods for inducing linguistic constructions and semantic frames, and representing their meaning using distributional semantics. In addition, the project will study the interaction between the automatically induced meaning representations and symbolic, knowledge-based resources such as frame and construction inventories, and use the representations in natural language processing (NLP) tools. It will combine two recent developments in unsupervised NLP: distributional methods for building and processing geometric meaning representations from corpora, and unsupervised semantic frame and role induction.

The results of the project will advance research in NLP and have practical benefits in applications: Corpus-induced semantic representations will be able to move beyond single words, and be formalized in terms of frame semantics and construction linguistics. Automatic syntactic and semantic analysis tools can be made more robust since they can use linguistic information beyond the word level. Linguistic resource building will benefit by the automatic methods for construction and frame discovery that the project will devise. NLP applications such as information extraction, opinion mining, grammar checking, and computer-assisted language learning can integrate the semantic frames and linguistic constructions discovered by the project, and use their distributional representations to understand their meaning.

VETENSKAPSRÅDET
THE SWEDISH RESEARCH COUNCIL

Kod
2013-42850-102405-55

Name of Applicant
Johansson, Richard

Date of birth
750709-2034

Abstract language
English

Keywords
språkteknologi, syntax, semantik, korpusar, maskininlärning

Research areas
*Nat-Tek generellt

Review panel
NT-2

Classification codes (SCB) in order of priority
10208

Aspects

Application is also submitted to

similar to:	**identical to:**

ANIMAL STUDIES

Animal studies
No animal experiments

OTHER CO-WORKER

Name(Last name, First name)	**University/corresponding, Department, Section/Unit, Addressetc.**
,	
Date of birth	**Gender**
Academic title	**Doctoral degree awarded (yyyy-mm-dd)**

Name(Last name, First name)	**University/corresponding, Department, Section/Unit, Addressetc.**
,	
Date of birth	**Gender**
Academic title	**Doctoral degree awarded (yyyy-mm-dd)**

Name(Last name, First name)	**University/corresponding, Department, Section/Unit, Addressetc.**
,	
Date of birth	**Gender**
Academic title	**Doctoral degree awarded (yyyy-mm-dd)**

Name(Last name, First name)	**University/corresponding, Department, Section/Unit, Addressetc.**
,	
Date of birth	**Gender**
Academic title	**Doctoral degree awarded (yyyy-mm-dd)**

VETENSKAPSRÅDET
THE SWEDISH RESEARCH COUNCIL

Kod
2013-42850-102405-55

Name of Applicant
Johansson, Richard

Date of birth
750709-2034

ENCLOSED APPENDICES

A, B, C, N, S

APPLIED FUNDING: THIS APPLICATION

Funding period (planned start and end date)
2014-01-01 -- 2018-12-31

Staff/ salaries (kSEK)

Main applicant	**% of full time in the project**	**2014**	**2015**	**2016**	**2017**	**2018**
Richard Johansson	40	415	418	430	443	511
Other staff						
Doktorand	55	137	138	597	643	711
Total, salaries (kSEK):		552	556	1027	1086	1222

Other projectrelated costs (kSek)	**2014**	**2015**	**2016**	**2017**	**2018**
Resor	28	57	29	58	30
Förbrukningsmaterial	3	3	3	3	3
Total, other costs (kSEK):	31	60	32	61	33

Total amount for which applied (kSEK)

2014	2015	2016	2017	2018
583	616	1059	1147	1255

ALL FUNDING

Other VR-projects (granted and applied) by the applicant and co-workers, if applic. (kSEK)

Funds received by the applicant from other funding sources, incl ALF-grant (kSEK)

POPULAR SCIENCE DESCRIPTION

Popularscience heading and description (max 4500 char)

De senaste åren har det skett stora framsteg inom konsten att tillverka datorprogram som analyserar grammatiken i meningar skrivna av människor, t.ex. på svenska. Dessa grammatikprogram är förstås användbara inom språkvetenskaplig forskning, men de är också mycket viktiga i praktiskt användbara datortillämpningar som används utanför akademin. Exempel på detta är hjälpmedel för automatisk översättning och stavningskontroll, eller de program som företag kan använda för att överblicka recensioner på nätet av sina produkter.

Dessa datorprogram har stora problem med att hänga med i språkets utveckling och hantera språkets stora variation. Språket är en organism som utvecklas snabbare och i fler riktningar än vad som är möjligt för språkvetenskapen att fullständigt beskriva. Nya ord och grammatiska fenomen uppstår och gamla tas ur bruk, och olika texttyper beter sig på olika sätt. Till exempel det språk som förekommer i svenskspråkig text på websidor och i bloggar uppvisar ofta grammatiska egenheter som inte finns beskrivna i standardiserade grammatiska beskrivningar som Svenska Akademiens

Grammatik. Detsamma gäller texter som kommer från specialistsammanhang, t.ex. avancerad vetenskaplig prosa. Hur kan man då hantera problemen med språkets heterogenitet och snabba utveckling om man vill bygga datorprogram som hanterar språk? Finns det något sätt att låta datorn själv studera texter och lära sig deras grammatik i stället för att beskriva grammatiken för hand?

För att studera detta forskningsproblem kommer vi att använda distributionella metoder. Denna idé innebär att man använder geometri för att studera ords betydelser och grammatiska beteeenden. Ordens geometri konstruerar man genom att låta en dator gå igenom stora mängder text och observera orden i de textsammanhang de förekommer. Om två ord förekommer i liknande sammanhang antas de vara språkligt besläktade och de kommer då också att ligga nära varann geometriskt. Vi kan då t.ex. mäta att "pizza" och "hamburgare" är två ord som har närbesläktade betydelser, eller att de två verben "uppmuntra" och "övertala" beter sig på liknande sätt ur ett grammatiskt perspektiv. Eftersom projektets syfte är att studera grammatik så kommer vi att behöva utöka de distributionella metoderna så att de kan hantera inte bara enstaka ord utan också ordkombinationer och mer abstrakta mönster, så kallade grammatiska konstruktioner.

För att tillämpa de distributionella metoderna kommer vi att datorbehandla de mycket stora samlingar av svensk text som under många år insamlats på Språkbanken vid Göteborgs universitet. Denna textsamling innehåller över en miljard ord och utökas ständigt med nytt textmaterial. Den innehåller texter av många olika typer och från många olika tidsperioder, t.ex. romaner, nyheter, vetenskapliga artiklar, medeltida lagtext samt texter från sociala medier som bloggar och Twitter.

VETENSKAPSRÅDET
THE SWEDISH RESEARCH COUNCIL

Kod
2013-42850-102405-55

Name of applicant
Johansson, Richard

Date of birth
750709-2034

Title of research programme
Distributional Methods to Represent the Meaning of Frames and Constructions

Appendix A

Research programme

Vetenskapsrådet, Box 1035, SE-101 38 Stockholm, tel. +46 (0)8 546 44 000, vetenskapsradet@vr.se

Research Program:
Distributional Methods to Represent the Meaning of Frames and Constructions

1 Purpose and Aims

This project aims to find automatic, corpus-based methods for inducing *linguistic constructions* and *semantic frames*, and representing their meaning using distributional semantics. In addition, the project will study the interaction between the automatically induced meaning representations and symbolic, knowledge-based resources such as frame and construction inventories, and use the meaning representations in natural language processing (NLP) systems.

