博士論文（要約）

Multilingual computational lexicography:
frame semantics meets distributional semantics

（多言語計算辞書学: フレーム意味論と分布意味論の接点）

Anna Rogers
ロージャズ　アンナ

The University of Tokyo

Multilingual computational lexicography:

frame semantics meets distributional semantics

（多言語計算辞書学: フレーム意味論と分布意味論の接点）

Author:
Anna Rogers

Supervisor:
Dr. Toshio Ohori

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

Graduate School of Arts and Sciences
Department of Language and Information Sciences

Dissertation committee:
Chair: Toshio Ohori (University of Tokyo, professor)
Kyoko Hirose Ohara (Keio University, professor)
Kaoru Koda (University of Tokyo, former professor)
Tsuneaki Kato (University of Tokyo, professor)
Brendan Wilson (University of Tokyo, professor)

April 2017

Abstract

This thesis discusses several issues in multilingual frame-based computational lexicography from the combined perspectives of frame semantics and distributional semantics. The case study under consideration is three posture verbs (*sit, lie, stand*) in English, and their counterparts in Russian and Japanese.

The first issue in frame-based accounts of these verbs concerns representation of aspectual classes and inchoativity/causativity. I develop an alternative model that can be implemented in the current Berkeley FrameNet architecture, but that is less subject to inconsistencies and is better equipped to deal with multilingual data. It is also more inference-friendly and thus would be more useful from the point of view of applications of FrameNets in Natural Language Processing.

Since implementing this model would require too large a number of frame-to-frame relations, I explore the possibilities of identifying verb classes with supervised machine learning over simple bag-of-words vector space models. I successfully recover Russian imperfective/perfective verbs, including some that are not registered in current dictionaries. I also demonstrate the possibility of identifying pairs of Japanese intransitive/transitive verbs with word analogies.

I further consider the problem of representing selectional preferences in a frame-based lexical resource, and I argue against the current Berkeley FrameNet model in which semantic types of frame elements are represented as independent ontological categories such as "sentients" or "physical objects". On the basis of data on individual variation in acceptability judgements on subjects of posture verbs, I propose a model based on extensions from the anthropocentric prototype. I further outline what semantic constraints define extensions of this prototype.

I consider an alternative approach to modeling selectional preferences, namely supervised classification over word vectors. In a series of experiments I show that a more linguistically informed training dataset can yield up to 50% increase in accuracy of such classification over the commonly used approach with random and unbalanced samples.

I conclude with a discussion of several theoretical issues in both frame semantics and distributional semantics, and potential venues for their collaboration. I also address the question of universality of frames. I argue that frames defined on the basis of one language are not a valid semantic interlingua, and may not even hold for all the speakers of the same language. Still, it is theoretically possible to construct an interlingual frame database if we take a more typological approach from the outset, integrating data from as many languages as possible.

Acknowledgements

I would like to express sincere gratitude to my supervisor and all committee members from whom I learned so much. It was my privilege to study linguistic analysis with Toshio Ohori and Kyoko Ohara, programming – with Kaoru Koda and Tsuneaki Kato, and philosophy – with Brendan Wilson.

This thesis draws heavily on Berkeley FrameNet, and it was the Japanese FrameNet that initially brought me to Japan. Without either of these FrameNets, my work would not have been possible.

I am deeply indebted to my co-author, Dr. Aleksandr Drozd of Tokyo Institute of Technology. Our joint research[1] was mutually insightful, and I am hoping that more linguists and computer scientists will be able to create such partnerships.

Many thanks are also due to many researchers to whom I was lucky to talk at conferences and other occasions, including William Croft, Colin Baker, Ellen Dodge, Alessandro Lenci, Yoav Goldberg, Marc Vilain, Stefan Evert, Anna Rumshisky, Ekaterina Vylomova, Omer Levy, Simon De Deyne, and many others. Among fellow students at the University of Tokyo, I got many insights from Kei Sakaguchi, Shiori Ikawa, Midori Wada, and Marzena Karpinska.

The material presented in chapter 6 of this thesis was collected from dozens of my friends all around the globe. Special mention for patience, imagination, and wonderful music goes to Misako Matsumoto, Eiichi Onda, Jeff Holtzkener, Robert Wallace, Katie Stevens, Andrew Shive, Olga Pokrovska, Boris Itzkovich, Aleksey Rozov, and Mark Kovnatsky. Among my teaching colleagues, I am particularly grateful to Junko Hirota, Akiko Yamanaka, John Porteous, Steven Green, Jennifer Nichole, Miho Takano, and Keiko Abe-Ford.

Finally, this work would not have been possible at all without support from my family, especially my husband and my mother.

[1] All due acknowledgements for particular projects are provided in footnotes and references in the body of the thesis.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	**A**rtificial **I**ntelligence
BCCWJ	**B**alanced **C**orpus of **C**ontemporary **W**ritten **J**apanese
BFN	**B**erkeley **F**rame**N**et
BNC	**B**ritish **N**ational **C**orpus
CG	**C**onstruction **G**rammar
DS	**D**istributional **S**emantics
FBCL	**F**rame-**B**ased **C**omputational **L**exicography
FE	**F**rame **E**lement
FI	**F**rame **I**nstance
FS	**F**rame **S**emantics
LU	**L**exical **U**nit
MWU	**M**ulti-**W**ord **U**nit
NLP	**N**atural **L**anguage **P**rocessing
RNC	**R**ussian **N**ational **C**orpus
SPFC	**S**elf-**P**ublished **F**iction **C**orpus
SVD	**S**ingular **V**alue **D**ecomposition
VSM	**V**ector **S**pace **M**odel

Chapter 1

Introduction

1.1 Computational lexicography: past and present

Computational lexicography is one of the most actively developed fields of linguistics. Due to the fast development of technology and the pace of news in an increasingly global society, a lot more speakers are interacting than ever before. This means that language changes fast: new terms, memes, and slogans are born every second and travel at lightning speed. For the first time in history linguists have the means to observe the speech of so many people so soon after its production.

All this new data comes in unprecedented amounts and new formats, including both written and spoken language. The digital era even removes the constraints imposed by the linear structure and limited space of paper dictionaries. However, the lexicographic toolbox has to adapt to the challenge.

In addition to the pressure of enormous volumes of new data, there is increasing peer pressure from other branches of knowledge. Linguistics, philosophy, psychology, and computer science are all working to produce more accurate models of meaning, potentially changing the very basis of lexicography.

Among numerous competing semantic theories that could make important contributions, this study focuses on two that come from completely different directions: frame semantics (FS) and distributional semantics (DS).

FS is so far the only theory in cognitive linguistics that has sparked off a large lexicographic project – the Berkeley FrameNet[1] (BFN) (C. F. Baker, Fillmore, & Lowe, 1998) and its numerous branches in other languages. The basic idea of FS is that the meanings of language units arise from the knowledge of the situations in which they are used, and can only be defined with respect to such situations. Frame, or "scene", is used to refer to both background knowledge not evoked by linguistic units (cognitive frames) and background knowledge evoked by linguistic units (linguistic frames). For example, the words "blackboard" and "teacher" are parts of the same linguistic frame (`School`) because they name the parts of the same block of experience. BFN is a resource for linguistic frames.

