Vetenskapsrådet

Ansökan

Tahmasebi, Nina

Information om sökande

Namn: Nina Tahmasebi
Födelsedatum: 19821029
Kön: Kvinna
Medelsförvaltare: Göteborgs universitet
Hemvist: Svenska språket, inst för

Dr-examen: 2013-11-13
Akademisk titel: Doktor
Arbetsgivare: Göteborgs universitet

Information om ansökan

Utlysningsnamn: Forskningsbidrag Stora utlysningen 2018 (Humaniora och samhällsvetenskap)
Bidragsform: Projektbidrag
Sökt inriktning: Fri
Ämnesområde utlysning: HS

Projekttitel (svenska): Mot automatiska metoder för att upptäcka språkförändring
Projektstart: 2019-01-01
Projektslut: 2022-12-31
Sökt beredningsgrupp: HS-J
Klassificeringskod: 60201. Jämförande språkvetenskap och allmän lingvistik, 10208. Språkteknologi (språkvetenskaplig databehandling)
Nyckelord: Lexical replacement, Semantic change, Sentiment change, Automatic detection of change

Sökta medel

År:	2019	2020	2021	2022
Belopp:	1 496 592	1 495 886	1 498 454	1 498 746

Medverkande

Namn: Richard Johansson
Födelsedatum: 19750709
Kön: Man
Land: Sverige

Dr-examen: 2008-12-05
Akademisk titel: Docent
Arbetsgivare: Göteborgs universitet

Namn: Maria Koptjevskaja Tamm
Födelsedatum: 19570612
Kön: Kvinna
Land: Sverige

Dr-examen: 1988-04-16
Akademisk titel: Professor
Arbetsgivare: Stockholms universitet

Namn: Susanne vejdemo
Födelsedatum: 19821113
Kön: Kvinna
Land: Sverige

Dr-examen: 2017-03-03
Akademisk titel: Doktor
Arbetsgivare: City University of New York: College of Staten Island

Beskrivande information

Projekttitel (svenska)*

Mot automatiska metoder för att upptäcka språkförändring

Projekttitel (engelska)*

Towards Automatic Detection of Language Change

Abstract (engelska)*

Today, we lack computational tools for studying lexical and semantic changes at a large scale. Current methods are limited and require huge amounts of text. Studies on semantic change capture only main changes of a single word and offer no possibility to capture the interplay of change in a semantic field. In this project, we aim to find automatic, corpus-based methods for detecting semantic change and lexical replacement for Swedish.

We will investigate the fundamental questions of how, when, and why languages change to allow us to quantify language change and shift lexical typological research from small case studies done on limited data sets to larger scales and over wider time spans using various media types and sources.

The results of the project will advance research in NLP and semantics and have practical benefits for researchers in other fields; We will be able to facilitate empirical study of language changes themselves, highlight changes for the public to avoid wrongful interpretations and account for language change in large-scale text mining applications such as information extraction and information retrieval. The results will also benefit the public as they access and interpret historical text as well as researchers that wish to track concepts over time without manually finding and accounting for language change.

Our group consists of a PI (50% + 30%), a researcher (40%), an ass. professor (15%), a professor (10%), and a software engineer (15% + 25%).

Populärvetenskaplig beskrivning (svenska)*

I takt med att vår värld och vår livsstil förändras, förändras även vårt språk. Vi lär oss nya ord, skaffar nya betydelser på existerande ord eller förändrar betydelser så att de passar in för att beskriva vår värld och vår tid. Vi glömmer fort och när vi tittar tillbaka, tex i gamla tidningsmaterial så är det inte alltid lätt att förstå vad som menats. Vem kommer tex ihåg *yuppienallen* eller vad ordet *guzz* betyder?

Generellt sett kan vi dela upp språkliga förändringar i två kategorier, den första rör ord vars betydelser ändras över tid medan den andra rör ord som ersätter varandra för samma betydelse. Ordet *rock* är ett exempel på ett ord som fått en tillagd betydelse, utöver att vara ett ytterplagg är det även en musikstil och faller därför i den första kategorin. I den andra kategorin faller en betydelse som 'slug'. Tidigare har ordet *fin* använts för denna betydelse, men ersatts av just *slug* eller *listig*. I denna senare kategori faller även namnförändringar, tex personer, städer och länder som byter namn.

När det gäller informationssökning i gammalt material, tex tidningar eller böcker, så orsakar ordbyten problem för att finna relevant material. Detta gäller oavsett om den som söker är en person,
eller ett datorprogram. Anta att vi vill hitta material om första världskriget från perioden då kriget pågick: vid den tiden kallades kriget inte för *första världskriget* och en sökning med denna sträng skulle inte ge oss alla relevanta dokument.

När vi väl har hittat relevant material måste vi kunna tolka innehållet korrekt och där ställer betydelseändringar till det för oss. *Han var en grym person*. Hur detta skall tolkas beror naturligtvis på när meningen skrevs. Att automatiskt finna dessa ord vars betydelser har ändrats över tid, samt att veta hur förändringarna skett är av högsta vikt för att hjälpa människor och datorprogram som behöver tolka äldre (och inte alltid så gamla) texter.

Vi kommer att bygga verktyg att studera vårt språk och dess förändringar i större skala. När får ett ord en ny betydelse och hur länge är betydelsen aktiv? Hur leder en förändring i ett ord till vidare förändringar i andra, besläktade ord? Problemet är av högsta vikt: allt mer historiskt material blir öppet och tillgängligt. Det är även intressant i sociala medier där språket ändras fort. Det lockar forskare från alla domäner, framförallt digial humaniora, att forska i historiskt material och leta svar på ett automatiskt och storskaligt vis. Dessa forskare ska inte behöva vara experter på historisk lingvistik för att kunna få tillgång till denna information. Allt ifrån attityden till retorik genom historien, till abstraktion av marknaden och olika politiska partiers användning av ord kan studeras och gynnas av att hantera språkliga förändringar automatiskt.

Problemen med att finna dessa förändringar är många och stora. Ordböcker och andra resurser kan användas till viss grad, men finns sällan i digitalt format, täcker inte alla epoker eller domäner och är tänkta som referenser. För att modellera den faktiska användningen av språket bör vi istället använda oss av automatiska metoder och börja med att finna betydelsen av ord ur en text. Detta mycket svåra problem kallas för *betydelseinduktion*. Vi kommer att studera ordens betydelser genom dess grannar enligt devisen "*You shall know a word by the company it keeps*" (Firth, J. R. 1957:11).

När vi väl har funnit vad orden betyder i varje tidsperiod så jämför vi betydelser över tid för att finna förändringar. Tidigare försök som gjorts har fokuserat på engelska och oftast delar av eller olika aspekter av problemet och ännu saknas t.ex. både automatiska utvärderingsmetoder och data att utvärdera på. Mycket fokus har legat på att hitta olika typer av förändringar utan att mäta eller filtrera brus. Nya, effektiva metoder använder distributionell semantik för att projicerar orden till vektorer och analyserar förändringar i dessa och kan då svara på *att* men inte *vad* som ändrats. Vi kommer att använda oss av en kombination av distributionell semantik och betydelseinduktion för kunna svara både på *vad* som ändrats och *när*. Vi kommer att använda de mycket stora samlingar av svensk text på Språkbanken, i ett världsunikt samarbete mellan semantiker och språkteknologer som med sina respektive expertiser har mycket goda förutsättningar att finna nya, automatiska metoder samt att i större skala svara på existerande hypoteser och deras generalisering till andra datamängder och tidsepoker.

Forskningsbeskrivning

Redogörelse för etiska överväganden*

There are no specific ethical considerations with this proposal, we will base our analysis on openly available corpora and text.

I projektet ingår hantering av persondata

Nej

I projektet ingår djurförsök

Nej

I projektet ingår humanförsök

Nej

Forskningsplan*

Se nästa sida för bilaga.

1 Purpose and aims

Today, we lack computational tools for studying lexical and semantic changes at a large scale. Current methods are limited in what they can find and require huge amounts of text that is typically not available for (historical) Swedish. Studies on automatic detection of semantic change detect only main changes of a single word and offer no possibility to capture the interplay of change in a semantic field. In this project, we aim to find automatic, corpus-based methods for detecting semantic change and lexical replacement, pinpoint time of change and handle smaller amounts of text. We will replicate existing manual studies at larger scale over a wide time span, and various media types and sources. We will help overcome hurdles based on diachronic and synchronic language change for the digital humanities and social sciences (DHSS). The focus language is Swedish, which has not been targeted for this kind of research, and for which we cannot satisfactorily apply methods developed for English. Tailor-made tools will strengthen the viability of Swedish in text analysis and data mining.

The potential of the project is greatly enhanced by the unique collaboration that forms its core: semanticists focused on high quality detail-oriented research, which generate wide-reaching hypotheses about how word change and semantic change intertwine; and language technology and data science researchers focused on using and advancing cutting-edge technology to work with Swedish digital language data. The project adds value through its core research contributions in both Natural Language Processing (NLP) and semantics; its application in DHSS for both researchers and nonprofessional users of digital archives; and for new tools for Swedish large-scale text analysis.

- **Linguistic research**: There are many open questions in the burgeoning field of quantitative semantics, which we cannot currently answer with existing computational methods; How does lexical change and semantic change interact? Why do different parts of the vocabulary change at different speeds? How does change spread throughout a word's semantic network? High quality case studies of change often produce hypotheses, and we will provide tools to test and quantify these hypotheses, tying lexical replacement to semantic change for a semantic field.
- **Applied DHSS Research**: Most researchers interested in historical content are not experts in historical language change. In particular, we aim to help researchers in DHSS who wish to investigate concepts, e.g. attitudes towards concepts, over time. Imagine investigating references to the *telegraph* and not knowing about sense and sentiment changes thus wrongly interpreting the content. Or investigating criminals over time not knowing about lexical replacements - e.g., that *varg* 'wolf' was used to describe violent criminals in the past, thus missing out on content.
- **Added value for nonprofessional users**: We will assist nonprofessional users of textual archives, to find and interpret content. For example, students are one of the largest user groups of Språkbanken ('the Swedish Language Bank'), and often study concepts in historical content without the ability to account for language change; good visual interfaces and easily understandable results will open up these resources even more.

The outcome of this project will advance NLP and semantics in several fields:

- **Word sense induction** (WSI), automatically finding the meaning of words from text, has not yet been studied in depth for Swedish. Here, we will establish the state-of-the-art, contribute to research on synchronic sense differences, and enable Swedish to stay a viable language in the digital realm. Unless we make significant research investments in core NLP, bilinguals will switch to the greater computational tools of English, and lose access to Swedish material. This is evident from the lack of e.g., sentiment mining resources and tools, word sense induction and disambiguation tools etc. for Swedish.
- **Semantic change** (SC): We will study word sense change detection, i.e., the process of automatically finding changes in a word's meanings over time. We will go beyond current limitations with 1-2 kinds of change, or a few studied time points, to be able to answer *what* happened, *how* the changes relate to what we already know and *when* the change took place. We will apply change detection to Swedish, to individual words and the interplay in a semantic field, setting the state-of-the-art for Swedish, and significantly furthering the research field internationally.
- **Sentiment change**: We will study changes in sentiment as a consequence of word sense change and have a unique opportunity to create a Swedish diachronic sentiment lexicon that can be used when studying concept change over time (e.g. computers, nuclear plants, women, immigration).