It will build on and cross-pollinate two recent important developments in unsupervised NLP:

- **Distributional methods** for building and processing geometric meaning representations from corpora. While this is a very well-established idea in NLP and IR, which has seen much use in applications, an important step has been taken in the last few years with the introduction of models that handle compositionality. However, these models currently cannot account for the meaning of larger units: constructions and frames, and their relation with lexical resources such as FrameNet is understudied.
- **Automatic semantic frame induction**. A number of corpus-based methods have been proposed recently, typically using LDA-derived learning methods. This work has so far been limited to single words, typically verbs: more complex linguistic units such as constructions are not handled. Furthermore, no extrinsic evaluation of the usability of the proposed frames has been carried out.

The results of the project will advance the research agenda in NLP and have practical benefits in a number of applications:

- **Corpus-induced semantic representations** will be able to move beyond single words, and be formalized in terms of frame semantics and construction linguistics.
- **Automatic syntactic and semantic analysis** tools for Swedish and other languages can be made more robust since they can use linguistic information beyond the word level.
- **Linguistic resource building** at the Swedish Language Bank and elsewhere will benefit by the automatic methods for construction and frame discovery that the project will devise.
- **Practical NLP applications** such as information extraction, opinion mining, grammar checking, and computer-assisted language learning can integrate the semantic frames and linguistic constructions discovered by the project, and make use of their distributional representations to understand their meaning.

To exemplify the goals we aim for, we consider a direct speech construction that is very common in Swedish informal language, in particular among young speakers. In this construction, the speech verb is omitted, and an adverb replaces it as a direct speech marker. By far, the most commonly used adverb in this situation is *ba*, derived from *bara* 'just', but sometimes other adverbs such as *typ* 'like' are used instead. This has some similarity to the American English *like* direct speech construction. The following two texts are written by young Swedish bloggers:

(1) [...] *mamma fråga ifall det gick att köpa dom nu och hon ba aah vilken storlek, jag ba, XS, hon ba nej dom är tyvärr slut och ger mig värsta hånfulla smilet!*

[...] mum asked if they could be bought now and she's like aah which size, I'm like, XS, she's like no they are out of stock and gives me the worst smirk!

(2) *Jag var under vatten sedan för Moa kom och typ "ÄR DU RÄDD FÖR VATTEN DIN LILLA PUSSY??" och jag typ "eh ja...." så blev hon sur* [...]

> Then I was under water because Moa came and was like "ARE YOU AFRAID OF WATER YOU LITTLE PUSSY??" and I'm like "errh yes..." then she was cross [...]

Handling this direct speech construction properly is of course meaningful for NLP systems focusing on spoken modern language or social media. However, we will have little success regardless of whether we use knowledge-based methods with hand-built grammars and lexicons, or corpus-driven methods: the construction is not described in standard grammars or lexicons since it occurs in informal language only, and while distributional methods in their current form can discern it indirectly, they only give an incomplete view restricted to the word level.

For instance, we applied a method similar to Sahlgren (2006) on a large collection of Swedish blog text, and examined the word *ba*, the adverb most commonly used in the direct speech construction. The five words closest to *ba* in a vector space, excluding spelling variants, were *utbrast* 'exclaimed', *utbrister* 'exclaims', *utropade* 'cried out', *hojtade*, 'shouted', and *svarade*, 'answered.' We are thus able to see the direct speech function of *ba*, but not the fact that it participates in a larger templatic unit where other adverbs can be used as well.

This example illustrates the goals of the project and the open research problems it tries to address: developing automatic, data-driven methods that

- *discover* this direct speech construction and suggest it to linguists for possible inclusion in construction inventories;
- *syntactically analyze* the construction on an appropriate level of generalization, and infer a construction pattern such as NP (*ba* | *typ*) DIRECTSPEECH;
- *semantically analyze* the construction in terms of distributional vector space semantics;
- *relate* the meaning of this construction to other direct speech constructions, as well as speech verbs such as *say*, *exclaim*;
- *link* the construction to a structured meaning representation such as FrameNet, e.g. that the construction corresponds to the STATEMENT frame, the NP to the frame element SPEAKER, and the DIRECTSPEECH to the frame elemtent MESSAGE;

The discovered construction can then be integrated in information extraction systems that extract opinions and statements expressed in the social web.

2 Survey of the Field

2.1 Construction linguistics

Theories involving constructions have recently revitalized the research agenda in theoretical syntax, lexicography, and argument structure theory (Fillmore et al., 1988; Goldberg, 1995). They can be seen as a principled but pragmatic way to handle the large grey zone between the realm of words as seen by lexicographers, and the realm of abstract grammatical rules as seen by syntacticians. This allows linguists to systematically describe formal idioms, syntactic patterns that are grammatically irregular with regard to either their interpretation or their syntactic composition. An example of a syntactically irregular formal idiom that has been discussed in the construction-linguistic literature is the correlative conditional or *the-the* construction, e.g., *The bigger they come, the harder they fall*. While this construction has conditional semantics, it does not fit into any other phrase-structure pattern for conditionals in English. Conversely, an example of a syntactically regular but semantically irregular formal idiom is the WXDY construction, e.g., *What's that fly doing in my soup?* Constructions are relevant in research on language learning: while a firm grasp of constructions is necessary for fluency when learning a language, they are often neglected in current language teaching approaches, leading to unidiomatic language production by the learner (Prentice and Sköldberg, 2011).

In modern large-scale and data-driven NLP, construction linguistics have so far had very little impact. In discourse and attitude analysis, Greene and Resnik (2009) showed that a speaker's attitude can be predicted from construction features such as the selection of a transitive or intransitive verb frame; a similar intuition has also been used for building practical sentiment analysis systems (Karlgren et al., 2010).

2.2 Frame and role semantics, and automatic semantic role analysis

Many types of information extraction tasks can be seen as instances of a more general problem: tagging events that are mentioned in a text, and determining which are the participants involved and the circumstances that surround the event. *Semantic roles* correspond to the semantic relations constituents in a sentence have with the event. For instance, Example 3 shows three sentences all describing the event that Microsoft (with the semantic role BUYER) bought the company Powerset (the GOODS). The fact that the semantic role structures in the three sentences are isomorphic shows us that they have the same meaning.

(3a) [Microsoft]$_{\text{BUYER}}$ ***bought*** [Powerset]$_{\text{GOODS}}$

(3b) [Powerset]$_{\text{GOODS}}$ was ***acquired*** [by Microsoft]$_{\text{BUYER}}$

(3c) [Microsoft's]$_{\text{BUYER}}$ ***purchase*** [of Powerset]$_{\text{GOODS}}$

This generalization is appealing and automatically analyzing the semantic role structure, or *semantic role labeling* (SRL), has proven useful in numerous NLP applications such as information extraction (Johansson et al. (2005), Christensen et al. (2011), *inter alia*) and opinion mining (Johansson and Moschitti, 2013).

FrameNet (Baker et al., 1998) is a lexical resource for English built according to the theory of frame semantics (Fillmore, 1976). It defines prototypical event types (*frames*) with corresponding semantic roles, and also defines the words corresponding to each event type (*lexical units*). For instance, the COMMERCE_BUY frame defines the BUYER and SELLER semantic roles, and lists several lexical units including *buy*, *acquire*, and *purchase*. The FrameNet release also includes an annotated corpus and a large collection of annotated examples.