[1]The BFN data presented in this dissertation comes from BFN 1.6.

On the other hand, DS views meanings as something emerging from word distributions. Even without any syntactic or semantic annotation of the source corpora this approach goes surprisingly far, capturing many relations which frame semanticists would consider their province.For example, Mikolov, Yih, and Zweig (2013) have shown that word embeddings capture many "encyclopedic" relations between pairs of words, such as countries and capitals: given *France* and *Paris*, it is possible to use vector arithmetic to find that the capital of *Japan* is *Tokyo*.

Unlike FS, DS derives its representation automatically, and is suitable for dealing with the flood of new linguistic data. Therefore it would be practically useful for FS to incorporate findings of DS. But this project also explores other possibilities of interaction between these two approaches to semantics that could help them both to move forward.

1.2 Objectives of the study

This study is motivated by the need to bring closer the fields of frame-based computational lexicography (as represented by the FrameNets community) and Natural Language Processing (NLP). The latter is typically performed by computer scientists with limited linguistic skills; the resources they create automatically or semi-automatically tend to be large-scale and produced quickly, but often lacking in quality. The former is typically performed by professional linguists with limited programming skills; they create lexicographic resources through a lot of manual annotation and qualitative analysis, which bounds such projects to be relatively small-scale and slow in production.

Berkeley FrameNet (BFN) is almost the only corpus project in cognitive linguistics, but it still does not compare even with traditional dictionaries because such coverage is not their goal. This prevents BFN from being as useful as it could be for various teaching and NLP applications.

The quality/quantity balance in the creation of language resources is a general problem pertaining not only to frame-based computational lexicography (FBCL). However, in this study I show that in case of FBCL, the "linguistic" workflow also masks several deeper methodological issues, stemming from the fact that FS, like any theory of semantics, has not yet come up with a unified account of language. FBCL currently focuses on the phenomena that it can best account for, using a relatively small dataset in combination with a cherry-picking approach to selecting data, and leaving out a wide range of linguistic phenomena (especially highly schematic linguistic units such as prepositions). This results in inconsistencies in practical analysis and a lack of systematicity in defining top-level frames (Osswald & Van Valin Jr, 2014).

In scope of this study I focus on three questions that are not unproblematic for the current FrameNets:

- establishing correspondences across languages,

- consistent representation of such features of verbal semantics as aspect and inchoative-causative alternation at the level of the "abstract" and "lexical" frames,
- providing a frame-based account for selectional constraints on the frame elements (FEs).

As a test case for cross-lingual FBCL I investigate posture verbs in English, Russian and Japanese, showing that a satisfactory account of the above issues would require a considerable revision of the current BFN architecture. I further show that DS methodology can help to supplement them, and in a way that would be beneficial to both approaches in both practical and theoretical perspective.

1.3 Approaches to multilingual frame-based lexicography

Before we can start, it is important to define the scope of this study. "Computational lexicography" in this work is understood in a very broad sense of creating resources that explicitly encode meanings of linguistic units and/or relations between them. Furthermore, following the construction grammar vein, I do not limit contemporary computational lexicography to lexical units (LUs), and consider resources such as Constructicons also lexicographic.

"Multilingual frame-based lexicography" at the moment is more of a research program than an ongoing project. At the moment, the only multilingual FrameNet in existence is the Kicktionary, a small FrameNet for the soccer domain (Schmidt, 2009). A bigger multilingual FrameNet is being planned in Berkeley[2].

However, "multilinguality" can also mean simply "multilinguality of language data refers to the existence of such resources for more than one language" (Lönneker-Rodman, 2007, p.3). In this sense FBCL is already multilingual, since there already are French, Spanish, Japanese and other FrameNets. Most of them follow the so-called "expand" approach, in which "a resource for one language, which is regarded as stable at that time, is transferred to another language" (Lönneker-Rodman, 2007, p.6). The role of the "stable" resource to be mapped to other languages usually falls to the English BFN. One hypothesis is that BFN frames can serve as semantic interlingua and, once all the resources are more or less complete, they could eventually be interlinked, with special transfer rules for cases of partial matches (Boas, 2009).

For a number of reasons, "expand" approach is more practical than the "merge" approach, in which resources for different languages are developed independently and then linked. However, starting from the English frame database creates a bias towards the English conceptualization scheme. In this dissertation I argue for an alternative to "expanding" and "merging": designing a frame database on the basis of data from

[2]Miriam Petruck, MetaNet tutorial at the conference of Association for Computational Linguistic (ACL 2016), `http://acl2016.org/index.php?article_id=61`

all target languages simultaneously, as it was done in Kicktionary. While this approach is more laborious at the initial stage, it is the only way to make sure that all necessary distinctions are taken into account, and it can be helped by the growing body of cross-linguistic data. In the scope of this dissertation I demonstrate how integrating data from English, Russian and Japanese can provide a more accurate frame-based account of posture verbs in these languages.

This study does not offer a complete methodology for developing a multilingual frame-based resource, but it shows that taking into account multilingual data significantly improves consistency of the database and highlights subtle distinctions that would be easy to miss in a monolingual resource. Furthermore, it demonstrates the possibility of leveraging DS tools for at least parts of the task that would otherwise be gargantuan.

1.4 Posture verbs as a test case for multilingual semantic resource

Posture verbs are a lexical group that has been shown to vary greatly across languages both grammatically and semantically (Ameka & Levinson, 2007; Newman, 2002b). The dimensions of variation discussed in this study include their aspectual properties, their patterns of lexicalization across languages, and their selectional constraints.

This study uses posture verb data from three languages of different families: English, Russian and Japanese. The "core" posture verbs in English are *sit*, *lie* and *stand*; their equivalents in Russian are *sidet'*, *lezhat'* and *stoyat'*, and in Japanese - *suwaru*, *tatsu* and *yoko-ni naru*. I also consider some dialectal variants of these verbs and their derivatives.

The main focus is on two dimensions of semantic variation in posture verbs: aspect and possibility of using them for indicating location of objects. Posture verbs also have numerous metaphorical extensions, such as *lie* as "rest", *stand* as "tolerate", *sit* as "conduct a meeting", etc.; but these extensions are not the focus of this study.

The first challenge concerns the fact that posture verbs, like other verbs, may correspond between languages in terms of describing the same situation, but they may vary with respect to the temporal "profile" of the posture events. For instance, in English the verb *sit* is aspectually polysemous, as it has both stative and dynamic meaning. Consider (1.1):

(1.1) I sat on the chair[3].

(1.1) could mean either "I sat myself down on the chair" or "I was sitting on a chair (and didn't move)" (Newman, 2002b, p. 4). In other languages the posture verbs may

[3]The examples cited in this study come from bilingual dictionaries, other studies, and corpora, including BFN (C. F. Baker et al., 1998), BNC ("The British National Corpus, Version 3 (BNC XML Edition)," 2007), RNC (the Russian National Corpus, (Apresjan et al., 2006)), and BCCWJ (the Balanced Corpus of Written Japanese, (Maekawa, 2008)). Where no reference is given, the examples are provided by the author.

have other aspectual properties, lexicalized or expressed with constructions. Accounting for such variation cross-linguistically is part of the general problem with providing a unified account of lexical and syntactic constructions, and integrating "verbal" frames with the general representation of event structure in cross-linguistically valid and consistent manner.