- **Lexical Replacement** (LR): We will have a unique opportunity to study Lexical Replacement, a class of change that is complex and extremely difficult to detect automatically. For Swedish, it has only been done by labor-intensive and detail-rich case studies e.g. on color word change (Vejdemo, 2017). Such studies generate hypotheses about general LR processes, which then necessitate computational approaches for falsification attempts. Here, classical diachronic semantics meet computational semantics and data science, one of the great strengths of this project.

Lexical Replacement and Semantic Change

The most common word for 'young female human' changed from *maiden* in Old and Middle English to *girl* in Modern English. This is a case of lexical replacement: a bundle of semantic material is first symbolized by one word, and later in time by another word (onomasiology). Parallel to this, *girl* 'young person' came to mean 'young female person'. This is a case of semantic change: a word stays the same over time [1] while the semantic material it symbolizes changes (semasiology). It is often useful to talk about the semantic material of a word as clustering into several sub-meanings: senses. Senses can be added, removed or changed. A particularly interesting kind of alteration is positive/negative sentiment change: while the morpheme *skit* in *skitdag* 'shitty, bad day' has a negative connotation, in the last few decades it has acquired a positive connotation as an intensifier in words like *skitgott* 'really good'. Sentiment analysis is increasingly important for commercial and political research and can greatly benefit from automatically handling lexical and semantic change.

All these intertwined processes make lexical and semantic change highly complex problems relying on defining a particular sense (and the allocation of senses to words), problems that are considered AI complete, i.e. equivalent of making computers as intelligent as people. Recent NLP advances based on the distributional hypothesis of meaning have proven extremely useful in assisting researchers in untangling SC processes. The distributional hypothesis links semantic similarity to distributional similarity - meaning can be induced from the set of words that appear in similar contexts. Automatically induced senses are approximations of an underlying word sense and vary naturally depending on which sentences that are used for the sense induction. A great challenge in automatic change detection is determining when two induced senses (for the same word at different times) are natural variations and when the differences represent sense change (cf *okasionelle* and *usuelle Bedeutung* in Paul (1886)).

The methods developed in this project will go beyond the state-of-the-art in the field in several aspects. Previously, **SC detection** projects have primarily focused on (i) a limited number of change types, e.g., only birth of senses; (ii) a few (far apart) time points, e.g., 50-year slots; or (iii) methods that find signals for change without differentiating between change types or separating the senses of a word (i.e., one topic/vector/cluster per word); and (iv) words in isolation, not their interplay within a semantic field. Existing techniques reduce complexity severely because considering yearly time buckets over two centuries and up to 5 senses per time period, the solution space is in the order of 5^{200} which is impossible to compute and evaluate. This project will build on the promising reduction techniques described by Tahmasebi and Risse (2017) to enable us to answer the *what*, *how* and *when* questions in full, and create a complete picture of all changes related to a word and its semantic field.

For **automatic LR**, we will set the state-of-the-art simply because there is almost no existing research. The problem is extremely complex because words must be linked based on their (stable) senses. In this project, we have a unique opportunity to study the LR problem because we are one of few research groups that will attempt to solve the problem of word sense change first. We will begin by working on word sense induction for Swedish, as this is the core of our methodology. Once we can induce word senses automatically, we can begin to detect change in senses (SC) and then, as a third step, find word replacement (LR). Using these tools, we can study the varying speed and different processes of LR and SC in e.g. different parts of the vocabulary and during different time periods.

Thus far, methods for detecting sentiment change, i.e., words changing their sentiment value, (Cook and Stevenson, 2010; Nguyen et al., 2012) have not differentiated between different senses: only the predominant value of a word has been considered. We will be able to overcome this hurdle by first solving word sense change. We build on ongoing effort at Språkbanken and our sentiment lexicon, SenSaldo, to tackle diachronic sentiment analysis for Swedish.

[1] In this project, we largely ignore problems caused by phonological (sound) change due to the shorter time spans of our data. For spelling variations, to detect when words that are spelled differently should be considered the same (diachronic and synchronic), we will rely on existing NLP tools and the PIs ongoing work at Språkbanken outside this project.

For all kinds of change targeted in this project, we will provide **textual evidence** to support our claims, e.g. , example sentences for each sense used to help users evaluate and understand the results. This is a prerequisite for uptake in the research community, in particular in the DHSS.

We envision several use cases that will help researchers and nonprofessional users to study language changes themselves, enabling them to search and explore archival content and improve downstream large-scale text mining applications. For researchers that are interested in language changes in general, our results will offer answers to what has changed as well as how and when it changed. It will also be possible to answer more complex questions like *how is change in one word connected to changes in others in the same semantic field?*

For researchers that have an interest in the resources but not necessarily in the changes themselves, e.g., researchers in DHSS, our methods will help to gather evidence for concepts by finding linked vocabulary and their senses, e.g. the word *handikappad* 'handicapped' has been replaced over time *(handikappad→ funktionshindrad→ f. nedsatt→ f. variation)*. The replacement aims to remove negatively connotated senses, but from the continuous replacements, we know that these still catch up. Upon completion of this project, it will be possible to study the lexical replacements on the one hand, and tie it with semantic change. E.g. Does a new word like *funktionshindrad* take over a subset of the senses of the previous word at first and then later add the negative senses? How fast are the negative senses added for each lexical replacement and does it speed up with additional replacements? Do we include more or less in each sense or add new senses over time?

The methods developed in the project will be generally applicable, but our primary target language is Swedish with equal focus on historical and modern text. E.g. Swedish historical newspapers, Kubhist 1750-1925, books (the Literature bank), parliamentary data (SUC, SOU) and modern newspapers but also social media text where there is evidence of high linguistic diversity and creativity (Goel et al., 2016). This makes the availability of large amounts of Swedish text crucial to the project. Språkbanken continuously collects texts written in Swedish, to date there is over 10 billion words of modern Swedish (e.g. fiction, news, politics and social media) and over one billion words of historical news materials. We will also use digital lexical resources, like SALDO, Svedbergs, Dahlin and SAOL. We will replicate our research using English text such as the Corpus of Historical American English and Google books. We will extend previous studies and quantify hypothesis regarding lexical and semantic change, work that can feed back into our tools for quality assessment and improvement.

2 State-of-the art

2.1 Vector-space models

In vector-space models of word meaning, each word is associated with a vector in a geometric space, also known as a *word embedding*, so that similarity of meaning can be defined in terms of geometric proximity (Sahlgren, 2006; Turney and Pantel, 2010). This is a statistical, knowledge-free method for computing a representation of word meaning, which means that the meaning can be *discovered* rather than described manually by experts. The vectors are derived either by an explicit statistical analysis of the word's co-occurrence patterns, or as a by-product of training a neural network (Baroni et al., 2014). While traditional vector spaces associate each word with one single meaning, which is not sufficient for our purposes, a number of attempts have been made to take ambiguity of word meaning into account (Schütze, 1998, *inter alia*). Recent examples include Nieto Piña and Johansson (2015), who derived a multi-sense modification of the well-known skip-gram method, Word2Vec, (Mikolov et al., 2013), and methods that rely on an expert-defined lexicon to build sense vectors (Johansson and Nieto Piña, 2015). Bartunov et al. (2016) presented a Bayesian variant of the multi-sense skip-gram model that automatically determines the necessary number of word senses.

Neural embeddings, like Word2Vec, require a large amount of instance of each word to produce sensible vectors. Therefore, new methods are developed with a high learning rate early on, that is later slowed, to find good neighborhood also for rare words, suitable for historical corpora (Herbelot and Baroni, 2017). Another limitation of neural embeddings is the inherent randomness; both the initialization as well as the order in which the training examples are seen affect the resulting vectors (Hellrich and Hahn, 2017). Bamler and Mandt (2017) point to overfitting when there is too little data. For Kubhist, there are only five 10-year periods (1850-1890) with over 100 million tokens, thus limiting the possibility of finding stable vectors, in particular if sense-differentiated embeddings are intended where textual evidence for each word must be further divided into senses (Tahmasebi, 2018).

2.2 Approaches for Sense Change Detection

A rich tradition of cross-linguistic, detailed studies in lexical semantics has generated hypotheses that are ripe for further quantitative exploration made possible by the outcomes of this project. For example, synchronically: there are cognitively different kinds of temperature sensations, which form overt or covert sub-senses (Koptjevskaja-Tamm, 2015), that may be handled by different grammatical constructions (Pustet, 2015). Diachronically: word senses often go from describing external reality to internal perception (Traugott and Dasher, 2002); change often proceeds from a higher sensory modality to a lower (Viberg, 1980); perception verbs develop into cognitive verbs (Sweetser, 1991); or the exact hues denoted by Swedish color terms may change in a predictable fashion over time (Vejdemo, 2017).

The first methods for automatic detection of word sense change were based on context vectors; they investigated semantic density (Sagi et al., 2009) and utilized mutual information (Gulordava and Baroni, 2011) to identify semantic change over time; both methods detect signals of change but neither aligns senses over time or determines what changed. Topic-based models (where topics are interpreted as senses) were used to detect novel senses by identifying new topics in a later corpus compared to an earlier one (Lau et al. (2012); Cook et al. (2014)), or by clustering topics over time (Wijaya and Yeniterzi, 2011). A dynamic topic model where topics for time t are learned with respect to topics from time t-1 is proposed by Frermann and Lapata (2016). With one exception, no alignment is made between topics to allow following the diachronic progression of a sense.

Graph-based models are utilized by Mitra et al. (2015, 2014) and Tahmasebi and Risse (2017) and aim to revealing complex relations between a word's senses by (a) modeling senses per se using WSI; and (b) aligning senses over time. The models allow differentiation of individual senses at different periods in time and the latter also group senses into linguistically related concepts (Cooper, 2005).

The largest body of work is done using word embeddings of different kinds in the last years (Basile et al., 2016; Kim et al., 2014; Zhang et al., 2016b). Embeddings are trained on different time-sliced corpora and compared over time. Kulkarni et al. (2015) project words onto their frequency, POS and semantic vectors and propose a model for detecting statistically significant changes between time periods. Hamilton et al. (2016) view both similarity between a priori known pairs of words, and between a word's vectors over time to detect change and have released HistWords, a set of pre-trained historical word embeddings. Basile et al. (2016); Hamilton et al. (2016); Kulkarni et al. (2015) all propose different methods for projecting vectors from different time periods onto the same space to allow comparison. Bamler and Mandt (2017) propose using dynamic embeddings to completely avoid projection, which results in smoother diachronic embeddings.

Existing methods for detecting change based on word embeddings do not allow us to recover individual senses; they model words with one vector per time unit and detect meaning change as change in the direction of those vectors. Once a change point is found, the most similar words around that time are used to illustrate the change. However, the most similar terms will only represent the dominant sense; they will not reflect the other senses or capture stable parts of a word. If multi-sense embeddings are used, to allow senses to be modeled individually, we again face multiple vectors for each time point, over hundreds of years, leading to an overwhelming result space. Solving this reduction and investigating the tradeoff between a large results space and information loss during reduction, is at the core of this field and largely ignored in previous work.

We will take a combined approach, utilizing the potential of embeddings with the expressiveness of sense-differentiated methods, using the reduction techniques proposed in our previous work.

2.3 Lexical Replacement (LR)

There are several impressive taxonomies that lists types of LR processes (Ullman, 1957), but these are mainly descriptive and make little attempt to predict what kind of lexical change might happen for a given concept. Some proposed general hypotheses are that nouns are replaced more readily than verbs, that more frequent words are insulated from replacement (Pagel et al., 2007) and that rich synonym networks speed up replacement (Vejdemo and Hörberg, 2016), hypotheses that we can investigate using quantitative approaches on large scale, diachronic text upon completion of this project.