Lexicon and corpus resources such as FrameNet makes it possible to build automatic SRL systems. Since the pioneering work by Gildea and Jurafsky (2002), there have been numerous FrameNet SRL systems described in literature (e.g. Johansson and Nugues (2007b) and Das et al. (2013)). A more theory-light alternative to FrameNet is PropBank (Palmer et al., 2005), which has proven easier to process for automatic systems (Johansson and Nugues, 2008a; Màrquez et al., 2008). SRL performance is tightly coupled to the problem of syntactic representation (Johansson and Nugues, 2008b) and it can therefore be effective to carry out syntactic and role-semantic analysis jointly (Johansson, 2009).

While most work has been carried out for English, there have been efforts in other languages as well (Hajič et al., 2009). For Swedish, the recent effort to build a FrameNet (Borin et al., 2010) has made it possible to construct SRL systems (Johansson et al., 2012; Johansson, 2012). Resources are limited for languages other than English, so system performances can be improved by semi-supervised learning and cross-lingual transfer (Johansson and Nugues, 2006).

2.3 Distributional semantics

While traditional research in artificial intelligence and cognitive science, and also lexical resources such as FrameNet, has tended to view concepts as symbols and the brain as a symbol-processing unit, a recent alternative defines the meaning of a concept as a point or region in a vector space. This has the advantage that the notion of graded similarity becomes natural: two concepts whose meanings are similar are associated with a pair of vectors that are close in

the vector space as defined by some distance metric. For instance, Gärdenfors (2000) proposed the notion of *conceptual spaces*; in his work, the dimensions of the space are assumed to be meaningfully interpretable by the human senses, e.g. color, weight, taste, time, etc.

In NLP, vector space representations of meaning are usually induced automatically from corpora. There are many methods to construct these vector spaces, but the essential underlying idea of all approaches is that words with similar meanings occur in similar linguistic contexts, e.g. in a similar set of documents or surrounding words: this is usually referred to as the *distributional hypothesis* (Harris, 1954). Distributional semantics is an appealing alternative or complement to manually defined meaning representations: in particular, it can be useful for dealing with the fluid and evolving nature of language, and for handling non-standard language varieties such as encountered in social media.

The most well-known approaches for inducing vector space representations are based on dimensionality reduction of matrices, e.g. Latent Semantic Analysis (LSA) (Deerwester et al., 1990) and Random Indexing (Sahlgren, 2006). An alternative to matrix-based methods is Latent Dirichlet Allocation (LDA), which is based on hierarchical Bayesian modeling (Blei et al., 2003). Due to its mathematical elegance and firm foundation in statistical theory, LDA and its many derivatives are increasingly popular for a very wide range of NLP tasks such as unsupervised WSD (Brody and Lapata, 2009). Semantic vector representations derived from neural models are also becoming influential (Collobert and Weston, 2007; Socher et al., 2012).

Classical models of distributional semantics have been restricted to representing the meaning of single words. However, in recent years there has been a growing interest in trying to account for the way we combine small units of meaning into larger units: *compositionality* (Mitchell and Lapata, 2010). In these models, there is typically a division between primitive concepts represented as vectors, and compositional functors represented geometrically as linear operators (matrices or tensors) that transform vectors. The primitive concepts can correspond e.g. to nouns and the compositional ones to verbs and adjectives (Baroni and Zamparelli, 2010).

Corpus-induced semantic representations are often used for building machine learning features that are more robust than the lexical features normally used in NLP (Turian et al., 2010). Such representations have been used in many applications, including parsing (Koo et al., 2008), FrameNet-based SRL (Johansson et al., 2012; Johansson, 2012), and sentiment analysis (Günther and Johansson, 2013). Compositional vector space semantics has been used to improve applications such as sentiment analysis and relation extraction (Socher et al., 2012).

2.4 Automatic induction of symbolic representations

Structured linguistic representations can be extracted from corpora even if they are of a symbolic nature and not distributional. For instance, there are methods for automatically expanding linguistic resources such as WordNet and FrameNet. Johansson and Nugues (2007c) described a method to expand the FrameNet lexicon, which gives automatic FrameNet analysis systems a better coverage on a wider range of texts.

In recent years, a number of ideas have been proposed to automatically induce FrameNet-like semantic frames from corpora. These ideas can be divided into two broad groups. The first group takes a top-down approach and understands semantic frames as narrative schemas or scripts (Chambers and Jurafsky, 2009), and the second group takes a bottom-up approach and views frames as clusters of words with similar syntactic–semantic linking behavior (Modi et al., 2012; Materna, 2012). Furthermore, there are approaches to automatically induce semantic roles, also outside the frame-semantic paradigm (Lang and Lapata, 2010).

There has been some tentative work on automatic discovery of constructions, see e.g. the workshop by Sahlgren and Knutsson (2010), but this work has so far not been related to the the-

ory of frame semantics or resources such as FrameNet, even though the notion of grammatical construction is becoming central to the theory of semantic frames (Fillmore et al., 2011).

3 Project Description

3.1 Method

We have organized the project work into three broad thematic areas based on the most important ideas of the project. The central task is to derive new distributional representations for constructions and frames, which then leads naturally to a number of spinoffs: using the new semantic representations to automatically induce complex linguistic units; integrating them into syntactic and frame-semantic analysis tools and practical NLP applications such as sentiment analysis.

Area 1: Corpus-induced representations for frames and constructions

As mentioned above, the core idea and the main goal of the project is to build corpus-induced semantic representations for complex linguistic objects such as constructions and frames. We will apply a wide range of learning methods when inducing the representations; however, the main focus will be to adapt two learning paradigms that have proven successful in recent research: the first is hierachical Bayesian modeling similar to the popular LDA approach (Blei et al., 2003), which have been applied in complex linguistic tasks such as unsupervised word sense disambiguation (Brody and Lapata, 2009) and frame induction (Modi et al., 2012; Materna, 2012); the second is neural or "deep-learning" modeling, which has recently been used to derive widely used word representations (Turian et al., 2010) and predictors for a wide range of NLP tasks (Collobert and Weston, 2007). They have recently been extended to handle compositionality, with applications such as sentiment analysis (Socher et al., 2012).

Furthermore, the project will explore the interaction between automatically induced linguistic representations and those defined by linguistic experts and encoded in resources such as FrameNet (Baker et al., 1998), in particular the Swedish FrameNet (Borin et al., 2010), and inventories of grammatical constructions such as the Swedish Constructicon (Lyngfelt et al., 2012). For instance, can we give geometric interpretations of frame-to-frame relations in FrameNet such as INHERITANCE or SUBFRAME?

The methods developed in the project will be generally applicable, but our primary target language will be Swedish. This makes the availability of a large amount of Swedish text crucial to the project. The Language Bank continuously collects texts written in Swedish, and makes them available for linguistically informed search via the Korp web service (Borin et al., 2012)[1]; this collection contains a very wide range of texts from many genres, including fiction, news, and specialized domains such as medicine. The collection also has a temporal dimension: it contains medieval texts about law and religion as well as texts taken from social media such as blogs and Twitter. In addition to the Swedish data, we will also replicate previous research using English text such as the Reuters corpus or the British National Corpus.

Area 2: Construction and frame discovery

The development of vector space semantical representations in Area 1 makes it possible to discover constructions and frames from corpora. There are several opportunities for developing discovery processes. Among the most important ones are anomaly detection, geometric operator clustering, compression, and structure pattern mining.

Anomaly detection. Similar to what has previously been done for multiword units, constructions can be discovered by detecting anomalies in the vector spaces, where we can measure that a complex unit has distributional properties that do not follow from its parts.