The second challenge is that languages differ with respect to whether verbs that indicate human posture can also be used for indicating location of objects (Viberg, 2013). When they do, they retain their original spatial characteristics to various extent, and they also vary with respect to which objects they are compatible with. For example, in English the verb *sit* is often used to indicate location of an object, irrespective of its spatial orientation (1.2), but *lie* or *stand* are more tied to the spatial context.

(1.2) The cup sat on the desk.

In Russian, on the other hand, it is possible to combine a posture verb with "cup", but it would have to "stand" rather than "sit". As for Japanese, it normally does not allow for any such combinations. In both cases, the restrictions are motivated by certain properties of the real-world entities that are denoted by the arguments of posture verbs, and, as such, should be accounted for by FS.

Finally, one more challenge posed by the posture verbs concerns the very nature of multilingual FBCL. The primary tenet of FS is that meanings are relativized to scenes; if scenes come from experience, they can be expected to vary from society to society, and even from person to person. Even something as basic as sitting is different for Japanese culture, where sitting on the floor is typical, and American culture, where it is less so. The actual sitting posture also differs in these two situations, ranging from extremely stiff and formal *seiza* pose to comfortable sinking in a huge sofa. Does that mean that English and Japanese have different *sitting* frames? If so, how do we establish the correspondence, and what does this mean for FBCL? In scope of this work I will consider this issue only briefly, but it needs to be kept in mind in the general perspective of multilingual FBCL.

1.5 Dissertation outline

This work generally takes a problem-driven approach. I start with the BFN perspective on the above issues, and identify where the current proposals can be improved. I further look for solutions in the collaboration between FS and DS. The contribution of this dissertation lies not only in the particular proposals I develop, but also in this workflow which I show to be mutually beneficial to both approaches to semantics, both practically and theoretically.

The overall structure of the dissertation is as follows.

Chapter 2 provides a historical overview of how FS was developed. I review some of the original ideas that Charles Fillmore brought forward in the 1970s, and their

current FrameNet incarnation. I also survey the current NLP applications that use FrameNets, and provide a brief introduction to distributional semantics.

Chapter 3 presents a practical test of the "expand" approach in the current FrameNets. The frames developed by BFN team for English posture verbs are projected onto Russian and Japanese data. This attempt suggests the necessity for a more consistent treatment of aspectual classes and the overarching event structure scenario, of which I present an alternative model. I also propose a basic mechanism for establishing cross-lingual correspondences in cases when in one language a frame is evoked lexically, and in another - by a complex expression.

Chapter 4 describes how distributional semantic models can be used to automatically induce verb classes relevant to event structure representation (which would help to ensure consistency of frame-to-frame relations in a multilingual FrameNet). I conduct two experiments on perfective/imperfective verbs in Russian and transitive/intransitive verbs in Japanese, with two different methods of retrieving verb classes both of which achieve over 90% accuracy on my test data. I show that such work would be mutually beneficial for both BFN and DS: the former can solve its practical tasks, and the latter gets to experiment with a wide range of linguistic relations, some of which correspond to unusual mathematical properties of the vector space.

Chapter 5 examines the current account of selectional preferences in BFN, showing that ontological semantic types cannot accommodate posture verb data (particularly not in cross-lingual perspective). I argue in favor of a prototype-based account, on the basis of the evidence of cross-lingual and individual variation in acceptability judgements for subjects of posture verbs. For this study I conducted a survey that was completed by 72 native speakers of English, Russian and Japanese (24 per language).

Chapter 6 explores the possibility of modeling selectional preferences of English posture verbs with supervised classification over word vectors. In this approach, the objective is to train a classifier in such a way that the probabilities it assigns to different potential arguments of posture verbs approximate human acceptability judgements. I show that supervised machine learning opens new avenues for the linguists, who can now experimentally study generalizations over word vectors produced by various training sets. In scope of this project I explored the effect of balanced subsampling of positive and negative training examples, and also the effect of introducing counter-examples. While the resulting models are limited in several ways, with a good training set it is possible to distinguish between, e.g., things that can and cannot *stand* with 80-90% accuracy.

Chapter 7 puts the findings of this study in the context of the problem of developing a unified semantic theory. I discuss the internal limitations of both FS and DS, and what could be expected of their collaboration. I also consider the problem of universality of frames, and the challenges of multilingual FBCL.

Chapter 8 concludes this study with discussion of its theoretical implications for FS and DS, its limitations, and suggestions for future work.

1.6 Summary

Of all the different theories in cognitive semantics, FS is the only theory applied to a large-scale computational lexicographic project: the family of FrameNets for various languages. Frame-based descriptions of vocabulary attempt to link lexical knowledge to world knowledge, rather than simply establish paradigmatic links between words such as synonymy or antonymy. This allows the linguist to avoid being trapped in words, and offers a very attractive platform for semantic analysis.

However, with the current FBCL methodology the FrameNets are making slow progress, and there are also a few methodological challenges. This project focuses on three such challenges: frame-based establishment of interlingual correspondences, consistency in linking lexical and abstract frames, and modeling of selectional constraints. I conduct a detailed case study of the above issues with the posture verb data from three languages from different families: English, Russian and Japanese. This lexical group has been shown to exhibit considerable cross-linguistic variation in terms of their selectional constraints, and their aspectual classes also offer a good test case for linking lexical frames with more abstract event structure templates.

I suggest that both of these issues can be tackled more efficiently if FS makes use of the methodology offered by DS, and I develop specific proposals about how this can be achieved. But the collaboration between FS and DS is by no means limited to the areas I investigate, and I hope that this work will show to both communities how fruitful such collaboration can be, in both theoretical and practical terms.

In order to reuse the data that has been already created by numerous FrameNets and make my contributions easy to incorporate, I aim to maintain as much compatibility with the current BFN architecture as possible. However, in chapters 5 and 6 I discuss alternative approaches to selectional preferences that are not possible to implement in the current FrameNets. In general, this study is not a part of the current FrameNets, but rather an alternative view of the future of both FS and DS. The current FrameNets are only one implementation of FS, and, as this theory continues to be developed in the work of many non-FN linguists, alternative implementations are bound to follow.

Chapter 2

Frames approach to semantic analysis

I will start with a brief review of the origin of FS in the early works of Fillmore, and then look at its current state in the FrameNet family of lexicographic projects and construction grammar. Sections 2.2 - 2.3 will look at early AI frame-based representations and modern NLP applications that make use of FS resources. Section 2.4 will provide an (extremely) brief overview of DS and distributional meaning models.

2.1 Frame semantics in cognitive linguistics

2.1.1 "Semantics of understanding"

Fillmore's FS can be seen as a development of his case grammar theory, both having the notion of "valence" as core. However, his 1970s works also show that it was rooted in dissatisfaction with feature-based or truth-conditional approaches (Fillmore, 1975). He was aiming at a "richer" semantics, and a semantics geared towards understanding rather than generating "valid" sentences.