Previous work on automatic detection of LR has been very limited and mainly focused on named entity changes. The interest has mostly been from an information retrieval (IR) perspective (Anand et al. (2012); Berberich et al. (2007, 2009); Morsy and Karypis (2016)). Kanhabua and Nørvåg (2010)

find semantically related named entities using Wikipedia links, limiting the method to modern, common domain entities and evaluate indirectly in an IR setting. Kaluarachchi et al. (2010) propose to find named entities via linked verbs that relate over time, referring the problem to diachronic linking of verbs. These attempts are computationally expensive, as they require recurrent computations because the target time is unknown beforehand. In our previous work, Tahmasebi et al. (2012b), we rely on bursts in frequency to detect time periods in which we search for name change thus eliminating recurrent computation. In all work, changes are only found pairwise, senses are not differentiated and there is no validity period associated with each name. By finding word replacements after having found word sense changes, we can generalize beyond names and overcome many of these obstacles.

3 Significance

This project will bring forth tools for computationally detecting lexical and semantic changes as well as word sense induction for Swedish. These tools give us a chance to study language changes in their own right but also overcome hurdles with research on historical text. We will bring together historical linguistics with NLP and data science, a unique collaboration needed to study SC and LR in full.

We believe the results of the project will advance the field in several research areas, offer benefits for researchers in the DHSS and lower the threshold for the public to make use of our textual resources. These areas include but are not limited to:

- Reliable methods for detecting semantic and lexical change in both synchronic and diachronic data contexts; how do language changes spread, in which media do they appear firstly and what is the change rate? We will further the studies in e.g., Vejdemo and Hörberg (2016).
- We offer researchers in the DHSS a method to gather evidence and analyze text related to their concept of interest without being experts in diachronic or synchronic language change, thus reducing the threshold to target large-scale text mining and the risk of drawing wrongful conclusions based on word sense and sentiment changes.
- We open up the vast resources of our digital archives, like the cultural treasures of Språkbanken, to the public and nonprofessional users. These users wish to use the resources without investing considerable time to (1) find all words related to their search query and then, once they have found the texts they are interested in, (2) look up words to find if their meanings have changed.
- Many downstream NLP applications, such as semantic role labeling and word sense disambiguation would benefit from robust methods to detect lexical and semantic change. The resulting tools will feed back into the Korp processing pipeline making the results directly usable. We will openly release modern and historical (sense-differentiated) word vectors, similar to HistWords (Hamilton et al., 2016) for direct inspection and use in other NLP applications.

Språkbanken follows an explicit policy of openness, we will publish software, papers and corpora (when possible) under an open-source license; when we have to work with proprietary material, we will still make our annotation layers publicly available in separate files. The scientific contributions of this project will be published at conferences like ACL, EACL, Coling, NAACL, EMNLP, as well as in journals like NLE, CL, LRE, Quantitative Linguistics, Cognitive semantics, Language Variation and Change, Semantics&Pragmatics, Journal of Semantics.

4 Preliminary Results

In (Tahmasebi and Risse, 2017; Tahmasebi, 2013), we present a word sense change (SC) detection method based on induced word senses that are tracked over time. We consider each change, e.g., addition of a novel sense and a later change of that sense, separately, and allow several change events per word. For novel senses, we differentiate between neologism (new word with a new sense, e.g. *Internet*) and existing word with a novel senses (e.g. *Rock* with its *music* sense). In addition, we take into consideration stable senses for words that later add a novel sense. The unique perspective is that we consider each sense of a word separately and group senses, e.g. the *Rock-and-Roll lifestyle* sense is grouped with the *music* sense of *Rock* and the *industrial stone* sense is grouped with *natural stone*.

Our method can determine what changed and when the change took place, on a yearly time scale (222 years). 85 % of all changes are found with a delay of 9-11 years after it appears in our induced

senses and between 11-35 years after appearing in a dictionary. Example results for English as well as our test set can be found here: `https://doi.org/10.5281/zenodo.495572`.

We also present a method on detecting name changes in Tahmasebi et al. (2012b) and released a test set to facilitate comparison, and encourage work in the field (Tahmasebi et al., 2012a).

We investigated Word2Vec applied to the Kubhist dataset and concluded that the vector space produced by Word2Vec cannot be used directly for word sense change detection, due to the small size and quality of the dataset (Tahmasebi, 2018). We have ongoing work to correct OCR errors, normalize spelling variations (e.g., kwinna, kuinna, qvinna, qwinna, quinna and kona are all variations of the modern form kvinna) to increase the quality and boost the resulting vectors.

In the project 'Towards a knowledge-based Culturomics', we have created a Swedish sentiment lexicon, SenSaldo, and are working on tools for sentiment analysis for Swedish (Rouces et al.).

5 Project Description

We have organized the project into four work packages. Because we need evaluation for all areas, it is integrated in all our work, and we will define datasets, test sets, and evaluation methods in all Wps. We are currently writing a survey paper on computational approaches for automatic detection of lexical and semantic change for the Computational Linguistics journal. As a part this, we are gathering test sets used in existing literature and in parallel, we will also develop a public Swedish test set to facilitate comparison and reproducibility.

Our work with quantitatively reproducing existing studies for lexical and semantic change and our collaborations with researchers from outside fields will be a part of all work packages. The outcome of this collaboration will help improve our tools iteratively.

Wp1: Word Sense Induction for Swedish Automatically derived word senses are the basis for our continued work. For English, evaluated methods (Pantel and Lin, 2002; Brody and Lapata, 2009; Lau et al., 2012) exist and new methods are developed, the latest is the utilization of word embeddings (Nieto Piña and Johansson, 2015; Trask et al., 2015; Pelevina et al., 2016; Li and Jurafsky, 2015). For Swedish, these methods have not been tested at large scale and might perform less well due to smaller amounts of available text. In our previous work, we performed small-scale evaluation of embedding-based methods and the curvature clustering method (Dorow et al., 2005) on our historical newspaper corpus Kubhist. Neither of these have proven very effective and it remains to investigate the reasons, e.g. corpora size, level of noise or differences in morphology and syntax. We will adapt, develop and evaluate methods that are tailored to Swedish; and keep links to original contexts, hence providing evidence of each sense. Extracted senses can be linked to dictionaries to aid lexicographical work.

Wp2: Semantic Change Detection Existing approaches to change detection derive sense representations first and compare these over time. The most recent approaches make use of vector representations, providing a methodology that can answer when something has changed, but not how. In addition, vector representations have the target word as one unit without differentiating between a word's senses and work poorly on smaller corpora (which our historical corpora are, but also our modern corpora might be if we split them up in e.g. yearly subsets). Therefore, we need to go beyond vector representations in the following ways, (i) model word senses separately e.g. using multi-sense embeddings; (ii) allow all senses to change individually; (iii) differentiate change types (novel /out-dated senses, broadening/narrowing, related/un-related senses as well as stable senses); (iv) model the relation between the senses; and (v) handle smaller amounts of (noisy) text.

Our experiments, e.g., Tahmasebi et al (2017), show that a word like the *telephone* have senses in actual usage indicating that the word goes from being a *property of a building* to the belonging to each *household* to being a *tool for communication* much like a radio, television or a newspaper. The word *travel* has always had the *moving from point A to point B* sense but was mostly used for describing literature; people read about traveling. First in the early 20th century, the word usage added the sense of actual travel by means of train, boat and eventually airplane, the literature sense still being valid but less frequent. This usage change is extremely interesting, reflects cultural change rather than explicit sense change, and is found only if we differentiate and align a word's senses over time.

We will investigate count-based methods like SVD_{ppmi} (Hamilton et al., 2016) that do not suffer from randomness, and the dynamic embeddings, which increase robustness by automatically aligning embeddings at different time points, will be extended to account for multiple word senses. In addition, we will go beyond analyzing individual words to capture changes in semantic fields.

Using sense-differentiated embeddings targeted for small amounts of data, and our reduction techniques, we will answer what happens to *all* senses of a word over time, allowing senses to change individually while others stay stable for the same word. We will focus on precision, which is particularly important for uptake, i.e., presenting good quality results without too much noise. We will find (and rank) good text passages that exemplify each meaning and change. As proof of concept, we will revisit and enlarge several previous manually executed case studies on synchronic word sense variation and diachronic word sense change in Swedish, and look deeper into active lexical semantics research fields such as temperature, color, and smell. We will approach the under-investigated topic of how word senses change with different speed and processes in different linguistic sources (the inertia of change differs in e.g. fiction and newspapers) and different time periods (some, but not all, published work have noted an increased level of SC after the world wars, see e.g. Juola (2003)).

Wp3: Lexical replacements Automatic detection of lexical replacements over time has only previously targeted named entities. Previous work like Zhang et al. (2016a) and Berberich et al. (2009) have used word context rather than sense information, which is suitable for (unambiguous) named entities but not for words in general. Contrary to word sense change, we are targeting senses that are stable over time, so we require the induced senses to exhibit only minor variation to represent the same underlying sense. The problem requires (i) deriving word senses; (ii) tracking word senses over time; and (iii) linking words to word senses to find a word that has been used to replace another for a given sense (not necessarily for all senses of a word).

We will investigate multi-sense embeddings for labeling, extend current word pairs to create chains of change, e.g. $\textit{ipod} \xleftarrow[2012-2001]{} \textit{mp3 player} \xleftarrow[2001-1996]{} \textit{minidisc} \xleftarrow[1996-1992]{} \textit{discman} \xleftarrow[1984-1979]{} \textit{walkman}$ for the word sense of *mobile music player*; and assign validity periods for follow up applications (e.g., Information retrieval) and understanding. We will analyze systematic errors to automatically reduce false positives, e.g., *ear phones, tape, disc, Sony, Apple* for the above, thus rendering the results useless.

Non-automatic research is producing interesting hypotheses on word replacement, identifying replacement affecting factors such as frequency (Pagel et al., 2007) and synonym network density (Vejdemo and Hörberg, 2016) but much of this is based on small over-used databases of core vocabulary. We will address these hypotheses with large, diachronic datasets. For research on language change in general, we will investigate methods to link SC and LR changes (as well as spelling variations) to present all changes to a word and other words in its semantic field, in a scrollable and clickable map with links to relevant text passages, with the aim that it should be understandable and searchable.

Wp4: Application The application areas aim to highlight all kinds of change and apply to researchers as well as the public. The use cases will be integrated into the Korp pipeline (Borin et al., 2012) where possible, and in the Strix environment, a new interface under development that offers a close-reading capacity where texts, not sentences, and their temporal comparison are in focus.

Close reading (Simplify research) In this use case, we will help users (researchers and laymen) to firstly *find*, and secondly *understand* content in digital archives by (1) implementing features that suggest extension to search queries with relevant word replacements and their validity periods, and (2) in a text, highlight words that have changed their meaning. Changes will be accompanied with the original passages of text such that users can verify the results, or get new entry points into the corpus.

Distant reading (Quantify Research hypothesis) Many researchers are moving into DHSS (as seen by the increasing number of centers for Digital Humanities across Swedish universities), drawn in by the promise of large amounts of data and automatic methods for analyzing them. In this use case, we will collaborate with three research groups to help quantify their hypotheses.

Firstly, we will collaborate with Sarah Valdez at the Institute for Analytical Sociology in Norrköping, who works with analyzing differences in meaning for concepts in politics (e.g. democracy, freedom, immigration) for different political parties, this relates primarily to word sense induction on 20th century Swedish political party programs and election manifestos.