[1] http://spraakbanken.gu.se/korp

Clustering geometric operators. Area 1 describes frames and constructions in terms of geometric regions and operators in a distributional space; a possible discovery process can be derived by clustering linguistic units with a similar behavior as geometric operators. This can be seen as a geometric interpretation of previous methods for frame-semantic discovery, which have used hierarchical Bayesian methods to cluster words based on their syntactic–semantic linking properties (Modi et al., 2012; Materna, 2012), and is also related to the compositional models for sentiment analysis by Socher et al. (2012).

Compression. One alternative is to induce constructions by applying compression. Given a large corpus, unprocessed or syntactically parsed, we could develop compression-inspired methods to find frequent patterns that simplify the description of the corpus. This is similar to grammatical inference methods based on Minimal Description Length (Kohonen et al., 2009). The new methods could use geometric distance in addition to symbolic distance measures (e.g. edit distance) as a way to introduce generalization.

Pattern mining. A further alternative is to mine for features in kernel-induced spaces. Convolution kernels implicitly represent large feature spaces derived from structured objects such as trees or sequences. Newly developed methods allow us to explore the implicit feature spaces and find the most informative features, and they have been used to extract informative patterns (Pighin and Moschitti, 2009). Again, the challenge in adapting these methods is to measure similarity in the semantic space in addition to the symbolic space.

The methods developed here can be exploited and evaluated by lexicographers and construction linguists at the Department of Swedish.

Area 3: Integration into NLP applications

The distributional representations developed in Area 1 and the automatically extracted constructions and frames from Area 2 can be evaluated *extrinsically* in NLP applications. The applications we primarily intend to use as testbeds are syntactic dependency parsing and semantic role labeling, which are already used in the infrastructure of the Language Bank.

Corpus-induced representations of words have been shown to improve syntactic parsers (Koo et al., 2008) as well as semantic role labelers (Johansson et al., 2012; Johansson, 2012). However, these representations have been limited to single-word units, and the project will develop methods to integrate complex linguistic objects (automatically extracted in Area 2) into the parsing process, and integrate their distributional representations (developed in Area 1) into the feature representations used in the machine learning models.

Linguistic constructions have a syntactic and semantic side, and this will lead to complex interactions between syntactic structure (constituents, dependencies) and semantic structure (semantic frames and roles); we will adapt algorithms by Johansson (2009) to handle the algorithmic problem of jointly maximizing syntactic and semantic scoring functions.

In addition to syntactic and semantic parsing, we will plug the semantic representations into other NLP applications with a more practical than linguistic purpose. An important example of such an application is opinion mining, where semantic structure as well as syntactic constructions have been shown to be informative (Johansson and Moschitti, 2013; Greene and Resnik, 2009; Karlgren et al., 2010); this application could benefit by correlating sentiment values to subspaces in a vectorial representation (Socher et al., 2012) or subclusters in an automatically induced taxonomy (Günther and Johansson, 2013), and by improved coverage when automatically extracted constructions and frames are used. Another example is computer-assisted language learning, where constructions are important for learners (Prentice and Sköldberg, 2011).

3.2 Personnel

The research team will consist of the principal investigator (PI) and a PhD student.

The principal investigator. Richard Johansson is currently a researcher at the University of Gothenburg. He is a cross-disciplinary researcher whose main interest is statistical NLP, but whose work has often been influenced by questions arising in theoretical linguistics. His most influential contributions to the field of NLP has been in the following areas, which are all important to this project:

- Machine learning models for shallow semantic analysis (Johansson and Nugues, 2008a; Johansson, 2012). He was among the first to stress the importance of syntactic dependencies in statistical models of frame semantics (Johansson and Nugues, 2007b).
- The question of choosing the most useful design of syntactic representation, and how syntax influences the automatic semantic interpretation (Johansson and Nugues, 2007a). He carried out one of the first extrinsic evaluations of the usefulness of syntactic parsers (using different underlying syntactic representation theories) for a downstream application (Johansson and Nugues, 2008b).
- Multilinguality of resources, in particular the automatic production of semantic lexicons and annotated corpora for new languages. He developed methods for transferring semantic structure annotation from English to other languages by means of parallel corpora (Johansson and Nugues, 2006). He organized large-scale evaluations of syntactic and semantic parsers, one for English and one for multiple languages (Hajič et al., 2009).
- The algorithmic question of how to design inference algorithms for statistical models where multiple representation layers interact (Johansson, 2009). This is crucial if we introduce constructions as a new layer of processing on top of basic dependencies.
- Application of shallow semantic extraction techniques in practical NLP applications such as information extraction and opinion mining (Johansson and Moschitti, 2013). In particular, his work in opinion mining relied on the idea that the selection of constructions in text encode the writer's attitudes and emotions.

The PI will lead the project intellectually and practically, supervise the PhD student, and coordinate the dissemination efforts. In the initial stages of the project, the student will use the research software frameworks already developed by the PI. The PI will work 40% per year.

PhD student. Most of the implementation will be carried out by a PhD student, who will be hired by the project and partially funded directly by the Department of Swedish. The student will work out the details of linguistic representations and statistical models defined by the PI, and carry out the burden of the experimental work. We will try to recruit a generalist with linguistic knowledge as well as NLP experience, ideally with knowledge of distributional methods. The PhD student will work 20% in the first two years and 80% in the remaining years.

3.3 Time Plan

The project is planned to last from the start of January 2014 to the end of December 2018. The following table shows a rough plan of how we believe that the effort will be distributed among the four research areas. In the table, a filled circle (●) means active research work, while an empty circle (○) means planning and surveying work.

Year	Area 1	Area 2	Area 3
2014	●	○	
2015	●	●	○
2016	●	●	●
2017	●	●	●
2018	●		●

Since Area 1 can be regarded as the core of the project and the other areas rely on its results, it is important that work starts there as soon as possible. The major part of the project effort is carried out in the later parts of the project, when the PhD student has finished the mandatory course part of the doctoral program.

4 Significance

We believe the results of the project will advance the field in several research areas in NLP:

Distributional semantics. While corpus-induced semantics has made very significant advances recently with the introduction of compositionality and new learning methods, these models currently cannot account for the meaning of larger linguistic units such as constructions and frames, and their relation with lexical resources such as FrameNet is not well understood.

Building linguistic resources. The project will devise new methods to discover semantic frames and syntactic constructions, and these methods can be applied and evaluated by lexicographers and syntacticians. For instance, the Language Bank and the Department of Swedish are developing frame-semantic and construction grammar resources where the automatic discovery tools can be applied.

Automatic syntactic and semantic analysis. The project will improve syntactic and semantic parsers in two ways. First, while it is well known that corpus-induced representations improve syntactic (Koo et al., 2008) as well as semantic (Johansson et al., 2012) analysis tools, so far these representations have been limited to single words only. Secondly, since it is widely recognized that current frame-semantic analysis tools suffer from a low coverage, automatic construction and frame discovery methods may help to make these tools more robust, in particular when applied to nonstandard text genres such as social media.

Applications. There are several types of NLP applications that could benefit from analyzing text in terms of constructions and representing their meaning with distributional semantics: information extraction could gain from improved frame-semantic analysis (Johansson et al., 2005); opinion mining is a rapidly expanding area that could benefit since it is known that attitudes and emotions influence the choice of constructions (Greene and Resnik, 2009; Karlgren et al., 2010; Johansson and Moschitti, 2013); grammar checking and computer-assisted language learning software are other examples. These applications could also extrinsically evaluate the syntactic and semantic analysis tools produced in the project.