Fillmore's basic idea was that understanding an utterance is only possible on condition of possessing the cognitive schema, parts of which the speaker has profiled by the words he used. At first he used the term "schema" to refer to "any coherent individuatable perception, memory, experience, action, or object" (Fillmore, 1977, p. 84). He considered frames as a linguistic rather than cognitive phenomenon: they are the "specific lexico-grammatical provisions in a given language for naming and describing the categories and relations found in schemata." (Fillmore, 1977, p. 127)

From the perspective of NLP, another important feature of this early version of FS was that it focused on the process of *understanding* language, including all the implicit information that goes beyond what is stated explicitly. Fillmore distinguished between "evoked" and "invoked" frames, describing it in the following way:

> A frame is invoked when the interpreter, in trying to make sense of a text segment, is able to assign it an interpretation by situating its content in a pattern that is known independently of the text. A frame is evoked by the text if some linguistic form or pattern is conventionally associated with the

> frame in question. For example, the sentence "We never open our presents until the morning" makes no mention of Christmas, yet interpreters who share certain cultural experiences, would immediately (in the terminology suggested here) invoke a Christmas context; replace the simple noun *presents* with *Christmas presents* and we have introduced a word which evokes that same context (Fillmore, 1985, p. 232).

This was a major step forward from semantic theories that were only dealing with the linguistic units explicitly present in a given context, because it put semantics in the "general knowledge" pool rather than some specialized linguistic ability. It is this feature of FS that accounts for much of its current success in AI and NLP: it offers access to commonsense reasoning.

Consider (2.3), another Fillmore's example:

(2.3) My dad wasted most of the morning on the bus.
(Fillmore, 1985, p. 230-231)

We understand a lot more from this sentence than what it actually says - that the addressee is probably not a member of speaker's family, that the speaker has a good relationship with his father, that something was probably wrong with the bus, that "morning" here denotes "working hours" rather than "dawn-till-noon" sense. Fillmore suggested that understanding a text consists of "giving it a maximally rich interpretation, an interpretation which draws everything out of the text that it can". One can argue that in actual communication the addressees do not always draw *all* possible information (e.g. because they are tired, uninterested, distracted, or uncooperative), but this approach is still very fruitful from the practical perspective of NLP - and it maintains a lot of cognitive plausibility.

This goal of maximally rich interpretation led Fillmore to abandon the distinction between "encyclopedic" and "linguistic" knowledge (Fillmore, 1985, p.233). It is the extra-linguistic nature of frames, their being "grounded" in reality that accounts for much of the appeal of FS.

2.1.2 The onset of frame-based lexicography

In subsequent Fillmore's work in the early 1990s there is a lot of discussion of corpora and new computational lexicons. Fillmore argued for the use of corpora, and warned about necessity of careful qualitative analysis of this data, calling himself "an armchair linguist who refuses to give up his old ways but who finds profit in being a consumer of some of the resources that corpus linguists have created" (Fillmore, 1992, p.35).

Fillmore's work on `Risk` with B.T Atkins (a lexicographic adviser at Oxford University press at that time) demonstrated the possibility of using a large corpus database to gather data about frame elements (FEs) and their syntactic realizations, and the benefit of this approach to lexicography: frame-based analysis would enable lexicographers to distinguish subtle differences that are best described "not necessarily in

terms of lexical semantic differences as such, but as differences in the manner of syntactic realization of the elements of their common frame" (Fillmore & Atkins, 1992, p.101).

Further examination of `Risk` in dictionaries and corpora databases led Fillmore to conclude that "the classical printed dictionary format is too restricted, in both length and dimension, to present an intelligible and truthful statement of the way the word `risk` is used" (Fillmore & Atkins, 1994, p.350). They showed that frame-based analysis required that the user of the new dictionary is not confined to a linear list of senses, and that corpus would need to be a part of the dictionary.

While the earlier papers discuss how frame-based analysis may be useful for lexicographic work, the `Risk` papers presented FS as an independent approach to building dictionaries of a new kind. Fillmore and Atkins envisaged that "an on-line frame semantics dictionary would be more than simply a multi-accessed print dictionary, though at the minimum it would certainly be that. The rich and flexible description underlying the lexical database, including the tagged and parsed corpus integral to it, would give the user a new kind of resource, one which would not only provide detailed and comprehensible answers to questions regarding word usage, synonymy, and antonymy, but would also be equipped to suggest various ways of expressing a complex concept: a true 'active' or 'encoding' dictionary." (Fillmore & Atkins, 1994, p.376).

2.1.3 Berkeley FrameNet

The development of Berkeley FrameNet (BFN) started almost twenty years ago (C. F. Baker et al., 1998). Its initial goal was "efficiently capturing human insights into semantic structure" in order to produce a frame-semantic lexicon for English. It aimed to cover general vocabulary in 13 semantic domains (including "Time", "Body", "Emotion", and "Transaction"). The project included three modules: the lexicon (descriptions of lexemes and their syntactic patterns), the frame database, and the example sentences, manually picked from the British National Corpus and annotated.

The second stage of the project introduced frame-to-frame relations and semantic types, as well as richer grammatical annotation which aims to be theory-neutral (Fillmore, 2007, p.158). An elaborate system was developed to deal with different kinds of mismatches between syntax and semantics, such as syntactic elements without an assigned semantic role, non-expressed semantic elements, semantic elements matching with several syntactic elements, support verbs, etc. (Fillmore, 2007).

As of 03.08.2016, BFN 1.5 contains 13542 LUs in 1223 frames, with about 202236 annotated sets (including full-text annotation). Release 1.6 is being prepared. Technically BFN is an SQL database, with data releases available in XML format. It is also integrated into some third-party NLP systems such as Natural Language Processing Toolkit[1].

[1] `http:www.nltk.org/howto/framenet.html`

Figure 2.1 presents the much-quoted `Commercial_transaction` scenario as an example of a frame entry in BFN. The position of this scenario in the overall frame network can be viewed in FrameGrapher, a BFN tool for visualizing frame-to-frame relations (Figure 2.2).

As shown in Figure 2.1, for each frame BFN provides an informal definition and a set of frame elements (FEs), divided into "core", "peripheral" and "extra-thematic". "Core" and "peripheral" FEs are frame-specific, even if their names happen to be shared between frames, while "extra-thematic" FEs are not frame-specific. Each FE is also provided with an informal definition and sometimes with example sentences. All of this information can be viewed on the BFN website in "frame reports".

Frame entries also contain the lists of frames that are related to the target frame. In case of `Commercial_transaction` only three such relations are present, but they place this frame in the wider frame network, as shown in Figure 2.2. There are currently 8 frame-to-frame relations: generalization relations (*Inherits*, *Using*, Perspective on), event structure relations (*Precedes* and *Subframe*), and systematic relations (*Inchoative_ of*, *Causative_ of*) (Fillmore & Baker, 2010, p. 329-330), plus a *See_ also* relation. Additionally, a new BFN relation for marking source and target frames in cognitive metaphors has just been introduced to enable markup of metaphors independently from the MetaNet project[2]. All of these relations are directed or assymmetrical, i.e. one of the frames in the relation is a "head" of the other. *Inheritance* relation is the chief mechanism for conceptual content to be transferred from frame to frame, and it entails that the child frames must also inherit the FEs of the parent frame.