Secondly, we will collaborate with a group of concept historians led by Henrik Björck who are investigating the rate and spread of abstraction of the *market*, that goes from a concrete time and place to an abstract concept, like job or stock markets. We will study the interplay of these concept and investigate when *the market* becomes an agent that affects people rather than the other way around. Once the first abstract market has been established as a concept, does the process move faster and faster with new abstract markets and can we quantify this rate?

Thirdly, we will work with historical linguists led by Lena Rogström to show that scientific texts in the 18th century attributed human-like features to animals and plants, see e.g., Linné and Bjerkander. By applying semantic change detection to the Royal Science Academy texts, as well as to texts from other contemporary genres, we can help quantify the hypothesis, and find spread and change rate. This collaboration will take place after the digitization of the relevant texts, pending a funding application.

Personnel

The research team will consist of the principal investigator (PI), a researchers, an associate professor and a full professor, and a software developer specialized in language technology.

Principal investigator Nina Tahmasebi is a researcher at the University of Gothenburg and a postdoc at the Center for Digital Humanities (2018-2019). She is a cross-disciplinary researcher whose main interest lies in statistical NLP and data science, which she has tackled both from a computer science, data-driven perspective and from a NLP, knowledge-driven perspective. She was a national project manager in Swe-Clarin, coordinating the work on language technology tools for the DHSS and the local chair of the Nodalida2017 conference. Her research has focused on SC detection and her most influential contributions to the NLP field have been the following, all important for this project:

- She was among the first to study modern word sense induction algorithms and the influence of noisy data on historical text (Tahmasebi et al., 2013).
- Nina studied the SC problem and its properties in a computational context (Tahmasebi et al., 2011, 2012c; Tahmasebi and Risse, 2013) and was one of the first to define SC with respect to automatic detection (Tahmasebi, 2009, 2013; Tahmasebi and Risse, 2017).
- She studied the properties and automatic detection of named entity changes both with respect to news corpora (Tahmasebi et al., 2012b) and blogs (Holzmann et al., 2015)
- She studied language changes with respect to the field of Culturomics (Tahmasebi et al., 2015) and digital humanities (Tahmasebi et al., 2016; Borin et al., 2017) in the Swe-Clarin context.

The PI (80% first year, then 50% yearly) will lead the project and coordinate the dissemination efforts. In the initial stages of the project, available software and research software already developed by the PI will be used and later on, new software developed by the engineer with be used.

Researcher Richard Johansson is an associate professor in data science at the University of Gothenburg and Chalmers University of Technology, mainly focusing on data-driven NLP methods and their interplay with representations of linguistic knowledge. He has made research contributions to core NLP technologies such as dependency parsing, semantic role labeling, and discourse processing, as well as in NLP applications such as information extraction and sentiment analysis. PI of the project "Distributional methods to represent the meaning of frames and constructions" aimed at constructing automatic, corpus-based methods for inducing and representing word meaning using distributional semantics, outcomes highly relevant for this project. He will work 15% the first year, then 10% yearly.

Researcher Susanne Vejdemo is a researcher in lexical semantics affiliated with both Stockholm University and CUNY CSI (New York). She combines qualitative and quantitative approaches to lexical and semantic change over time. She has studied complex contact-induced LR in the color domain in seven related languages, as well as how LR intertwines with SC in the course of a generational shift. She has also developed a statistical model to explain why the speed of LR over time is different for different parts of the vocabulary (see Sec. 2 for references). She will work 40% yearly.

Researcher Maria Koptjevskaja Tamm is professor of general linguistics at the Dep. of Linguistics at Stockholm University. She is an internationally leading expert in semantically oriented typology, where she often combines synchronic and diachronic approaches. A large portion of her work focuses on the interplay between lexical and grammatical semantics. MKT was the PI and coordinator of the collaborative project "Core vocabulary in a typological perspective: semantic shifts and form /meaning correlations (TypVoc)", involving partners from five countries. She is currently authoring the book "Temperature in language: typology, evolution and extended uses" where cross-linguistic variability in semantic systems, in space and time are explored. She will work 7.5% yearly.

Software engineer. We will make use of a software engineer to implement and integrate the use case scenarios into Korp and Strix, to increase dissemination of the research in this project and make use of the rich research infrastructural efforts that are ongoing in the national infrastructure of Språkbanken. The engineer will work 15% yearly with parallel work at Språkbanken for 25%.

Time Plan: The project is planned from January 2019 to the end of December 2022. A rough time plan is given in the table showing the distribution of work in each area (• active , ◦ planning / survey).

Year	Wp2	Wp3	Wp4	Wp5
2019	•	•		•
2020		•	◦	•
2021		•	•	•
2022			•	•

Research Environment: The team is distributed over Språkbanken at the Department of Swedish at the University of Gothenburg (GU), Computer Science at Chalmers (CTH) and Linguistics at Stockholm University (SU). The department of Swedish fits the project well since it is home of the national infrastructure of Språkbanken. There will be synergies with ongoing projects; *Towards a Knowledge-based Culturomics* (a project aimed at developing LT resources and methods for deep linguistic processing of Swedish text), and *Swe-Clarin* (an LT-based eScience infrastructure for DHSS). The collaboration between SU and Språkbanken unites a lexical semantics expertise with a strong NLP group, a collaboration that is, to the best of our knowledge, unique in the world when it comes to its focus on diachronic and synchronic lexical semantics and language variation and change. The team at SU have ongoing collaborations with Gavagai, a text analytics company, on sentiment analysis and quantitative semantics. **International collaboration:** The PI has ongoing collaborations with Adam Jatowt, a professor at the Kyoto University who focuses on computational history and who has made valuable contributions to language change detection, starting with a joint survey paper for the CL journal.

Need for infrastructure: The research to be conducted in this project has a dual role with respect to the recently established national research infrastructure Språkbanken: (1) it will use the existing language tools and language resources (diachronic corpora and lexical resources) made available by the Språkbanken Text division; and (2) being a project with a strong LT component, it will produce new resources, methods and tools which can be integrated into the infrastructure to support future research on Swedish historical linguistics. To optimally support both these aims, the PI and Språkbanken have agreed on an arrangement where the PI will work in Språkbanken at 30% (funded by Språkbanken) for the duration of the project, as coordinator between the project and Språkbanken, and further that a systems developer will be shared (15% project +25% Språkbanken).

Dissemination plan: We will disseminate the results of this project through publications, conference presentations and tutorials; at least two workshops (the second collocated with an appropriate conference) to bring together researchers from NLP with cultural and historical semantics; public software for use at Språkbanken; during the annual autumn workshop of Språkbanken; and through our collaboration with researchers from other fields and universities. We will network with high school teachers through the Linguistic Olympiad network.

References

A. Anand, S. Bedathur, K. Berberich, and R. Schenkel. Index maintenance for time-travel text search. In *SIGIR*, 2012.

R. Bamler and S. Mandt. Dynamic word embeddings. In *ICML*, 2017.

M. Baroni, G. Dinu, and G. Kruszewski. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL*, 2014.

S. Bartunov, D. Kondrashkin, A. Osokin, and D. Vetrov. Breaking sticks and ambiguities with adaptive skip-gram. In *AISTATS*, 2016.

P. Basile, A. Caputo, R. Luisi, and G. Semeraro. Diachronic analysis of the italian language exploiting google ngram. In *Italian Conf. on Comp. Linguistics*, 2016.

K. Berberich, S. Bedathur, T. Neumann, and G. Weikum. A time machine for text search. In *SIGIR*, 2007.

K. Berberich, S. J. Bedathur, M. Sozio, and G. Weikum. Bridging the Terminology Gap in Web Archive Search. In *WebDB'09 workshop*, 2009.

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

L. Borin, N. Tahmasebi, and E. Volodina et al. Swe-clarin: Language resources and technology for digital humanities. In *Digital Humanities 2016. Vol-2021*, 2017.

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

P. Cook and S. Stevenson. Automatically Identifying Changes in the Semantic Orientation of Words. In *LREC*, 2010.

P. Cook, J. H. Lau, D. McCarthy, and T. Baldwin. Novel word-sense identification. In *COLING*, 2014.

M. C. Cooper. A Mathematical Model of Historical Semantics and the Grouping of Word Meanings into Concepts. *Computational Linguistics*, 32(2):227–248, 2005.

B. Dorow, J.-p. Eckmann, and D. Sergi. Using curvature and markov clustering in graphs for lexical acquisition and word sense discrimination. In *MEANING*, 2005.

L. Frermann and M. Lapata. A bayesian model of diachronic meaning change. *TACL*, 4:31–45, 2016.

R. Goel, S. Soni, N. Goyal, J. Paparrizos, H. Wallach, F. Diaz, and J. Eisenstein. The social dynamics of language change in online networks. In *ICSI*, 2016.

K. Gulordava and M. Baroni. A distributional similarity approach to the detection of semantic change in the google books ngram corpus. In *GEMS workshop*, 2011.

W. L. Hamilton, J. Leskovec, and D. Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *ACL*, 2016.

J. Hellrich and U. Hahn. Bad company - neighborhoods in neural embedding spaces considered harmful. In *COLING*, 2017.

A. Herbelot and M. Baroni. High-risk learning: acquiring new word vectors from tiny data. In *EMNLP*, 2017.

H. Holzmann, N. Tahmasebi, and T. Risse. Named entity evolution recognition on the blogosphere. *International Journal on Digital Libraries*, 15(2-4):209–235, 2015.

R. Johansson and L. Nieto Piña. Embedding a semantic network in a word space. In *NAACL*, 2015.

P. Juola. The Time Course of Language Change. *Computers and the Humanities*, 37(1):77–96, 2003.

A. C. Kaluarachchi, A. S. Varde, S. J. Bedathur, G. Weikum, J. Peng, and A. Feldman. Incorporating terminology evolution for query translation in text retrieval with association rules. In *CIKM*, 2010.

N. Kanhabua and K. Nørvåg. Exploiting time-based synonyms in searching document archives. In *JCDL*, 2010.

Y. Kim, Y.-I. Chiu, K. Hanaki, D. Hegde, and S. Petrov. Temporal analysis of language through neural language models. In *LACSS*, 2014.

M. Koptjevskaja-Tamm. *Introducing the linguistics of temperature*, volume 107 of *Typological Studies in Language*, page 1–40. John Benjamins Publishing Company, 2015.

Kubhist. Språkbanken, University of Gothenburg. `https://spraakbanken.gu.se/korp/?mode=kubhist`.

V. Kulkarni, R. Al-Rfou, B. Perozzi, and S. Skiena. Statistically significant detection of linguistic change. In *WWW*, 2015.

J. H. Lau, P. Cook, D. McCarthy, D. Newman, and T. Baldwin. Word Sense Induction for Novel Sense Detection. In *EACL*, 2012.

J. Li and D. Jurafsky. Do multi-sense embeddings improve natural language understanding? In *EMNLP*, 2015.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

S. Mitra, R. Mitra, M. Riedl, C. Biemann, A. Mukherjee, and P. Goyal. That's sick dude!: Automatic identification of word sense change across different timescales. In *ACL*, 2014.

S. Mitra, R. Mitra, S. K. Maity, M. Riedl, C. Biemann, P. Goyal, and A. Mukherjee. An automatic approach to identify word sense changes in text media across timescales. *NLE*, 21(05):773–798, 2015.

S. Morsy and G. Karypis. Accounting for language changes over time in document similarity search. *Trans. on Information Systems*, 35(1):1, 2016.

L. T. Nguyen, P. Wu, W. Chan, W. Peng, and Y. Zhang. Predicting collective sentiment dynamics from time-series social media. In *WISDOM*, 2012.