Multilinguality. Almost all previous work in distributional semantics and linguistic representation induction has been built and evaluated using English data only. The application to Swedish will demonstrate the wider applicability of these methods in languages with a lower resource availability.

To ensure that the results are disseminated and remain widely used after the end of the project, the software developed in the project will be published under an open source license.

5 Preliminary Results

The experimental platform to be used in the project will be based on previous work developed during several years. This platform includes NLP tools such as dependency parsers, frame-semantic analysis tools, and an opinion extractor. Using these tools, we have already achieved a number of preliminary results relevant to the project:

- Johansson and Nugues (2007c) presented a method automatically inducing new lexical units in FrameNet. The expanded lexicon led to increased coverage in a frame-semantic analysis system (Johansson and Nugues, 2007b).
- Johansson et al. (2012) and Johansson (2012) showed the utility of corpus-induced representations for Swedish frame-semantic analysis, and Günther and Johansson (2013) for sentiment analysis in social media text.
- Johansson (2013) showed that simple construction-like features, extracted automatically, make it possible to train parsers on treebanks using different annotation formalisms.
- Johansson and Moschitti (2013) used role-semantic and construction patterns to improve a system for the extraction of opinions and their holders.

6 The Research Environment at the Language Bank

The team will be employed at the Language Bank (*Språkbanken*) at the Department of Swedish at the University of Gothenburg. The department fits the project well since it has specialists in language resource building and NLP system implementation as well as construction-oriented theoretical linguists, and there will be synergies with ongoing projects at the department (Borin et al., 2010; Lyngfelt et al., 2012). Furthermore, the corpus collection hosted by the Language Bank will be very important for the project since it relies on large corpora.

NLP is one of eight focus areas (*styrkeområden*) considered particularly important by the University of Gothenburg. The focus area is realized practically as the Centre for Language Technology, a network consisting of research groups in several departments at the University of Gothenburg as well as the Chalmers University of Technology.

The Language Bank follows an explicit policy of openness of resources and software. The software will be published under an open-source license, and annotated corpora will be freely downloadable when possible; when we have to work with proprietary material, we will still make our annotation layers publicly available in separate files. The publications resulting from the research will be published in open-access conference proceedings and journals.

7 International collaborations

The PI has ongoing collaborations with groups at the University of Trento, Italy. Primarily the PI collaborates with the group of Alessandro Moschitti, a well-known expert on machine learning and in particular convolution kernels to handle large feature spaces. In the project, expertise from Moschitti's group will be used in the development of construction discovery methods, where we will use kernels and feature space analysis methods (Pighin and Moschitti, 2009). In addition, the Department of Swedish collaborates with Fillmore's group that develops the English construction and frame-semantic lexicon and corpora (Baker et al., 1998).

References

C. Baker, C. Fillmore, and J. Lowe. 1998. The Berkeley FrameNet project. *ACL/COLING*.

M. Baroni and R. Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. *EMNLP*.

D. Blei, A. Ng, and M. Jordan. 2003. Latent Dirichlet allocation. *J. Machine Learning Research*, 3.

L. Borin, D. Dannélls, M. Forsberg, M. T. Gronostaj, and D. Kokkinakis. 2010. The past meets the present in the Swedish FrameNet++. *EURALEX*.

L. Borin, M. Forsberg, and J. Roxendal. 2012. Korp – the corpus infrastructure of språkbanken. *LREC*.

S. Brody and M. Lapata. 2009. Bayesian word sense induction. *EACL*.

N. Chambers and D. Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. *ACL*.

J. Christensen, Mausam, S. Soderland, and O. Etzioni. 2011. An analysis of open IE based on SRL. *KCAP*.

R. Collobert and J. Weston. 2007. Fast semantic extraction using a novel neural network architecture. *ACL*.

D. Das, D. Chen, A. Martins, N. Schneider, and N. Smith. 2013. Frame-semantic parsing. *Computational Linguistics*, to appear.

S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6).

C. Fillmore, P. Kay, and M. C. O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: the case of let alone. *Language*, 64.

C. Fillmore, R. L. Goldman, and R. Rhodes. 2011. The FrameNet constructicon. H. Boas and I. Sag, editors, *Sign-based construction grammar*. CSLI.

C. Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language*, 280.

P. Gärdenfors. 2000. *Conceptual spaces: The Geometry of Thoughs*. Bradford Books.

D. Gildea and D. Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3).

A. Goldberg. 1995. *Constructions. A construction grammar approach to argument structure*. U. of Chicago Press.

S. Greene and P. Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. *NAACL*.

T. Günther and R. Johansson. 2013. Generalizing beyond the lexical level in sentiment classification of Twitter messages. *ACL*. Submitted.

J. Hajič, M. Ciaramita, R. Johansson, D. Kawahara, M. A. Martí, L. Màrquez, A. Meyers, J. Nivre, S. Padó, J. Štěpánek, P. Straňák, M. Surdeanu, N. Xue, and Y. Zhang. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. *CoNLL*.

Z. Harris. 1954. Distributional structure. *Word*, 10(23).

R. Johansson and A. Moschitti. 2013. Relational features in fine-grained opinion analysis. *Comp. Ling.*, 39(3).

R. Johansson and P. Nugues. 2006. A FrameNet-based semantic role labeler for Swedish. *COLING/ACL*.

R. Johansson and P. Nugues. 2007a. Extended constituent-to-dependency conversion for English. *NODALIDA*.

R. Johansson and P. Nugues. 2007b. Semantic structure extraction using nonproj. dependency trees. *SemEval*.

R. Johansson and P. Nugues. 2007c. Using WordNet to extend FrameNet coverage. *Building Frame-semantic Resources for Scandinavian and Baltic Languages, at NODALIDA*.

R. Johansson and P. Nugues. 2008a. Dependency-based semantic role labeling of PropBank. *EMNLP*.

R. Johansson and P. Nugues. 2008b. The effect of syntactic representation on semantic role labeling. *COLING*.

R. Johansson, A. Berglund, M. Danielsson, and P. Nugues. 2005. Automatic text-to-scene conversion in the traffic accident domain. *IJCAI*.

R. Johansson, K. Friberg Heppin, and D. Kokkinakis. 2012. SRL with the Swedish FrameNet. *LREC*.

R. Johansson. 2009. Statistical bistratal dependency parsing. *EMNLP*.

R. Johansson. 2012. Non-atomic classification to improve a SRL for a low-resource language. **SEM*.

R. Johansson. 2013. Training parsers on incompatible treebanks. *NAACL/HLT*. To appear.

J. Karlgren, G. Eriksson, M. Sahlgren, and O. Täckström. 2010. Between bags and trees – constructional patterns in text used for attitude identification. *ECIR*.

O. Kohonen, S. Virpioja, and K. Lagus. 2009. Constructionist approaches to grammar inference. *Proceedings of the NIPS Workshop on Grammar Induction, Representation of Language and Language Learning*.

T. Koo, X. Carreras, and M. Collins. 2008. Simple semi-supervised dependency parsing. *ACL*.

J. Lang and M. Lapata. 2010. Unsupervised induction of semantic roles. *NAACL*.

B. Lyngfelt, L. Borin, M. Forsberg, J. Prentice, R. Rydstedt, E. Sköldberg, and S. Tingsell. 2012. Adding a constructicon to the Swedish resource network of Språkbanken. *KONVENS*.