BFN represents polysemy by attributing different lexical units (LUs) to different frames. LUs may consist of single words of multi-word units (MWUs). Figure 2.3 presents an example of a lexical entry for *transaction*, a noun evoking the above `Commercial_transaction` frame (it is currently the only LU in this frame). For each entry there is an informal definition. BFN aims to provide all lexical entries with example sentences that have annotations of both syntactic and semantic elements. The idea is that once the database is big enough, the combined annotations of FEs and their syntactic realizations will provide insights into the interface of syntax and semantics. For each lexical entry there are also tables of valence patterns. However, since BFN is work-in-progress, many such annotations have been initiated, but not actually completed. In Figure 2.3, only 2 out of 5 example sentences have annotations of the syntactic realizations of FEs (and only one such per sentence).

Some FEs are associated with semantic types, which are used "to record information that is not representable in our frame and FE hierarchies". They indicate "the basic typing of fillers of FE", such as "Sentient" for the Cognizer FE. They are propagated to the FEs of child frame, irrespective of the type of frame-to-frame relation, but are said to be independent of the frame network *per se*, "since FEs which are arbitrarily far away according to the frame hierarchy, such as the Experiencer of `Perception_body` and the Perpetrator of the Piracy frame, are often marked as the same semantic type (in this case, Sentient). BFN currently uses very broad ontological classes to denote semantic types, and it is said to be "desirable" to also

[2]Miriam Petruck, MetaNet tutorial at the conference of Association for Computational Linguistic (ACL 2016), `http://acl2016.org/index.php?article_id=61`

`Commercial_transaction`
Definition: These are words that describe basic commercial transactions involving a BUYER and a SELLER who exchange MONEY and GOODS. The individual words vary in the frame element realization patterns. For example, the typical patterns for the verbs buy and sell are: BUYER buys GOODS from the SELLER for MONEY. SELLER sells GOODS to the BUYER for MONEY.

[$_{Buyer}$ His] [$_{Money}$ $20] **transaction** [$_{Seller}$ with Amazon.com] [$_{Goods}$ for a new TV] had been very smooth.

Core FEs:
BUYER [Byr]: The BUYER wants the GOODS and offers MONEY to a SELLER in exchange for them.

GOODS [Gds]: The FE GOODS is anything (including labor or time, for example) which is exchanged for MONEY in a transaction.

MONEY [Mny]: MONEY is the thing given in exchange for GOODS in a transaction.

SELLER [Slr]: The SELLER has possession of the GOODS and exchanges them for MONEY from a BUYER.

Non-Core FEs:
MEANS [Mns]: The means by which a commercial transaction occurs. Semantic Type: State_of_affairs

RATE [Rate]: Price or payment per unit of GOODS.

UNIT [Unit]: The Unit of measure of the GOODS according to which the exchange value of the Goods (or services) is set. Generally, it occurs in a by-PP.

Frame-frame Relations:
Inherits from: `Reciprocality`

Is Inherited by:

Perspective on:

Is Perspectivized in:

Uses:

Is Used by:

Subframe of: `Commerce_scenario`

Has Subframe(s): `Commerce_goods-transfer`, `Commerce_money-transfer`

Precedes:

Is Preceded by:

Is Inchoative of:

Is Causative of:

See also:

Lexical Units: transaction.n

FIGURE 2.1: `Commercial_transaction` in BFN: definition and FEs.

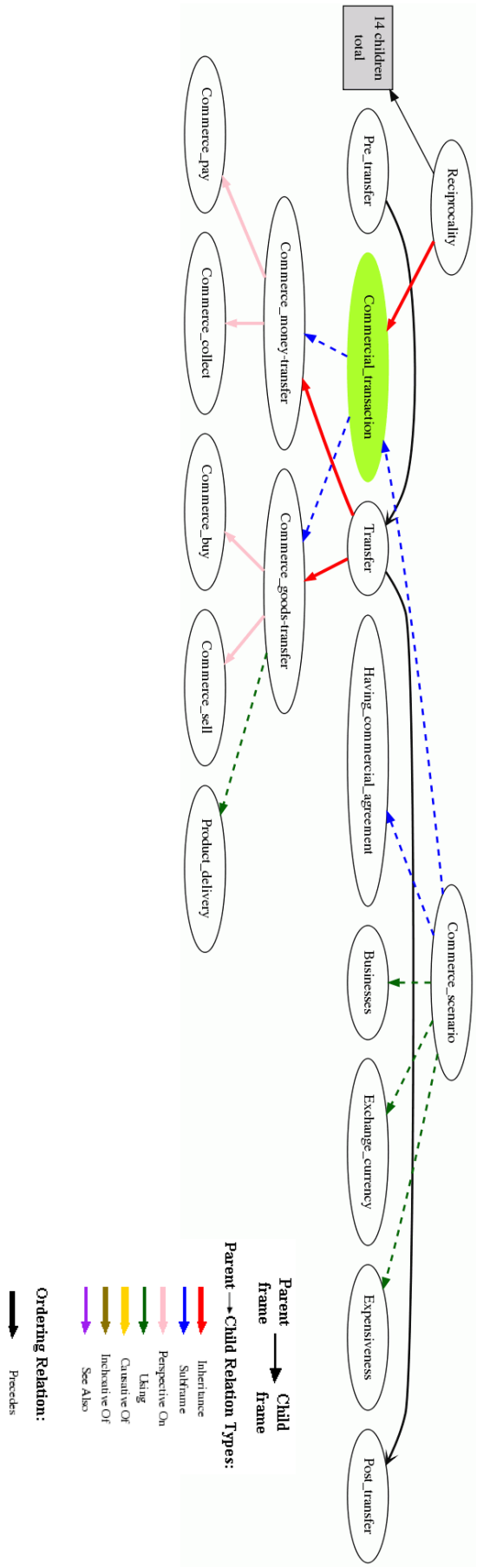


FIGURE 2.2: `Commercial_transaction` in the BFN frame network.

transaction.n

Frame: `Commercial_transaction`

Definition: COD: an instance of buying or selling.

Frame Elements and Their Syntactic Realizations

The Frame Elements for this word sense are (with realizations):

Frame Element	Number Annotated	Realization(s)
BUYER	(5)	INI.– (4), Poss.Gen (1)
GOODS	(5)	INI.– (4), PP[of].Dep (1)
MONEY	(5)	INI.– (5)
SELLER	(5)	INI.– (5)

Valence Patterns:

These FEs occur in the following syntactic patterns:

Number Annotated	Patterns			
5 TOTAL	BUYER	GOODS	MONEY	SELLER
(3)	INI –	INI –	INI –	INI –
(1)	INI –	PP[of] Dep	INI –	INI –
(1)	Poss Gen	INI –	INI –	INI –

However, reports of **transactions** [$_{Goods}$ of various dual-use material] is publicly known. INI INI INI

There were rumors in 2003 of Burmese plans to purchase ballistic missiles from North Korea, but it is unclear whether any **transactions** have been completed. INI INI INI INI

QN : Has the **transaction** taken place? INI INI INI INI

However, reports of **transactions** [$_{Goods}$ of various dual-use material]$_{\textbf{\textit{PP[of] Dep}}}$ is publicly known. INI INI INI

But why might potential buyers of financial assets delay [$_{Buyer}$ their]$_{\textbf{\textit{Poss Gen}}}$ **transactions** and instead hold their wealth in a perfectly liquid form with a zero or low yield, e.g. cash or current account deposits? INI INI INI

FIGURE 2.3: *Transaction.n*: an example of a lexical entry in BFN.

classify them with WordNet categories (Ruppenhofer, Ellsworth, Petruck, Johnson, & Scheffczyk, 2006, p. 79-90).