L. Nieto Piña and R. Johansson. A simple and efficient method to generate word sense representations. In *RANLP*, 2015.

M. Pagel, Q. D. Atkinson, and A. Meade. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature*, 449:717–720, 2007.

P. Pantel and D. Lin. Discovering word senses from text. In *KDD*, 2002.

H. Paul. *Prinzipien der Sprachgeschichte*. Max Niemeyer, 1886.

M. Pelevina, N. Arefyev, C. Biemann, and A. Panchenko. Making sense of word embeddings. In *RepL4NLP*, 2016.

R. Pustet. *The syntax of temperature predications*, page 889–916. John Benjamins Publishing Company, 2015.

J. Rouces, N. Tahmasebi, L. Borin, and S. R. Eide. SenSALDO: Creating a sentiment lexicon for Swedish. In *LREC 2018, forthcoming*.

E. Sagi, S. Kaufmann, and B. Clark. Semantic density analysis: comparing word meaning across time and phonetic space. In *GEMS workshop*, 2009.

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

H. Schütze. Automatic word sense discrimination. *Comp. Ling.*, 24(1):97–123, 1998.

E. Sweetser. *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge University Press, July 1991.

N. Tahmasebi. Automatic Detection of Terminology Evolution. In *OTM 2009 workshop*, 2009.

N. Tahmasebi. *Models and Algorithms for Automatic Detection of Language Evolution*. PhD thesis, Gottfried Wilhelm Leibniz Universität Hannover, 2013.

N. Tahmasebi. A Study on Word2Vec on a Historical Swedish Newspaper Corpus. In *DHN*, 2018.

N. Tahmasebi and T. Risse. The role of language evolution in digital archives. In *SDA workshop*, 2013.

N. Tahmasebi and T. Risse. Finding individual word sense changes and their delay in appearance. In *RANLP*, 2017.

N. Tahmasebi, T. Risse, and S. Dietze. Towards automatic language evolution tracking, A study on word sense tracking. In *EvoDyn workshop*, 2011.

N. Tahmasebi, G. Gossen, N. Kanhabua, H. Holzmann, and T. Risse. Named Entity Evolution Dataset. `https://www.l3s.de/neer-dataset/`, 2012a.

N. Tahmasebi, G. Gossen, N. Kanhabua, H. Holzmann, and T. Risse. NEER: An Unsupervised Method for Named Entity Evolution Recognition. In *COLING*, 2012b.

N. Tahmasebi, G. Gossen, and T. Risse. Which Words Do You Remember? Temporal Properties of Language Use in Digital Archives. In *TPDL*, 2012c.

N. Tahmasebi, K. Niklas, G. Zenz, and T. Risse. On the applicability of word sense discrimination on 201 years of modern english. *IJDL*, 13(3-4):135–153, 2013.

N. Tahmasebi, L. Borin, G. Capannini, and D. Dubhashi et al. Visions and open challenges for a knowledge-based culturomics. *IJDL*, 15(2-4):169–187, 2015.

N. Tahmasebi, L. Borin, C. Jordan, and S. Ekman. SWE-CLARIN – the Swedish CLARIN project – aims and activities. In *DHN*, 2016.

A. Trask, P. Michalak, and J. Liu. sense2vec - A fast and accurate method for word sense disambiguation in neural word embeddings. *CoRR*, abs/1511.06388, 2015.

E. Traugott and R. B. Dasher. *Regularity in semantic change*. Cambridge University Press, Cambridge ; New York, 2002.

P. D. Turney and P. Pantel. From frequency to meaning: Vector space models of semantics. *JAIR*, 37:141–188, 2010.

S. Ullman. *The principles of semantics.* Blackwell Publishers, Oxford, 1957.

S. Vejdemo. *Triangulating Perspectives on Lexical Replacement: From Predictive Statistical Models to Descriptive Color Linguistics*. Stockholm University, 2017.

S. Vejdemo and T. Hörberg. Semantic Factors Predict the Rate of Lexical Replacement of Content Words. *PLOS ONE*, pages 1–15, Jan. 2016.

A. Viberg. *Studier i kontrastiv lexikologi: perceptionsverb*. Stockholms Universitet, Inst. för lingvistik, 1980.

D. T. Wijaya and R. Yeniterzi. Understanding semantic change of words over centuries. In *DETECT*, 2011.

Y. Zhang, A. Jatowt, S. S. Bhowmick, and K. Tanaka. The past is not a foreign country: Detecting semantically similar terms across time. *KDE*, 2016a.

Y. Zhang, A. Jatowt, and K. Tanaka. Detecting evolution of concepts based on cause-effect relationships in online reviews. In *WWW*, 2016b.

Budget och forskningsresurser

Aktivitetsgrad i projektet*

Roll i projektet	Namn	Procent av heltid
1 Projektledare	Nina Tahmasebi	50%
2 Medverkande forskare	Richard Johansson	11%
3 Medverkande forskare	Maria K Tamm	7%
4 Medverkande forskare	Susanne Vejdemo	40%
5 Övrig ej disputerad personal	systemutvecklare oklart vem	15%

Löner inklusive sociala avgifter

Roll i projektet	Namn	Procent av lönen
1 Projektledare	Nina Tahmasebi	38%
2 Medverkande forskare	Richard Johansson	11%
3 Medverkande forskare	Maria K Tamm	7%
4 Medverkande forskare	Susanne Vejdemo	40%
5 Övrig ej disputerad personal	systemutvecklare oklart vem	15%
Totalt		0

	2019	2020	2021	2022	Totalt
1	0	366 500	375 663	385 054	1 127 217
2	139 967	95 177	97 081	99 022	431 247
3	108 765	83 613	79 990	58 564	330 932
4	381 615	248 917	255 140	261 518	1 147 190
5	185 069	67 418	68 766	70 142	391 395
Totalt	815 416	861 625	876 640	874 300	3 427 981

Lokaler

Typ av lokal	2019	2020	2021	2022	Totalt
1 arbetsrum	62 510	103 719	105 764	107 857	379 850
Totalt	62 510	103 719	105 764	107 857	379 850

Driftskostnader

Driftskostnader	Beskrivning	2019	2020	2021	2022	Totalt
1 Resor, konferenser		85 000	40 000	25 000	30 000	180 000
2 IT-avgift	IT service, system	9 500	15 000	15 000	15 000	54 500
Totalt		94 500	55 000	40 000	45 000	234 500

Avskrivningar utrustning

Avskrivning	Beskrivning	2019	2020	2021	2022
Ingen information ifylld					

Total budget*

	Specificerade kostnader	2019	2020	2021	2022	Totalt, sökt
1	Löner inkl. sociala avgifter	815 416	861 625	876 640	874 300	3 427 981
2	Driftskostnader	94 500	55 000	40 000	45 000	234 500
3	Avskrivningar utrustning					0
4	Lokaler	62 510	103 719	105 764	107 857	379 850
5	Delsumma	972 426	1 020 344	1 022 404	1 027 157	4 042 331
6	Indirekta kostnader	524 166	475 542	476 050	471 589	1 947 347
7	Total projektkostnad	1 496 592	1 495 886	1 498 454	1 498 746	5 989 678

	Annan kostnad	Total kostnad
1		3 427 981
2		234 500
3		0
4		379 850
5	0	4 042 331
6		1 947 347
7	0	5 989 678

Motivering av sökt budget*

Alla involverade forskare kommer dels att bidra till att bygga upp systemet samt publicera projektets resultat vid vetenskapliga konferenser och tidskrifter.

Nina Tahmasebi (NT) kommer att leda projektet och utföra och analysera undersökningar. NT kommer också att utveckla delar av de datorprogram som kommer att användas i projektet. Projektet föreslås bekosta 50% av NTs heltidstjänst i perioden januari 2020 – december 2022. De övriga 50\% kommer att täckas av arbete inom Språkbankens forskningsinfrastruktur (30%) och institutionstjänstgöring. Under 2019 bekostas NTs tjänst av en postdoc på Center for Digital Humaniora, där 80% kan ägnas åt projektet och 20% åt institutionstjänstgöring.

Richard Johansson (RJ) kommer att bidra till projektet med sin kunskap inom distributionella semantiska metoder. Projektet föreslås bekosta 15% av RJs heltidstjänst under första året, därefter 10% resterande tre år. De övriga 85% kommer att täckas av institutionstjänstgöring och och andra projekt.

Susanne Vejdemo (SV) kommer att utföra och analysera undersökningar ihop med NT, med utgångspunkt i lexikaltypologisk forskning. Projektet föreslås bekosta 40% av SVs heltidstjänst under projektets hela löptid.

Maria Koptjevskaja Tamm (MKT) kommer att rådge projektet med utblick från den absoluta framkanten för lexikaltypologisk forskning, utvärdera forskningsresultat och designa use cases för testning. Projektet föreslås bekosta 7.5% av MKTs heltidstjänst under projektet. De övriga 90% kommer att täckas av institutionstjänstgöring och och andra projekt.

En systemutvecklare kommer att implementera och integrera metoderna som utvecklas inom projektet, samt integrera kod som utvecklas av NT, RJ och SV inom Korp, Språkbankens forskningsinfrastruktur. Systemutvecklaren föreslås bekostas till 15% av projektet och arbeta parallellt till 25% på relaterad forskningsinfrastruktur på Språkbanken.

Projektets resultat kommer att publiceras vid internationella konferenser. Vi räknar med ungefär 3 konferensresor per år. Vi planerar för projektmedlemmarna att träffas två gånger per år, antingen i Stockholm eller i Göteborg. Utöver detta kommer vi att arrangera minst en workshop för att föra ihop datavetare, språkteknologer samt lingvister och framförallt semantiker där vi hoppas att kunna bygga upp ett starkt forskningsnätverk i Sverige och Norden.

Annan finansiering för detta projekt

Finansiär	Sökande/projektledare	Typ av bidrag	Status	Dnr eller motsv.
	2019	2020	2021	2022
Ingen information ifylld				

Publikationer

Sökandes publikationslista (pdf)*

Se nästa sida för bilaga.

Selection of publications

Peer-reviewed original articles

Nina Tahmasebi, Lars Borin, Gabriele Capannini, Devdatt Dubhashi, Peter Exner, Markus Forsberg, Gerhard Gossen, Fredrik D. Johansson, Richard Johansson, Mikael Kågebäck, Olof Mogren, Pierre Nugues, and Thomas Risse. 2015. Visions and open challenges for a knowledge-based culturomics. *International Journal on Digital Libraries*, 15(2-4), pages 169–187.

I contributed one full chapter, coordinated the writing and edited the text. The paper shows the need for detecting lexical and semantic change for the field of Culturomics (the study of language and cultural phenomena over time in large scale corpora.)

Helge Holzmann, **Nina Tahmasebi**, and Thomas Risse. 2015. Named entity evolution recognition on the blogosphere. *International Journal on Digital Libraries*, 15(2-4):209–235.

I contributed ideas and writing. The paper shows an alternative method for finding named entity changes (name replacements) using external resources like Wikipedia.

Nina Tahmasebi, Kai Niklas, Gideon Zenz, and Thomas Risse. On the applicability of word sense discrimination on 201 years of modern english. 2013. *International Journal on Digital Libraries*, 13(3-4):135–153.

I contributed ideas, experiments, code and writing. This paper studies the quality of modern word sense induction on historical corpora, and the impact of OCR errors and spelling variations for the analysis. The word sense induction algorithm evaluated in this paper has been preliminarily evaluated for Swedish, without fully convincing results, and will be used as a starting point and a possible baseline for our continued work.