L. Màrquez, X. Carreras, K. Litkowski, and S. Stevenson. 2008. Semantic role labeling: An introduction to the special issue. *Computational Linguistics*, 34(2).

J. Materna. 2012. LDA-frames: An unsupervised approach to generating semantic frames. *CICLING*.

J. Mitchell and M. Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8).

A. Modi, I. Titov, and A. Klementiev. 2012. Unsupervised induction of frame-semantic representations. *NAACL-HLT 2012 Workshop on Inducing Linguistic Structure (WILS)*.

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1).

D. Pighin and A. Moschitti. 2009. Reverse engineering of tree kernel feature spaces. *EMNLP*.

J. Prentice and E. Sköldberg. 2011. Figurative word combinations in texts written by adolescents in multilingual school environments. *Young urban Swedish. Variation and change in multilingual settings*.

M. Sahlgren and O. Knutsson, editors. 2010. *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics*.

M. Sahlgren. 2006. *The Word-Space Model*. Ph.D. thesis, Stockholm University.

R. Socher, B. Huval, C. Manning, and A. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. *EMNLP*.

J. Turian, L. Ratinov, and Y. Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. *ACL*.

VETENSKAPSRÅDET
THE SWEDISH RESEARCH COUNCIL

Kod
2013-42850-102405-55

Name of applicant
Johansson, Richard

Date of birth
750709-2034

Title of research programme
Distributional Methods to Represent the Meaning of Frames and Constructions

Appendix B

Curriculum vitae

Vetenskapsrådet, Box 1035, SE-101 38 Stockholm, tel. +46 (0)8 546 44 000, vetenskapsradet@vr.se

Curriculum Vitae

Richard Johansson

1. University Degrees

Licentiate in Computer Science, Lund University, 2006. Supervised by Pierre Nugues.

M.Sc. in Computer Science, Lund University, 2003. Supervised by Pierre Nugues.

2. Doctoral Degree

Ph.D. in Computer Science, Lund University and the Swedish National Graduate School of Language Technology (GSLT), defended on December 5, 2008. Supervised by Pierre Nugues.

3. Postdoctoral Positions

Department of Information Engineering and Computer Science, University of Trento, Italy. 2009 – 2011. Supervised by Alessandro Moschitti.

5. Current Position

Postdoctoral researcher, Language Bank and Centre for Language Technology, University of Gothenburg, Sweden. 2011 – 2013. Research time: about 95%.

6. Previous Positions

PhD student at the Department of Computer Science, Lund University, Sweden. 2003–2008.

Postdoctoral researcher at the Department of Information Engineering and Computer Science, University of Trento, Italy. 2009 – 2011.

9. Other Information

Collaboration with Companies

XTRANORMAL, Montreal. Collaboration and consultation on text-to-scene conversion in 2006–2008.

MICROSOFT, Copenhagen. Consultation on semantic role labeling and information extraction for legal documents, 2007–2008.

IBM, Watson. Parser adaptation for the DeepQA project on the application of question answering techniques to Jeopardy answers in 2009–2010.

LOCALOT RESEARCH, Cambridge, MA. Consultation on parsing and semantic role labeling in 2010–2011.

Organization of Conferences and Workshops

NODALIDA workshop on Building Frame Semantics Resources for Scandinavian and Baltic Languages, 2007.

CoNLL Shared Task on Joint Learning of Syntactic and Semantic Dependencies, 2008.

CoNLL Shared Task on Syntactic and Semantic Dependencies in Multiple Languages, 2009.

Reviewing

Reviewer or PC member for the following international conferences: ACL (2009 – 2013), Coling (2010), CoNLL (2009 – 2012), EACL (2009, 2012), ECIR (2011 – 2013), ECML (2010), EMNLP (2009 – 2012), IJCAI (2007, 2011), IJCNLP (2011), LAW (2009 – 2011), LREC (2010, 2012), NAACL (2009, 2012), NODALIDA (2007), *SEM (2013), TextGraphs (2010, 2011).

Reviewer for the following journals: ACM Transactions on Speech and Language Processing, Computer Speech and Language, ACM Transactions on Intelligent Systems and Technology, Natural Language Engineering, Journal of Artificial Intelligence Research, Language Resources and Evaluation, Information Systems.

Awards

Best participating system in the SemEval-2007 task on Frame-semantic Structure Extraction (among 4 participants).

Best participating system in the CoNLL-2008 Shared task on Joint Learning of Syntactic and Semantic Dependencties (among 22 participants).

Invited talks

Department of Computational Linguistics, University of Saarland, October 2009: *Dependency-based Semantic Role Labeling.*

Google research, Zurich, June 2011: *Linguistically Driven Methods for the Extraction of Fine-grained Opinion Structure.*

Symposium at ESSIR 2011 – Bias and Diversity in IR: *Overview of Opinion Extraction Methods*.

Uppsala university, February 2012: *Linguistically Driven Methods for the Extraction of Fine-grained Opinion Structure.*

VETENSKAPSRÅDET
THE SWEDISH RESEARCH COUNCIL

Kod
2013-42850-102405-55

Name of applicant
Johansson, Richard

Date of birth
750709-2034

Title of research programme
Distributional Methods to Represent the Meaning of Frames and Constructions

Appendix C

List of publications

Vetenskapsrådet, Box 1035, SE-101 38 Stockholm, tel. +46 (0)8 546 44 000, vetenskapsradet@vr.se

Publications 2005–2013

More details can be found on my publications page:

`http://www.df.lth.se/~richardj/publications.cgi`

Citation counts can be found on my Google scholar page:

`http://scholar.google.se/citations?user=FvhWYU8AAAAJ`

My Google scholar *h*-index as of March, 2013 is 15.

The five most relevant publications for the project are marked with an asterisk.

1. Peer-reviewed Articles

(*) Johansson, Richard and Alessandro Moschitti. 2013. Relational features in fine-grained opinion analysis. To appear in *Computational Linguistics*, 39(3).

2. Peer-reviewed Conference Contributions

Johansson, Richard. 2013. Training parsers on incompatible treebanks. To appear in *2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, United States.

Ju, Qi, Alessandro Moschitti, and Richard Johansson. 2013. Structural reranking for hierarchical text classification. In *Advances in Information Retrieval; 35th European Conference on IR Research, ECIR 2013*, Lecture Notes in Computer Science 7814, pages 183–194. Moscow, Russia.

Johansson, Richard. 2012. Bridging the gap between two different Swedish treebanks. In *Proceedings of the Fourth Swedish Language Technology Conference (SLTC 2012)*, pages 40–41. Lund, Sweden.

(*) Johansson, Richard. 2012. Non-atomic classification to improve a semantic role labeler for a low-resource language. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 95–99. Montréal, Canada.

Moschitti, Alessandro, Qi Ju, and Richard Johansson. 2012. Modeling topic dependencies in hierarchical text categorization. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, pages 759–767. Jeju, Republic of Korea.

Ghosh, Sucheta, Giuseppe Riccardi, and Richard Johansson. 2012. Global features for shallow discourse parsing. In *Proceedings of the 13th annual SIGdial Meeting on Discourse and Dialogue*, pages 150–159. Seoul, Republic of Korea.

Borin, Lars, Markus Forsberg, Richard Johansson, Kristiina Muhonen, Tanja Purtonen, and Kaarlo Voionmaa. 2012. Transferring frames: Utilization of linked lexical resources. In *Proceedings of the Workshop on Induction of Linguistic Structure (WILS 2012)*, pages 8–15. Montréal, Canada.