BFN proceeds frame-by-frame and creates all the LUs for a given frame before moving onto the next one. The general criteria for attributing lexemes to the same frame include shared aspectual profiles, sets of arguments, their types and relations between them, and the same denotation (Ruppenhofer et al., 2006, p. 9-14), although the

authors admit that not all of these criteria are applied consistently.

BFN seems to be decreasingly "cognitive" in its goals. FS came about as the study of how conceptual structures are associated with language units (Fillmore & Baker, 2010, p. 313-314). BFN frames are linguistic frames, that is, frames evoked by linguistic expressions, and their selection is guided by linguistic criteria rather than psychological evidence. The BFN homepage currently defines[3] frame as simply "a description of a type of event, relation, or entity and the participants in it."

In this study the term *frame* (marked graphically as `Frame`) refers to the FS unit of description, and conceptual structures are referred to as "cognitive frames", for the lack of better term. I will focus on the former, but overall there are many important questions about the cognitive aspects of FrameNets that need answers. To my knowledge, there have been no attempts to give the overall linguistic workflow of BFN (or any other cognitive linguistic project) a psychological footing, i.e. require that any posited linguistic distinction is grounded in psycholinguistic evidence. For example, BFN does not distinguish between synonyms and antonyms, creating frames like `Agree_or_refuse_to_act`, but it is not clear whether this decision is psychologically warranted.

2.1.4 The Constructicon

Fillmore believed that "not only words and fixed phrases, but also various kinds of grammatical features and syntactic patterns presuppose particular structured understandings of cultural institutions, beliefs about the world, shared experiences, standard or familiar ways of doing things and ways of seeing things" (Fillmore, 1985, p.231). As such, they should all be analyzable in terms of frames, and Construction Grammar (CG, e.g. A.E. Goldberg, 1995) offers a framework for doing that. CG is the earliest constructional approach to grammar, and there is a quickly growing family of construction grammars, including Sign-based Construction Grammar (I. A. Sag, Boas, & Kay, 2012), Embodied Construction Grammar (Feldman, Dodge, & Bryant, 2015), and others.

The main tenet of CG is that all linguistic units are treated as pairings of meaning and form, whether they would be considered lexical or syntactic in structural linguistics. This necessitates providing meanings even for very abstract syntactic phenomena, such as ellipsis or predication, but it also makes CG perfectly suited for work on the syntax-semantics interface. Like Langacker's cognitive grammar, CG rejects the Chomskian autonomy of syntax, as "lexicon is not neatly differentiated from the rest of grammar" (A. E. Goldberg, 1995, p. 4).

BFN now has a Constructicon, i.e. a part of the frame database dedicated to constructions. But BFN Constructicon differs from the above-mentioned grammar theories in that it is is not a theory but rather a collection of grammatical constructions. It aims to catalogue constructions in a way "compatible with the development of full grammar of the language" (Fillmore, Lee-Goldman, & Rhomieux, 2012, p. 310), but it

[3] `https://framenet.icsi.berkeley.edu/fndrupal/about`

does not aim to develop a full-fledged syntactic theory. As of 25.08.2016, it included 77 constructions[4].

Figure 2.4 shows an example of a Constructicon entry in BFN. Similarly to the frame reports exemplified in Figure 2.1, there is an informal definition of the construction. It is followed by a list of its components (CEs, construct elements) that is headed by a CEE (constructicon evoking element), and it includes annotated example sentences.

Degree_so

The construction is evoked by the CEE **so**. **So** is an adverbial modifier of a SCALAR_PREDICATE (usually an adjective or adverb), indicating the degree to which a particular ITEM has a property. In particular, this construction states that the Item has that property to an extent greater than some contextual standard. This property it retains from the use of bare adjectives, in contrast with comparative constructions, which replace the contextual standard). The extent to which the ITEM has the property is minimally bound by the RESULT_CLAUSE. Out of context and with no RESULT_CLAUSE it may be impossible to tell the exact degree indicated by the construction.

CEE (cee): The word so.

ITEM(ite): The ITEM has a scalar property to a particular extent.

RESULT_CLAUSE (res): The RESULT_CLAUSE gives an indication of the extent to which the SCALAR_PREDICATE holds of the ITEM. It takes the form of a finite clause (marked by that or not), or a non-finite verb phrase marked with as (as to make it hard to read.

The external argument of a **(so) as to...** clause may be identified with the ITEM (*that theory so complicated as to be incomprehensible*), but sometines is not, as in *that theory is so complicated as to render it useless for our purposes*, in which it is (roughly) the degree of complicatedness which renders the theory useless, not the theory per se.

SCALAR_PREDICATE (sca): The Scalar_predicate is modified by the DEGREE_MARKER. It may be inherently scalar, or construed as such.

ex.: I could never [$_{Item}$ rise] [$_{cee}$ **so**] [$_{sca}$ high] [$_{res}$ that I would forget about my loved ones].

ex.: It was not comparable to cycling, in fact [$_{ite}$ it] was [$_{cee}$ **so**] [$_{sca}$ different] [$_{res}$ that comparison was a nonsense].

ex.: That movie is so [$_{sca}$ Canadian] that it's not even a parody anymore.

FIGURE 2.4: *Degree_so*: an example of a construction entry in BFN Constructicon.

Constructicon, like FN itself, currently pursues the cherry-picking approach, and focuses on constructions that have been discussed in literature. It is probably this general lexicographic agenda that directs the project to constructions with more tangible content. Some construction grammarians try to find the semantic commonalities

[4]As displayed at `http://sato.fm.senshu-u.ac.jp/frameSQL/cxn/CxNeng/cxn00/21colorTag/index.html`.

even among abstract constructions like "Subject Auxiliary Inversion", but Fillmore et al. (2012) tentatively call them "meaningless": "while a 'metagrammar' of English might find some motivating concept that underlies uses of this pattern, the actual work of building the FrameNet Constructicon is proceeding under an assumption of the legitimacy of semantically null constructions." (Fillmore et al., 2012, p.325).

The "lexicographic" focus of the BFN Constructicon implies also that currently there are no attempts to account for sentence meanings by specifying how different parts of a complex linguistic expression combine in a single unit. The Constructicon lists different constructions, and at the moment they are simply represented as different layers of annotation. No claims are made about deriving complete interpretation of sentences from this annotation. According to Hasegawa, Lee-Goldman, and Fillmore (2014, p.197), "since FrameNet itself is a lexicographic resource, it does not provide a complete account of frame semantics."

However, this situation might change in the future releases. Fillmore and Baker (2010, p.339) state that "future FrameNet activities will be moving into the semantics of grammar, both general and abstract (negation, tense, aspect) and phraseological (constructions and syntactic idioms), making it possible in principle to test methods of integrating lexical meanings and grammatical meanings into a complete account of the language-based interpretations of texts." This dissertation makes a contribution towards this goal with regards to aspectual meanings (chapter 3).