Peer-reviewed conference contributions

Jacobo Rouces, **Nina Tahmasebi**, Lars Borin, Stian Rødven Eide 2018. SenSALDO: Creating a Sentiment Lexicon for Swedish. To appear in *Language Resources and Evaluation Conference (LREC 2018)*.

I contributed ideas, evaluation and writing. This is a Swedish sentiment lexicon based on SALDO that lists word senses rather than conflating all senses of a word into one, and will be extended in this project.

Nina Tahmasebi 2018. A Study on Word2Vec on a Historical Swedish Newspaper Corpus. In *Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018)* .

I contributed ideas, experiments, code and writing. This paper shows the limited applicability of neural embedding methods like Word2Vec on smaller datasets (such as Swedish historical newscorpora) with a high rate of OCR errors and spelling variations.

Nina Tahmasebi and Thomas Risse. 2017. Finding Individual Word Sense Changes and their Delay in Appearance. In *Recent Advances in Natural Language Processing, RANLP 2017*.

I contributed ideas, experiments, code and writing. This paper presents methods for detecting semantic change for words on the basis of induced word senses. The reduction techniques and the main framework will be used as a starting point for our work on semantic change for Swedish.

Mikael Kågebäck, Olof Mogren, **Nina Tahmasebi** and Devdatt Dubhashi. 2014. Extractive summarization using continuous vector space models. In *Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, pages 31–39.

I contributed ideas and writing. This paper uses neural embeddings for creating text summaries (automatically capturing the essence of a set of texts). Summaries could be used for offering textual evidence for lexical or semantic change.

Nina Tahmasebi, Gerhard Gossen, Nattiya Kanhabua, Helge Holzmann, and Thomas Risse. 2012. NEER: An Unsupervised Method for Named Entity Evolution Recognition. In *24th International Conference on Computational Linguistics (COLING 2012)*, pages 2553–2568.

I contributed ideas, experiments, code and writing. A method for detecting named entity change (name replacements) only relying on the corpus (without use of external resources like Wikipedia). Burst detection is used for identifying important periods for a name to search for replacements, a technique that will be evaluated also for words from other classes.

Nina Tahmasebi, Gerhard Gossen, and Thomas Risse. 2012. Which words do you remember? Temporal properties of language use in digital archives. In *Theory and Practice of Digital Libraries*, volume 7489 of *Lecture Notes in Computer Science*, pages 32–37.

I contributed ideas, experiments, code and writing. A study on the properties of words in text in social media compared to other sources like newspapers, to show a higher variety in word usage over time in social media (similar to spoken language).

Peer-reviewed books and book chapters

Nina N. Tahmasebi. 2013. *Models and Algorithms for Automatic Detection of Language Evolution*. PhD thesis, Gottfried Wilhelm Leibniz Universität Hannover.

I contributed ideas, experiments, code and writing. Here both semantic change and lexical change is studied and many of the techniques presented here will be evaluated for this project directly, and some will be used as starting points. All techniques here have been tested for English, large-scale corpora and the techniques will need adaptation for the Swedish language and smaller scale texts.

Total number of publications

Peer-reviewed original articles

Nina Tahmasebi, Lars Borin, Gabriele Capannini, Devdatt Dubhashi, Peter Exner, Markus Forsberg, Gerhard Gossen, Fredrik D. Johansson, Richard Johansson, Mikael Kågebäck, Olof Mogren, Pierre Nugues, and Thomas Risse. 2015. Visions and open challenges for a knowledge-based culturomics. *International Journal on Digital Libraries*, 15(2-4), pages 169–187.

Helge Holzmann, **Nina Tahmasebi**, and Thomas Risse. Named entity evolution recognition on the blogosphere. 2015. *International Journal on Digital Libraries*, 15(2-4):209–235.

Nina Tahmasebi, Kai Niklas, Gideon Zenz, and Thomas Risse. 2013. On the applicability of word sense discrimination on 201 years of modern english. *International Journal on Digital Libraries*, 13(3-4):135–153.

Gideon Zenz, **Nina Tahmasebi**, and Thomas Risse. 2013. Towards mobile language evolution exploitation. *Multimedia Tools and Applications*, 66(1):147–159.

Bogdan Pogorelc, Artur Lugmayr, Björn Stockleben, Radu-Daniel Vatavu, **Nina Tahmasebi**, Estefanıa Serral, Emilija Stojmenova, Bojan Imperl, Thomas Risse, Gideon Zenz, and Matjaz Gams. 2012. Ambient bloom: new business, content, design and models to increase the semantic ambient media experience. *Multimedia Tools and Applications*, 66(1):7–32.

Peer-reviewed books and book chapters

Nina N. Tahmasebi. 2013. *Models and Algorithms for Automatic Detection of Language Evolution*. PhD thesis, Gottfried Wilhelm Leibniz Universität Hannover.

Peer-reviewed conference contributions

Jacobo Rouces, **Nina Tahmasebi**, Lars Borin, Stian Rødven Eide 2018. SenSALDO: Creating a Sentiment Lexicon for Swedish. To appear in *Language Resources and Evaluation Conference (LREC 2018)*.

Jacobo Rouces, **Nina Tahmasebi**, Lars Borin, Stian Rødven Eide 2018. Generating a Gold Standard for a Swedish Sentiment Lexicon. To appear in *Language Resources and Evaluation Conference (LREC 2018)*.

Nina Tahmasebi 2018. A Study on Word2Vec on a Historical Swedish Newspaper Corpus. In *Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018)* .

Jacobo Rouces, Lars Borin, **Nina Tahmasebi** and Stian Rødven Eide 2018. Defining a Gold Standard for a Swedish Sentiment Lexicon: Towards Higher-Yield Text Mining in the Digital Humanities. In *Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018)* .

Nina Tahmasebi and Thomas Risse. 2017. Finding Individual Word Sense Changes and their Delay in Appearance. In *Recent Advances in Natural Language Processing, RANLP 2017*, pages 741–749.

Sallam Abualhaija, **Nina Tahmasebi**, Diane Forin, and Karl-Heinz Zimmermann. 2017. Parameter Transfer across Domains for Word Sense Disambiguation. In *Recent Advances in Natural Language Processing, RANLP 2017*, pages 1–8.

Nina Tahmasebi and Thomas Risse. 2017. On the uses of word sense change for research in the digital humanities. In *21th International Conference on Theory and Practice of Digital Libraries 2017 (TPDL 2017)*, pages 246–257.

Lars Borin, **Nina Tahmasebi**, Elena Volodina, Stefan Ekman, Caspar Jordan, Jon Viklund, Beáta Megyesi, Jesper Näsman, Anne Palmér, Mats Wirén, Kristina N Björkenstam, Gintarė Grigonytė, Sofia Gustafson Capková, and Tomasz Kosiński. 2017. Swe-Clarin: Language Resources and Technology for Digital Humanities. In *Digital Humanities Symposium Proceedings*, pages 29–51.

Nina Tahmasebi, Lars Borin, Caspar Jordan, and Stefan Ekman. 2016. Swe-clarin – the Swedish Clarin project – aims and activities. In *Digital Humanities in the Nordic countries, Oslo*, pages 122–123.

Bianka Nusko, **Nina Tahmasebi**, and Olof Mogren. 2016. Building a sentiment lexicon for swedish. In

Linköping Electronic Conference Proceedings. Digital Humanities 2016. From Digitization to Knowledge 2016: Resources and Methods for Semantic Processing of Digital Works/Texts, volume 126, pages 32–37.

Stian Rødven Eide, **Nina Tahmasebi**, and Lars Borin. The swedish culturomics gigaword corpus: A one billion word swedish reference dataset for NLP. In *Linköping Electronic Conference Proceedings. Digital Humanities 2016. From Digitization to Knowledge 2016: Resources and Methods for Semantic Processing of Digital Works/Texts*, volume 126, pages 8–12.

Malin Ahlberg, Peter Andersson, Markus Forsberg, and **Nina Tahmasebi**. 2015. A case study on supervised classification of Swedish pseudo-coordination. In *20th Nordic Conference of Computational Linguistics, NODALIDA 2015*, pages 11–19.

Mikael Kågebäck, Olof Mogren, **Nina Tahmasebi** and Devdatt Dubhashi. 2014. Extractive summarization using continuous vector space models. In *Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, pages 31–39.

Dimitris Spiliotopoulos, Thomas Risse, and **Nina Tahmasebi**. 2013. Sms 2013 pc co-chairs message. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 587–587.

Nina Tahmasebi and Thomas Risse. 2013. The role of language evolution in digital archives. In *3rd International Workshop on Semantic Digital Archives*, pages 16–27.

Helge Holzmann, **Nina Tahmasebi**, and Thomas Risse. 2013. Blogneer: Applying named entity evolution recognition on the blogosphere. In *3rd International Workshop on Semantic Digital Archives*, volume 1091 of *CEUR Workshop Proceedings*, pages 28–39.

Elena Demidova, N Barbieri, Stefan Dietze, Adam Funk, Gerhard Gossen, Diana Maynard, N Papailiou, V Plachouras, W Peters, Y Stavrakas, **Nina Tahmasebi**, et al. 2013. Analysing entities, topics and events in community memories. In *International Workshop on Archiving Community Memories*.

Nina Tahmasebi, Gerhard Gossen, Nattiya Kanhabua, Helge Holzmann, and Thomas Risse. 2012. NEER: An Unsupervised Method for Named Entity Evolution Recognition. In *24th International Conference on Computational Linguistics (COLING 2012)*, pages 2553–2568.

Helge Holzmann, Gerhard Gossen, and **Nina Tahmasebi**. 2012. *fokas*: Formerly Known As – A Search Engine Incorporating Named Entity Evolution. In *24th International Conference on Computational Linguistics (COLING 2012), Demonstration Papers*, pages 215–222.

Nina Tahmasebi, Gerhard Gossen, and Thomas Risse. 2012. Which words do you remember? Temporal properties of language use in digital archives. In *Theory and Practice of Digital Libraries*, volume 7489 of *Lecture Notes in Computer Science*, pages 32–37.

Nina Tahmasebi, Thomas Risse, and Stefan Dietze. 2011. Towards automatic language evolution tracking, A study on word sense tracking. In *Workshop on Knowledge Evolution and Ontology Dynamics (EvoDyn'11)*.

Thomas Risse, Stefan Dietze, Diana Maynard, **Nina Tahmasebi**, and Wim Peters. 2011. Using Events for Content Appraisal and Selection in Web Archives. In *Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011)*.

Christopher Kunz, **Nina Tahmasebi**, Thomas Risse, and Matthew Smith. 2011. Detecting Credential Abuse in the Grid Using Bayesian Networks. In *12th IEEE/ACM International Conference on Grid Computing (GRID)*, pages 114 –120.

Nina Tahmasebi, Kai Niklas, Thomas Theuerkauf, and Thomas Risse. 2010. Using word sense discrimination on historic document collections. In *Joint International Conference on Digital Libraries, (JCDL'10)*, pages 89–98.

Gideon Zenz, **Nina Tahmasebi**, and Thomas Risse. 2010. Language Evolution On The Go. In *the 3rd International Workshop on Semantic Ambient Media Experience (SAME 2010)*.

Nina Tahmasebi. 2009. Automatic detection of terminology evolution. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, volume 5872 of *Lecture Notes in Computer Science*, pages 769–778.

Nina Tahmasebi, Sukriti Ramesh, and Thomas Risse. 2009. First results on detecting term evolutions. In *9th International Web Archiving Workshop*.

Nina Tahmasebi, Tereza Iofciu, Thomas Risse, Claudia Niederée, and Wolf Siberski. 2008. Terminology evolution in web archiving: Open issues. In *8th International Web Archiving Workshop*.