Borin, Lars, Markus Forsberg, Karin Friberg Heppin, Richard Johansson, and Annika Kjellandsson. 2012. Search result diversification methods to assist lexicographers. To appear in *LAW*

2012 : The 6th Linguistic Annotation Workshop. Jeju, Republic of Korea.

(*) Johansson, Richard, Karin Friberg Heppin, and Dimitrios Kokkinakis. 2012. Semantic role labeling with the Swedish FrameNet. In *Proceedings of the Eighth Conference on International Language Resources and Evaluation (LREC-2012)*, pages 3697–3700. Istanbul, Turkey.

Ghosh, Sucheta, Richard Johansson, Giuseppe Riccardi, and Sara Tonelli. 2012. Improving the recall of a discourse parser by constraint-based postprocessing. In *Proceedings of the Eighth Conference on International Language Resources and Evaluation (LREC-2012)*, pages 2791–2794. Istanbul, Turkey.

Bennaceur, Amel, Falk Howar, Valérie Issarny, Richard Johansson, Alessandro Moschitti, Romina Spalazzese, Bernhard Steffen, and Daniel Sykes. 2012. Machine learning for emergent middleware. To appear in *Proceedings of the Joint Workshop on Intelligent Methods for Software System Engineering (JIMSE)*. Montpellier, France.

Bennaceur, Amel, Richard Johansson, Alessandro Moschitti, Romina Spalazzese, Daniel Sykes, and Valérie Issarny. 2012. Automatic service categorisation through machine learning in emergent middleware. To appear in *Proceedings of Software Technologies Concertation on Formal Methods for Components and Objects (FMCO)*, LNCS. Springer, Heidelberg.

Ghosh, Sucheta, Richard Johansson, Giuseppe Riccardi, and Sara Tonelli. 2011. Shallow discourse parsing with conditional random fields. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 1071–1079. Chiang Mai, Thailand.

Ju, Qi, Richard Johansson, and Alessandro Moschitti. 2011. Towards using reranking in hierarchical classification. In *Proceedings of the Joint ECML/PKDD – PASCAL Workshop on Large-Scale Hierarchical Classification*. Athens, Greece.

Ghosh, Sucheta, Sara Tonelli, Giuseppe Riccardi, and Richard Johansson. 2011. End-to-end discourse parser evaluation. In *Proceedings of the Fifth IEEE International Conference on Semantic Computing (ICSC 2011)*. Palo Alto, United States.

Johansson, Richard and Alessandro Moschitti. 2011. Extracting opinion expressions and their polarities – exploration of pipelines and joint models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 101–106. Portland, United States.

Bennaceur, Amel, Richard Johansson, Alessandro Moschitti, Romina Spalazzese, Daniel Sykes, Rachid Saadi, and Valérie Issarny. 2012. Inferring affordances using learning techniques. In *EternalS 2011*, CCIS 255. Springer, Heidelberg.

Bennaceur, Amel, Valérie Issarny, Richard Johansson, Alessandro Moschitti, Daniel Sykes, and Romina Spalazzese. 2012. Machine Learning for Automatic Classification of Web Service Interface Descriptions. In *ISoLA 2011 Workshops*, CCIS 336, pages 220–231. Springer, Heidelberg.

Dupplaw, David, Michael Matthews, Richard Johansson, and Paul Lewis. 2012. LivingKnowledge: a platform and testbed for fact and opinion extraction from multimodal data. In *EternalS 2011*, CCIS 255. Springer, Heidelberg.

Johansson, Richard and Alessandro Moschitti. 2010. Reranking models in fine-grained opinion analysis. In *Proceedings of the 23rd International Conference of Computational Linguistics (Coling 2010)*, pages 519–527. Beijing, China.

Johansson, Richard and Alessandro Moschitti. 2010. Syntactic and semantic structure for opinion expression detection. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL)*, pages 67–76. Uppsala, Sweden.

Johansson, Richard and Alessandro Moschitti. 2010. A flexible representation of heterogeneous annotation data. In *Proceedings of the Seventh Conference on Language Resources and Evaluation (LREC'10)*, pages 3712–3715. Valetta, Malta.

Johansson, Richard. 2010. Self-adaptation and evolution by learning – description of EternalS task force 3. In *ISoLA 2010, Part II, LNCS 6416*, pages 30–31. Heraklion, Greece.

(*) Johansson, Richard. 2009. Statistical bistratal dependency parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 561–569. Singapore.

Hajič, Jan, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 Shared Task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 1–18. Boulder, United States.

Johansson, Richard and Alessandro Moschitti. 2009. LivingKnowledge: Exploring the spectrum of opinions over time. In *Proceedings of the Workshop on Advanced Technologies for Digital Libraries (AT4DL 2009)*. Trento, Italy.

Persson, Jacob, Richard Johansson, and Pierre Nugues. 2009. Text categorization using predicate–argument structures. In *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009)*, pages 142–149. Odense, Denmark.

Johansson, Richard and Pierre Nugues. 2008. Dependency-based semantic role labeling of PropBank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 69–78. Honolulu, United States.

Johansson, Richard and Pierre Nugues. 2008. Dependency-based syntactic–semantic analysis with PropBank and NomBank. In *CoNLL 2008: Proceedings of the Twelfth Conference on Natural Language Learning*, pages 183–187. Manchester, United Kingdom.

(*) Johansson, Richard and Pierre Nugues. 2008. The effect of syntactic representation on semantic role labeling. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 393–400. Manchester, United Kingdom.

Johansson, Richard and Pierre Nugues. 2008. Comparing dependency and constituent syntax for frame-semantic analysis. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 480–484. Marrakech, Morocco.

Surdeanu, Mihai, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *CoNLL 2008: Proceedings of the Twelfth Conference on Natural Language Learning*, pages 159–177. Manchester, United Kingdom.

Johansson, Richard. 2007. Logistic online online learning methods and their application to incremental dependency parsing. In *Proceedings of the ACL 2007 Student Research Workshop*, pages 49–54, Prague, Czech Republic.

Johansson, Richard and Pierre Nugues. 2007. Extended constituent-to-dependency conversion

for English. In *NODALIDA 2007 Conference Proceedings*, pages 105–112. Tartu, Estonia.

Johansson, Richard and Pierre Nugues. 2007. Incremental dependency parsing using online learning. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 1134–1138. Prague, Czech Republic.

Johansson, Richard and Pierre Nugues. 2007. Semantic structure extraction using nonprojective dependency trees. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 227–230. Prague, Czech Republic.

Johansson, Richard and Pierre Nugues. 2007. Syntactic representations considered for frame-semantic analysis. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories*. Bergen, Norway.

Johansson, Richard and Pierre Nugues. 2007. Using WordNet to extend FrameNet coverage. In *Proceedings of the Workshop on Building Frame-semantic Resources for Scandinavian and Baltic Languages, at NODALIDA*, pages 27–30. Tartu, Estonia.

Johansson, Richard and Pierre Nugues. 2006. Automatic annotation for all semantic layers in FrameNet. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 135–138. Trento, Italy.

Johansson, Richard and Pierre Nugues. 2006. Construction of a FrameNet labeler for Swedish text. In *Proceedings of LREC-2006*. Genoa, Italy.

Johansson, Richard and Pierre Nugues. 2006. A FrameNet-based semantic role labeler for Swedish. In *Proceedings of COLING/ACL 2006*, pages 436–443. Sydney, Australia.