2.2 Frame semantics in Artificial Intelligence

While Fillmore was working on the early versions of FS in the 1970s, similar ideas were put forward in AI, although with focus on information rather than language. Minsky (1974, p. 2) defined a frame as "a data-structure for representing a stereotyped situation". Schank's theory of conceptual dependencies also focused on representation of procedural knowledge, with scripts defined as "standartized generalized episodes" (Schank & Abelson, 1977, p. 19).

In the 1970s, a group of computer scientists including Terry Winograd were regularly visiting Berkeley to meet with Fillmore and Lakoff (Lakoff, 2014). They later presented a "frame-based" knowledge representation language KL-ONE. This approach to frames is different from the current FrameNet in that it is not a static relational database, but an active system for deducing information from known information. For example, CLASSIC, one of later languages in the KL-ONE family (Brachman, McGuinness, Patel-Schneider, Resnick, & Borgida, 1991), creates a self-organizing taxonomy which changes dynamically as new frames are added.

The early script-based AI approaches were one of the first attempts to tackle the problem of commonsense reasoning. Later it was mostly considered in the context of building knowledge bases with various relations between entities. One of the best-known examples of that approach is Cyc[5], the open-source release of which became

[5] `http://www.cyc.com/`

one of the sources of FreeBase[6], which in its turn became the Google Knowledge Graph[7] that we all use every day.

While the early script-based approaches were not successful, the knowledge bases are not perfect either; they typically aim at automatic extraction of relations tuples from large masses of texts (such as IsA(Lincoln, president)), and reduce the actual task of, e.g., question answering, with matching input to database entries. The commonsense reasoning still remains the Holy Grail of AI, and is unlikely to be tackled by any one discipline alone.

Interestingly, one specific type of commonsense reasoning came to be known as the Frame Problem (Hayes, 1971): it is the problem of an AI agent being unable to tell which components of a situation are affected by an event, and which of the consequences are important and worth noticing. For example, if a child kicks a ball, the dog will run after the ball, but this event actually has many other consequences: the friction will cause wear on both ball and the floor, there may be some noise or even a broken window, the ball will no longer be visible where it was visible only a moment ago, etc. Some of these consequences are more likely to be noticed by a spectator than others, but it is not clear how to predict - which ones.

From the point of view of Fillmore's FS, it is possible to model all these consequences through various scenarios, but it remains to be demonstrated how they can be linked together in a single event, and how the attention selects which ones are to be "activated". The Frame Problem is interesting not only for AI experts, but also for linguists and psychologists, and solving it requires collaboration between all these fields.

2.3 Frame semantics in Natural Language Processing

Natural Language Processing (NLP) is a cross-disciplinary research area involving linguists, computer scientists, and experts in AI. It broadly focuses on computer interpretation and generation of human language, typically viewed through the lens of smaller "tasks" or applications, such as question answering, text summarization, semantic parsing, part-of-speech tagging, sentiment analysis, etc.

BFN attracted attention of NLP specialists from its very beginning. Its past co-investigators include D. Jurafsky (question answering), J.Mark Gawron (machine translation), and S. Narayanan (information extraction). This interest is not dying off: as of 09.10.2015, searching for "FrameNet" in the electronic library of Association for Computing Machinery brings up 982 hits. To give a few examples, BFN is being used as the "golden standard" of semantic annotation (Chambers & Jurafsky, 2009), as a training dataset for classifiers (Riaz & Girju, 2014), or as a source of heuristic rules for discovering relations in texts (Aharon, Szpektor, & Dagan, 2010). As of now, most of these approaches are "symbolic" in the sense that they rely on text patterns

[6] https://developers.google.com/freebase/

[7] https://www.google.com/intl/es419/insidesearch/features/search/knowledge.html

and lexicographic resources, but there already is at least one proposal that attempts to boost the traditional semantic parsing with word embeddings (Y.-N. Chen, Wang, & Rudnicky, 2014).

BFN is not only a source for many NLP applications, but also a recipient. There already exist several automatic frame-semantic parsers including Shalmaneser (Erk & Pado, 2006) and SEMAFOR (Das, Schneider, Chen, & Smith, 2010). There is a wide range of proposals for using FrameNet data for cross-lingual transfer (Annesi & Basili, 2010; B. Chen & Fung, 2004; Kim, Hahm, & Choi, n.d.; Padó & Lapata, 2005), semantic role generalization (Matsubayashi, Okazaki, & 'ichi Tsujii, 2009) and for linking FrameNet to other resources such as Wikipedia (Tonelli & Giuliano, 2009), WordNet (Burchardt, Erk, & Frank, 2005), VerbNet (M. Palmer, Bonial, & McCarthy, 2014), among others.

However, this is not to say that there is nothing left to do. By far the most frequent "complaint" by NLP researchers that use BFN is its low coverage (Kaisser & Webber, 2007; A. Palmer & Sporleder, 2010). For example, in the experiments of Wang, McAllester, Bansal, and Gimpel (2015), only 50% of target vocabulary could be mapped to the correct frames using BFN. This explains the numerous proposals for its automatic extensions, such as the system by Rastogi and Van Durme (2014). Other general complaints that have been voiced include the lack of formalization (Chang, Narayanan, & Petruck, 2002) and missing annotations (and with them – syntactic patterns) (Litkowski, 2010, p. 303).

Concerning more task-specific issues, by now there is a rather long list. While no resource can ever hope to cater to the needs of all applications, it is interesting to see what tasks called for what features in the context of future developments of BFN and other FS-based resources.

- *paraphrasing task:* FrameNet is a powerful tool for linguistic analysis of paraphrases (Hasegawa, Lee-Goldman, Ohara, Ellsworth, & Fillmore, 2012), but Ellsworth and Janin (2007, p. 148) found that lumping synonyms and antonyms in the same frame leads to contradictory paraphrases;
- *semantic role labeling task*: lack of selectional restrictions limits BFN's utility for disambiguating polysemous words (Ovchinnikova et al., 2014);
- *semantic role labeling task*: the grammatical annotations of BFN are not compatible with dependency parsers (Fürstenau & Lapata, 2012);
- *implicit semantic role labeling task*: absence of annotations for antecedents of definite null instantiations limits BFN's usefulness as training data (Feizabadi & Padó, 2015, p.41);
- *question answering task*: lack of annotations of peripheral adjuncts prevents FrameNet from handling "When" or "Where" questions, and it is considerably outperformed by PropBank (Kaisser & Webber, 2007, p. 46);
- *event identification task*: problems with frame-to-frame relations, particularly with causative/inchoative distinction, interfere with identification of complex events (Chambers & Jurafsky, 2009, p. 608);

Many of the issues mentioned above stem from the fact that BFN was designed and intended to be a lexicographic project rather than, e.g., an inference-friendly resource. Some of the distinctions it makes (e.g. "Uses" frame-to-frame relation) could be unnecessary in a particular NLP application, and some that it needs (e.g. antonymy) would be missing.

It goes without saying that the design of any research project is governed by its goals, and it is impossible to encompass everything. It is also true that many of the tasks mentioned above were established after BFN appeared. However, now there is a niche for a resource that would "fit the bill" of such NLP applications.