Peer-reviewed edited volumes

Jörg Tiedemann and **Nina Tahmasebi**, editors. 2017. *21st Nordic Conference on Computational Linguistics, NODALIDA 2017, Gothenburg, Sweden*. Association for Computational Linguistics.

Popular science publications and Open Access computer programs

Nina Tahmasebi and Thomas Risse. 2017. Word Sense Change Test Set. `https://doi.org/10.5281/zenodo.495572`.

Nina Tahmasebi, Gerhard Gossen, Nattiya Kanhabua, Helge Holzmann, and Thomas Risse. 2012. Named Entity Evolution Dataset. `http://www.l3s.de/neer-dataset/`.

Bifoga samtliga eventuella medverkande forskares publikationslistor i pdf-format, upprättad enligt instruktionerna i utlysningstexten. Du kan endast bifoga en fil på maximalt 10 MB. Klicka först på mappsymbolen för att söka fram filen på din hårddisk, klicka sedan på plustecknet för att ladda upp filen i formuläret.

Medverkande forskares publikationslistor (pdf)

Se nästa sida för bilaga.

Peer-reviewed original articles

Susanne Vejdemo and Thomas Hörberg. 2016. Semantic factors predict the rate of lexical replacement of content words. In *PLoS ONE 11(1): e0147924*.

Susanne Vejdemo, Carsten Levisen, Thorhalla G. Beck, Cornelia von Scherpenberg, Åshild Næss, Martina Zimmerman, Linnaea Stockall, and Matthew Whelpton. 2015. Two kinds of pink: Development and difference in germanic colour semantics. In *Language Sciences. 49:19–34. 10.1016/j.langsci.2014.07.007.*

Mikael Vejdemo-Johansson, **Susanne Vejdemo**, and Carl-Henrik Ek. 2014. Comparing distributions of color words: Pitfalls and metric choices. In *PLoS ONE 9(2):e89184*.

Peer-reviewed conference contributions

Susanne Vejdemo. 2016. To database meaning: Building the typological database of temperature terms. In *Annual Meeting of the Michigan Linguistic Society, the University of Michigan*

Susanne Vejdemo. 2010. Cross-linguistic lexical change: Why, how and how fast? In *Proceedings of WIGL 2010*, volume 8 of *LSO Working Papers in Linguistics*.

Peer-reviewed books and book chapters

Susanne Vejdemo. 2017. *Triangulating Perspectives on Lexical Replacement: From Predictive Statistical Models to Descriptive Color Linguistics.* PhD thesis, Stockholm University, Stockholm.

Susanne Vejdemo and Sigi Vandewinkel. 2016. Extended uses of temperature terms across languages. In Maria Koptjevskaja-Tamm and Päivi Juvonen, editors, *Lexicotypological approaches to semantic shifts and motivation patterns in the lexicon*, pages 249–284.

Hunter Lockwood and **Susanne Vejdemo**. 2015. "There is no thermostat in the forest" – the ojibwe temperature term system. In Maria Koptjevskaja-Tamm, editor, *The Linguistics of Temperature*, volume 107 of *Typological Studies in Language*, pages 721–741. John Benjamins Publishing Company.

Other publications

Susanne Vejdemo. 2007. *Skarp, vass och sharp – semantiska relationer hos tre perceptionsadjektiv*. Master thesis, University of Stockholm.

Peer-reviewed original articles

Maria Koptjevskaja-Tamm, Martine Vanhove and Peter Koch. 2007 Typological approaches to lexical semantics. In *Linguistic Typology, 11-1: 159 – 186*.

Peer-reviewed original articles/Peer-reviewed book chapters

Maria Koptjevskaja-Tamm and Magnus Sahlgren. 2014. Temperature in the Word Space: sense exploration of temperature expressions using word-space modeling. In Szmrecsanyi, B. and B. Wälchli (eds), *Linguistic variation in text and speech, within and across languages* (Series: Linguae et Litterae: Publications of the School of Language and Literature, Freiburg Institute for Advanced Studies). Berlin: de Gruyter, 231 – 267.

Maria Koptjevskaja-Tamm, Dagmar Divjak and Ekaterina Rakhilina. 2010. Aquamotion verbs in Slavic and Germanic: A case study in lexical typology. In Driagina-Hasko, V. and R. Perelmutter (eds.), *New approaches to Slavic verbs of motion*. Amsterdam: Benjamins, 315-342.

Maria Koptjevskaja-Tamm. 2009. "A lot of grammar with a good portion of lexicon": towards a typology of partitive and pseudo-partitive nominal constructions. In Helmbrecht, J., N. Yoko, S. Yong-Min, S. Skopeteas and E.Verhoeven (eds.), *Form and Function in Language Research*. Berlin: Mouton de Gruyter, 329 – 346.

Maria Koptjevskaja-Tamm. 2008. Approaching lexical typology. In Vanhove, M. (ed.), *From polysemy to semantic change: a typology of lexical semantic associations*. Amsterdam: Benjamins, 3–52.

Peer-reviewed edited volumes

Päivi Juvonen and **Maria Koptjevskaja-Tamm** (eds.). 2016. *The lexical typology of semantic shifts*. Berlin – New York: de Gruyter Mouton.

Maria Koptjevskaja-Tamm (ed.) 2015 *The linguistics of temperature.* Amsterdam: John Benjamins.

Maria Koptjevskaja-Tamm and Martine Vanhove (eds.). 2012 New directions in lexical typology. In *A special issue of Linguistics, 50–3: 373–743*.

Peer-reviewed book chapters / Research review articles

Maria Koptjevskaja-Tamm and Henrik Liljegren. 2017. *Lexical semantics and areal linguistics*. In Hickey, R. (ed.), The Cambridge Handbook of Areal Linguistics. Cambridge: Cambridge University Press, 204 – 236.

Maria Koptjevskaja-Tamm. 2015. *Semantic typology*. In Dabrowska, E. and D. Divjak (eds.), Handbook of Cognitive Linguistics, 453 – 472. Handbooks of Linguistics and Communication Sciences (HSK), 39, Berlin – New York: de Gruyter Mouton.

List of publications: Richard Johansson

Peer-reviewed original articles

Johansson, Richard. 2014. Automatic expansion of the Swedish FrameNet lexicon. *Constructions and Frames*, 6(1), pages 92–113.

Forsberg, Markus, **Richard Johansson**, Linnéa Bäckström, Lars Borin, Benjamin Lyngfelt, Joel Olofsson, Julia Prentice. 2014. From construction candidates to constructicon entries. An experiment using semi-automatic methods for identifying constructions in corpora. *Constructions and Frames*, 6(1), pages 114–135.

Johansson, Richard and Alessandro Moschitti. 2013. Relational features in fine-grained opinion analysis. *Computational Linguistics*, 39(3), pages 473–509.

Peer-reviewed conference contributions

Nieto-Piña, Luis and **Richard Johansson**. Training word sense embeddings with lexicon-based regularization. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 284–294. Taipei, Taiwan.

Ehrlemark, Anna, **Richard Johansson**, and Benjamin Lyngfelt. 2016. Retrieving occurrences of grammatical constructions. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 815–824. Osaka, Japan.

Johansson, Richard, Yvonne Adesam, Gerlof Bouma, and Karin Hedberg. 2016. A multi-domain corpus of Swedish word sense annotation. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 3019–3022. Portorož, Slovenia.

Nieto Piña, Luis and **Richard Johansson**. 2015. A simple and efficient method to generate word sense representations. In *Proceedings of Recent Advances in Natural Language Processing*, pages 465–472. Hissar, Bulgiaria.

Ghanimifard, Mehdi and **Richard Johansson**. 2015. Enriching word-sense embeddings with translational context. In *Proceedings of Recent Advances in Natural Language Processing*, pages 208–215. Hissar, Bulgiaria.

Johansson, Richard and Luis Nieto Piña. 2015. Combining relational and distributional knowledge for word sense disambiguation. In *Proceedings of the 20th Nordic Conference of Computational Linguistics*, pages 69–78. Vilnius, Lithuania.

Johansson, Richard and Luis Nieto Piña. 2015. Embedding a semantic network in a word space. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies*, pages 1428–1433. Denver, United States.

CV

CV - Nina Tahmasebi

Namn: Nina Tahmasebi
Födelsedatum: 19821029
Kön: Kvinna
Land: Sverige

Dr-examen: 2013-11-13
Akademisk titel: Doktor
Arbetsgivare: Göteborgs universitet

Doktorsexamen

Examen	Organisation	Avhandlingens titel (originalspråk)	Handledare
10208. Språkteknologi (språkvetenskaplig databehandling), 2013-11-13	Gottfried Wilhelm Leibniz Universität Hannover, Inst. för elektroteknik och datavetenskap		Wolfgang Nejdl

Utbildning

Forskarutbildning

Examen	Organisation
Doktorsexamen, 10208. Språkteknologi (språkvetenskaplig databehandling), 2013-11-13	Gottfried Wilhelm Leibniz Universität Hannover, Inst. för elektroteknik och datavetenskap

Utbildning på grund- och avancerad nivå

År	Examen
2008	10199. Annan matematik, Masters examen, Chalmers tekniska högskola
2007	10106. Sannolikhetsteori och statistik, Kandidatexamen, Göteborgs universitet

Arbetsliv

Anställningar

Period	Anställning	Del av forskning i anställningen (%)	Arbetsgivare
september 2014 - Nuvarande	Forskare, Tillsvidareanställning	100	Göteborgs universitet, Svenska språket, inst för

Postdoktorvistelser

Period	Organisation	Ämne
januari 2018 - december 2019	Göteborgs universitet, Litteratur, idéhistoria och religion, inst för	60201. Jämförande språkvetenskap och allmän lingvistik
september 2013 - augusti 2014	Chalmers tekniska högskola, 3724 - Computing Science	10299. Annan data- och informationsvetenskap

Forskarutbyten

Period	Typ	Organisation	Ämne
september 2009 - december 2009	Gästforskare	Max-Planck Institute for Informatics, Databases and Information Systems	10202. Systemvetenskap, informationssystem och informatik (samhällsvetenskaplig inriktning under 50804)

Uppehåll i forskningen

Period	Beskrivning
2016-09-05 - 2017-05-31	Föräldraledighet 20%
2015-10-01 - 2016-09-05	Föräldraledighet 50%
2015-05-18 - 2015-10-01	Föräldraledighet 100%

Meriter och utmärkelser

Handledda personer

År	Handledda personer	Roll
2022	Doktorand, Stian Eide, Göteborgs universitet	Bihandledare
2018	Student, Axel Almquist	Bihandledare
2017	Student, Peter Sumbler	Bihandledare
2016	Student, Antonio Ramis, Göteborgs universitet	Huvudhandledare
2016	Student, Bianka Nusko, Göteborgs universitet	Huvudhandledare
2014	Student, Karina Bunyik, Chalmers tekniska högskola	Huvudhandledare
2011	Student, Zhivko Asenov, L3S Research Center	Bihandledare
2010	Student, Kai Niklas, L3S Research Center	Bihandledare
2010	Student, Thomas Theuerkauf, L3S Research Center	Bihandledare

Bidrag erhållna i konkurrens

Period	Finansiär	Projektledare	Din roll	Totalt belopp (kr)
2016 - 2016	VR - Vetenskapsrådet	Nina Tahmasebi	Projektledare	77000

Priser och utmärkelser

År	Namn på priset/utmärkelsen	Utfärdare
2009	Best paper award	OnTheMove Federated Conferences & Workshops