Johansson, Richard and Pierre Nugues. 2006. Investigating multilingual dependency parsing. In *Proceedings of the Shared Task Session of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 206–210. New York City, United States.

Berglund, Anders, Richard Johansson, and Pierre Nugues. 2006. Extraction of temporal information from texts in Swedish. In *Proceedings of LREC-2006*. Genoa, Italy.

Berglund, Anders, Richard Johansson, and Pierre Nugues. 2006. A machine learning approach to extract temporal information from texts in Swedish and generate animated 3D scenes. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 385–392. Trento, Italy.

Johansson, Richard, Anders Berglund, Magnus Danielsson, and Pierre Nugues. 2005. Automatic text-to-scene conversion in the traffic accident domain. In *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*, pages 1073–1078. Edinburgh, United Kingdom.

Johansson, Richard and Pierre Nugues. 2005. Automatic conversion of traffic accident reports into 3D animations. In *SIGRAD-05*. Lund, Sweden.

Johansson, Richard and Pierre Nugues. 2005. Sparse Bayesian classification of predicate arguments. In *CoNLL-2005: Proceedings of the Ninth Conference on Computational Natural Language Learning, 43rd Annual Meeting of the Association of Computational Linguistics*, pages 177–200. Ann Arbor, United States.

Johansson, Richard and Pierre Nugues. 2005. Using parallel corpora for automatic transfer of FrameNet annotation. In *Proceedings of the 1st ROMANCE FrameNet Workshop*. Cluj-Napoca, Romania.

3. Books, Dissertations, Conference Proceedings

Johansson, Richard. 2008. Dependency-based semantic analysis of natural-language text. Doctoral dissertation, Lund University.

Johansson, Richard. 2006. Natural language processing methods for automatic illustration of text. Licentiate thesis, Lund University.

Nugues, Pierre and Richard Johansson. 2007. Building frame semantics resources for Scandinavian and Baltic languages. Conference proceedings. Tartu, Estonia.

5. Publicly Available Software

Richard Johansson was the single designer and implementer of the following pieces of software.

LTH CONSTITUENT-TO-DEPENDENCY CONVERSION TOOL: a program that converts Penn Treebank trees into labeled dependency trees. Used to create the English data in the CoNLL-2007, 2008, and 2009 Shared Tasks. Also used to prepare dependency annotations in the Open American National Corpus.
URL: `http://nlp.cs.lth.se/software/treebank_converter`

LTH SYSTEM FOR FRAME-SEMANTIC STRUCTURE EXTRACTION: a tool for frame-semantic analysis of English text that participated in the SemEval-2007 task on Frame-semantic Structure Extraction.
URL: `http://nlp.cs.lth.se/software/semantic_parsing:_framenet_frames`

LTH SYNTACTIC–SEMANTIC DEPENDENCY PARSER: a tool for dependency-syntactic parsing and semantic role labeling of English text that participated in the CoNLL-2008 shared task.
URL: `http://nlp.cs.lth.se/software/semantic_parsing:_propbank_nombank_frames`

UNITN FINE-GRAINED OPINION ANALYSIS: A system for extraction of opinion expressions with polarity, intensity, and opinion holder features.

RVM-TRAIN: a C/BLAS implementation of the Relevance Vector Machine training algorithm for binary kernel classifiers. URL: `http://www.df.lth.se/~richardj/rvm.cgi`

J-SVM^{struct}: a Java binding to the SVM^{struct} trainer for structure SVMs.
URL: `http://www.df.lth.se/~richardj/jsvmstruct.cgi`

VETENSKAPSRÅDET
THE SWEDISH RESEARCH COUNCIL

Kod
2013-42850-102405-55

Name of applicant
Johansson, Richard

Date of birth
750709-2034

Title of research programme

Appendix N

Budget and research resources

Vetenskapsrådet, Box 1035, SE-101 38 Stockholm, tel. +46 (0)8 546 44 000, vetenskapsradet@vr.se

Motivering av budget

Richard Johansson
Språkbanken, Institutionen för svenska språket, Göteborgs universitet

1 Projektbudget

Typ av bidrag	Status	Finansiär	Projektledare	Bidragsperiod	Totalbelopp
Projektbidrag	Ansökt	VR	Richard Johansson	2014-2018	4.659 tkr

2 Lönekostnader

Richard Johansson (RJ) kommer att leda projektet, utföra och analysera undersökningar, och publicera projektets resultat i vetenskapliga konferenser och tidskrifter. RJ kommer också att utveckla delar av de datorprogram som kommer att användas i projektet.

Projektet föreslås bekosta 40% av RJs heltidstjänst i perioden januari 2014 – december 2018. De övriga 60% kommer att täckas av institutionstjänstgöring och andra projekt.

En doktorand kommer att anställas på Språkbanken vid institutionen för svenska språket på Göteborgs universitet. Doktoranden kommer att utföra huvuddelen av arbetet med att implementera de idéer som definierats av projektledaren, och genomföra projektets vetenskapliga experiment. Doktorandens finansiering täcks av GU för den tid doktoranden läser kurser och undervisar, så projektet kommer att bekosta 20% under de två första åren och 80% under de tre återstående.

3 Inköp av böcker

Relevant litteratur kommer att inhandlas i de fall den inte finns att tillgå lokalt.

4 Resor

Projektets resultat kommer att publiceras vid internationella konferenser. Vi räknar med ungefär 3 konferensresor per år. Vi planerar dessutom att doktoranden kommer att ha två utlandsvistelser för forskningssamarbeten.

VETENSKAPSRÅDET
THE SWEDISH RESEARCH COUNCIL

Kod
2013-42850-102405-55

Dnr
2013 -

Name of applicant
Johansson, Richard

Date of birth
750709-2034

Reg date
2013-04-10 14:16:10

Project title
Distributional Methods to Represent the Meaning of Frames and Constructions

NT

Research grants NT April 11, 2013
Project Grant Junior Researchers
*Nat-Tek generellt
Göteborgs universitet

NT-2

Appendix S

Signatures

2014	2015	2016	2017	2018	
583	616	1059	1147	1255	2014-01-01 -- 2018-12-31

Applicant | **Date**

Head of department at host University | **Clarifi cation of signature** | **Telephone**

A signature on the application is required not only from the applicant but also from the authorised representative of the administrating organisation/institution or equivalent (normally the head of the department or establishment where the research is to be conducted). The signature confirms that the department can accommodate the proposed research, position or equipment; that the costing in the application is approved for the department's part, that any proposed experimentation on human or animal subjects has been reported, and that the applicant has reported any secondary occupations and commercial ties (s)he may have, and nothing inconsistent with good research practice has thereby emerged. The applicant must have discussed these conditions with the representative of the administrating organisation/institution or equivalent before the latter approves and signs the application.

NOTE the exceptions!
Please note that a different procedure applies for the following types of grants:

- Infrastructure- Operation Grants
- Infrastructure- Grants for Expensive Equipment
- Infrastructure- Grants for Large Databases
- Framework Grant SIMSAM

For these types of grants the vice-chancellor, instead of the head of department, must sign Appendix S. Thereby (s)he confirms the commitments of the administrative organization. The details are specified in the Call for Proposals, published at www.vr.se.

Vetenskapsrådets noteringar
Kod

2013-42850-102405-55

Vetenskapsrådet, Box 1035, SE-101 38 Stockholm, tel. +46 (0)8 546 44 000, vetenskapsradet@vr.se