Övriga meriter

Period	Typ av merit	Beskrivning
2016 - 2017	Chair lokal organisationskommitté	Chair för den lokala organisationskommittén för Nodalida 2017 (Nordic Conference on Computational Linguistics)
2016 - 2017	Programkommitté	I den nationella programkommittén för den andra konferensen i Digital Humanities in the Nordic Countries
2014 - 2016	Nationell samordnare	Nationell samordnare för Swe-Clarin, den svenska Clarin noden (Common Language Resources and Technology Infrastructure) https://sweclarin.se/
2013 - 2014	Lokal organisationskommitté	Projektledare för den lokala organisationen av EACL2014 (European Chapter of the Association for Computational Linguistics)

Namn: Richard Johansson
Födelsedatum: 19750709
Kön: Man
Land: Sverige

Dr-examen: 2008-12-05
Akademisk titel: Docent
Arbetsgivare: Göteborgs universitet

Utbildning

Forskarutbildning

Examen	Organisation	Avhandlingens titel (originalspråk)
Doktorsexamen, 10208. Språkteknologi (språkvetenskaplig databehandling), 2008-12-05	Lunds universitet, Datavetenskap 107121	Dependency-based Semantic Analysis of Natural-language Text

Utbildning på grund- och avancerad nivå

År	Examen
2003	10208. Språkteknologi (språkvetenskaplig databehandling), Civilingenjörsexamen/motsv, Lunds universitet

Arbetsliv

Anställningar

Period	Anställning	Del av forskning i anställningen (%)	Arbetsgivare
april 2016 - Nuvarande	Lektor, Tillsvidareanställning	50	Göteborgs universitet, Data- och informationsteknik, inst för
september 2015 - mars 2016 (Nuvarande)	Forskare, Tillsvidareanställning	50	Göteborgs universitet, Svenska språket, inst för
september 2013 - augusti 2015	Lektor	0	Göteborgs universitet, Svenska språket, inst för
september 2013 - augusti 2015	Forskare	50	Göteborgs universitet, Svenska språket, inst för
september 2011 - augusti 2013	Postdoktor	100	Göteborgs universitet, Svenska språket, inst för

Postdoktorvistelser

Period	Organisation	Ämne
januari 2009 - augusti 2011	Università degli Studi di Trento, Department of Information Engineering and Computer Science	10208. Språkteknologi (språkvetenskaplig databehandling)

Meriter och utmärkelser

Docentur

År	Ämne	Organisation
2015	602. Språk och litteratur	Göteborgs universitet, Svenska språket, inst för

Handledda personer

År	Handledda personer	Roll
2019	Doktorand, Luis Nieto Piña, Göteborgs universitet	Huvudhandledare
2018	Doktorand, Mikael Kågebäck, Chalmers tekniska högskola	Bihandledare
2018	Doktorand, Prasanth Kolachina, Göteborgs universitet	Bihandledare
2018	Doktorand, Olof Mogren, Chalmers tekniska högskola	Huvudhandledare

Bidrag erhållna i konkurrens

Period	Finansiär	Projektledare	Din roll	Delbelopp (kr)	Totalt belopp (kr)
2014 - 2016	RJ - Riksbankens Jubileumsfond	Yvonne Adesam	Medverkande	0	5605000
2014 - 2018	VR - Vetenskapsrådet	Richard Johansson	Projektledare	0	3760000

CV - Maria Koptjevskaja Tamm

Namn: Maria Koptjevskaja Tamm
Födelsedatum: 19570612
Kön: Kvinna
Land: Sverige
Dr-examen: 1988-04-16
Akademisk titel: Professor
Arbetsgivare: Stockholms universitet

Utbildning

Forskarutbildning

Examen	Organisation	Avhandlingens titel (originalspråk)
Doktorsexamen, 60201. Jämförande språkvetenskap och allmän lingvistik, 1988-04-16	Stockholms universitet, Institutionen för lingvistik	A typology of action nominal constructions

Utbildning på grund- och avancerad nivå

År	Examen
1979	60201. Jämförande språkvetenskap och allmän lingvistik, Högskoleexamen, MGU

Arbetsliv

Anställningar

Period	Anställning	Del av forskning i anställningen (%)	Arbetsgivare	Övrig information
juli 2001 - Nuvarande	Professor, Tillsvidareanställning	50	Stockholms universitet	Har varit anställd vid Stockholms universitet sedan 1988. Blev befordrad till professor i 2001

Forskarutbyten

Period	Typ	Organisation	Ämne
oktober 2015 - oktober 2015	Gästprofessor	University of Ghana-Legon, Department of Linguistics	60201. Jämförande språkvetenskap och allmän lingvistik
mars 2003 - april 2003	Gästforskare	Max-Planck Institute for Evolutionary Anthropology, Linguistics / Typology	60201. Jämförande språkvetenskap och allmän lingvistik
december 1999 - december 1999	Gästprofessor	L'Università di Pavia, Department of linguistics	60201. Jämförande språkvetenskap och allmän lingvistik

Meriter och utmärkelser

Docentur

År	Ämne	Organisation
1993	602. Språk och litteratur	Stockholms universitet, Institutionen för lingvistik

Handledda personer

Handledda personer	Roll	Antal
Doktorand	Huvudhandledare	8
Doktorand	Bihandledare	2

Bidrag erhållna i konkurrens

Period	Finansiär	Projektledare	Din roll	Totalt belopp (kr)
2018 - 2018	RJ - Riksbankens Jubileumsfond	Maria Koptjevskaja Tamm	Projektledare	0
2014 - 2016	Sverige - Övriga statliga myndigheter	Maria Koptjevskaja Tamm	Projektledare	0
2012 - 2012	VR - Vetenskapsrådet	Maria Koptjevskaja Tamm	Projektledare	0
2009 - 2012	VR - Vetenskapsrådet	Maria Koptjevskaja Tamm	Projektledare	0
2006 - 2009	Europeiska Unionen (EU)	Maria Koptjevskaja Tamm	Projektledare	0
2001 - 2004	VR - Vetenskapsrådet	Maria Koptjevskaja Tamm	Projektledare	0
1991 - 1998	Sverige - Universitet och högskolor	Maria Koptjevskaja Tamm	Projektledare	0

Priser och utmärkelser

År	Namn på priset/utmärkelsen	Utfärdare
2018	Filosofisk-filologoska klassens Rettigska pris	Kung. Vitterhetsakademien
2010	ledamot i Academia Europaea	NB: "land" är Europa, men detta omfattas inte av valmöjligheterna Academia Europaea

Övriga meriter

Period	Typ av merit	Beskrivning
2018 - 2024	Huvudredaktör för tidskrifter "Linguistic Typology"	"Linguistic Typology" är organ för Association for Linguistic Typology och den viktigaste internationella tidskriften inom typologiområdet.
2016 - 2022	Medlem av "Board of Consulting editors" för "Linguistics"	"Linguistics" är en av de internationellt ledande lingvistiska tidskrifterna
2013 - 2020	medlem av Sektionskommitté för lingvistik inom Academia Europaea	Medlemmarna i Sektionskommittén nominerar och röstar för nya ledamöter till Akademien
2008 - 2012	Member of the International Review Panel for the EUROCORES (European Collaborative Research) Programme "EuroBABEL: Better Analyses Based on Endangered Languages", ESF	

CV - Susanne vejdemo

Namn: Susanne vejdemo
Födelsedatum: 19821113
Kön: Kvinna
Land: Sverige

Dr-examen: 2017-03-03
Akademisk titel: Doktor
Arbetsgivare: City University of New York: College of Staten Island

Utbildning

Forskarutbildning

Examen	Organisation	Avhandlingens titel (originalspråk)
Doktorsexamen, 60201. Jämförande språkvetenskap och allmän lingvistik, 2017-03-03	Stockholms universitet, Institutionen för lingvistik	Triangulating Perspectives on Lexical Replacement: From Predictive Statistical Models to Descriptive Color Linguistics

Arbetsliv

Anställningar

Period	Anställning	Del av forskning i anställningen (%)	Arbetsgivare
augusti 2017 - mars 2018 (Nuvarande)	Lektor, Projektanställning	0	City University of New York: College of Staten Island, English and World Languages
augusti 2017 - december 2017	Lektor, Projektanställning	0	City University of New York: Queens College, Language and Communicative Disorders
augusti 2011 - januari 2017	Doktorand, Tillsvidareanställning	100	Stockholms universitet, Institutionen för lingvistik

Publikationer

Publikationer - Nina Tahmasebi

Namn: Nina Tahmasebi	**Dr-examen:** 2013-11-13
Födelsedatum: 19821029	**Akademisk titel:** Doktor
Kön: Kvinna	**Arbetsgivare:** Göteborgs universitet
Land: Sverige	

Publikationer är avstängt för Tahmasebi, Nina på den här ansökan.

Publikationer - Richard Johansson

Namn: Richard Johansson	**Dr-examen:** 2008-12-05
Födelsedatum: 19750709	**Akademisk titel:** Docent
Kön: Man	**Arbetsgivare:** Göteborgs universitet
Land: Sverige	

Publikationer är avstängt för Johansson, Richard på den här ansökan.

Publikationer - Maria Koptjevskaja Tamm

Namn: Maria Koptjevskaja Tamm	**Dr-examen:** 1988-04-16
Födelsedatum: 19570612	**Akademisk titel:** Professor
Kön: Kvinna	**Arbetsgivare:** Stockholms universitet
Land: Sverige	

Publikationer är avstängt för Koptjevskaja Tamm, Maria på den här ansökan.

Publikationer - Susanne vejdemo

Namn: Susanne vejdemo	**Dr-examen:** 2017-03-03
Födelsedatum: 19821113	**Akademisk titel:** Doktor
Kön: Kvinna	**Arbetsgivare:** City University of New York: College of Staten Island
Land: Sverige	

Publikationer är avstängt för vejdemo, Susanne på den här ansökan.

Registrera

Villkor

Ansökan ska förutom av den sökande även signeras av behörig företrädare för medelsförvaltaren. Företrädaren är vanligtvis prefekten vid den institution där forskningen ska bedrivas men beror på medelsförvaltares organisationsstruktur.

Signering av den *sökande* innebär en bekräftelse av att:

- uppgifterna i ansökan är korrekta och följer Vetenskapsrådets instruktioner
- eventuella bisysslor och kommersiella bindningar har redovisats för medelsförvaltaren och att det där inte framkommit något som strider mot god forskningssed
- sökande inte fällts för oredlighet i forskning under de två senaste åren räknat från sista ansökningsdag
- de tillstånd och godkännanden som krävs finns innan forskningen påbörjas, exempelvis tillstånd från Läkemedelsverket eller godkännande från etikprövningsnämnd respektive djurförsöksetisk nämnd
- sökande kommer att följa samtliga övriga villkor som gäller för bidraget.

Signering av *medelsförvaltaren* innebär en bekräftelse av att:

- den beskrivna forskningen eller forskningsstödjande verksamheten kan beredas plats vid medelsförvaltaren under den tid och i den omfattning som anges i ansökan
- den sökande kommer vara anställd vid medelsförvaltaren under den tid som ansökan avser
- medelsförvaltaren godkänner kostnadsberäkningen i ansökan
- sökande inte fällts för oredlighet i forskning av signerande medelsförvaltare under de senaste två åren innan sista ansökningsdag
- medelsförvaltaren kommer att följa samtliga övriga villkor som gäller för bidraget.

Ovanstående punkter ska ha diskuterats mellan parterna innan företrädaren för medelsförvaltaren godkänner och signerar ansökan.