The above "laundry list" brings up the question of what kind of frame-semantic resource could accommodate more practical concerns of NLP applications than BFN does. In this thesis, I am focusing on the features that would make FBCL more useful to inference-driven NLP applications, and I would argue that consistent frame-to-frame relations, distinction between synonymy and antonymy, and support for selectional preferences are essential for this.

And, of course, the biggest concern with BFN and FBCL in general is its limited coverage. Since development of frames requires so much time and resources, it is essential that NLP, in its turn, helps FN to develop ways of automatic or at least semi-automatic induction of frames and frame-to-frame relations. By this I mean not just ways to extend the current database with more synonyms or translations, but ways to further develop it, identifying new frames, examples and relations.

2.4 Distributional semantics: a very brief introduction[8]

2.4.1 Building vector space models

DS, like FS, is currently not a full-fledged semantic theory, but rather an approach to representing and working with meaning based on word distributions in corpora. Conceptually it is based on the so-called distributional hypothesis, which considers "meaning as a function of distribution" (Harris, 1954, p. 155). Another famous quotation is "You shall know the word by the company it keeps!" (Firth, 1957, p. 10), i.e. by the context it appears in. For example, the word *croissant* is more likely to be found in the context of words like *sweet*, *butter*, or *breakfast* than with words like *megabytes*, *cylinders* or *decibel*. Following Wittgenstein's "meaning is use" slogan, some researchers suggest that "the representation that captures much of how words are used in natural context will capture much of what we mean by meaning" (Landauer & Dumais, 1997, p. 218).

[8] A part of material in this section has been published: Drozd, Gladkova, and Matsuoka (2015a, © 2015 IEEE), Gladkova and Drozd (2016, © 2016 ACL). For a general overview of the field, see Erk (2012), P. D. Turney and Pantel (2010). The author's last name changed from "Gladkova" to "Rogers" in April 2017.

Although distributional models originated in the middle of the 20th century, their utility was limited by how much data could be processed. The bigger the source corpus, the more information the resulting model has. The "national" corpora of the BNC type that started appearing in the 1990s contain roughly 100 million tokens, which allows for a much larger vocabulary than the early 1-million word corpora like Brown. But, according to Zipf's law, most words will still occur only a few times, and in case of a BNC-sized corpus there will be many relatively frequent and useful words for which we will simply not have enough distributional information (e.g. the word *beetroot* occurs 35 times in BNC). When building distributed representations, words occurring under 100 or even 1000 times are typically discarded. Thus to cover words beyond the top-frequent ones we need to process corpora containing billions of words. Only very recently this became possible with consumer-grade computers.

In vector space models (VSMs), every word is represented as a vector in multi-dimensional space. In the simplest case, each dimension of a vector is a possible context where the corresponding word can occur. Based on the size and type of context, vector space models can be classified into document-based, window-based, or syntax-based. There also are various combinations of different types of context via vector concatenation or combining several vectors in second/third-order tensors. Each of these models can be further specified by various parameters, including the size of the window and penalty for the larger distance to the target word inside the window (triangular, Gaussian, etc. (Lund & Burgess, 1996)). It is also important whether the words appearing before and after the target word are counted together or independently.

From the point of view of how VSMs are constructed, they can be divided into count-based and neural-net based, or explicit vs implicit models. While the term "word embeddings" was initially applied only to the latter kind, it is now often used to refer to both, and this is how it is used in this work, interchangeably with (vector space) models.

Count-based models start with building co-occurrence matrices, where each row represents a word, and each column - a context in which it occurs. These matrices are very sparse, and also suffer from bias by total corpus frequency (as frequent words will be more frequent in any context). A typical workflow for constructing a count-based model involves some kind of normalizing the frequencies, such as Pointwise Mutual Information (PMI) (Church & Hanks, 1990). PMI quantifies the discrepancy between probability of the coincidence of two random variables, given their joint distribution and their individual distributions, and assuming their independence: $pmi(c,t) \equiv \log \frac{p(c,t)}{p(c)p(t)} = \log \frac{p(c|t)}{p(c)}$.

Sparse vectors are very large, since the number of dimensions is basically defined by the number of possible contexts. For example, a 1.2 billion token corpus yields 2 million possible contexts of words with frequency above 5, but the resulting co-occurrence matrix has only 0.5 billion non-zero elements (0.02% sparsity) and could be stored in about 2 Gb of memory space in compressed sparse row format (Drozd et al., 2015a). This kind of data can already be used in practical applications, but usually some kind of dimensionality reduction technique is used, such as Principal Component Analysis or Singular Value Decomposition (SVD).

SVD is a factorization of an $m \times n$ real or complex matrix M in a form $M = U\Sigma V^*$ (Golub & Van Loan, 1996), where U is $m \times m$ real or complex unitary matrix, Σ is $m \times n$ rectangular diagonal matrix with non-negative real numbers on the diagonal, and V^* (the conjugate transpose of V, or simply the transpose of V if V is real) is $n \times n$ real or complex unitary matrix. The diagonal entries σ_i of Σ are known as the singular values of M and are typically sorted in descending order. The matrices U and V contain left-singular vectors and right-singular vectors of M respectively.

Essentially what dimensionality reduction techniques do is trying to identify which "raw" dimensions are similar, and can be unified with the least loss of the information. This improves computational efficiency and also mitigates the effect of random noise in the data. The positive effect of dimensionality reduction has been reported in many applications that use vector space models (Bullinaria & Levy, 2012; Landauer & Dumais, 1997; Rapp, 2003). However, the downside is that the resulting space is no longer "transparent", in the sense that we can no longer directly interpret what each of the numbers means. It also means that different vector space models are no longer directly comparable: even with the same number of reduced dimensions in different models each of the dimensions is likely to stand for something different.

With implicit word embeddings, the number of dimensions is set from the beginning, and they are never "transparent". Such models map words (or other types of linguistic units) to vectors in a low dimensional space directly, without counting the co-occurrences (Pennington, Socher, & Manning, 2014). A prominent subclass of implicit VSMs is based on artificial neural networks, which are trained to correctly predict words in a given context. Thus conceptually this approach is also rooted in the distributional hypothesis.

Neural word embeddings recently attracted much attention after Mikolov, Sutskever, Chen, Corrado, and Dean (2013) showed that the vectors obtained by this method capture "linguistic regularities", i.e. that certain semantic or syntactic relations between the words correspond to linear offset between word vectors (to be discussed in detail in section 4.2.1). However, Levy and Goldberg later showed that popular neural word embeddings exhibit much of the same properties as the traditional count-based models (O. Levy, Goldberg, & Ramat-Gan, 2014), and that they essentially perform implicit co-occurrence matrix factorization (O. Levy & Goldberg, 2014b). They also behave similarly in many practical tasks, including word analogies (Gladkova, Drozd, & Matsuoka, 2016).

2.4.2 Distributional meaning representations

Distributional representations of words have the advantage of being inherently "fuzzy", which brings them closer to the connectionist models than to the neat lists of features in symbolic approaches (Lenci, 2008, p. 12). Unlike traditional binary semantic features, the dimensions of an embedding are very fine-grained, they take on continuous rather than discrete values, and each of them may represent a weak signal that works in ensemble with others (Boleda & Erk, 2015, p. 2), perhaps in complex patterns.

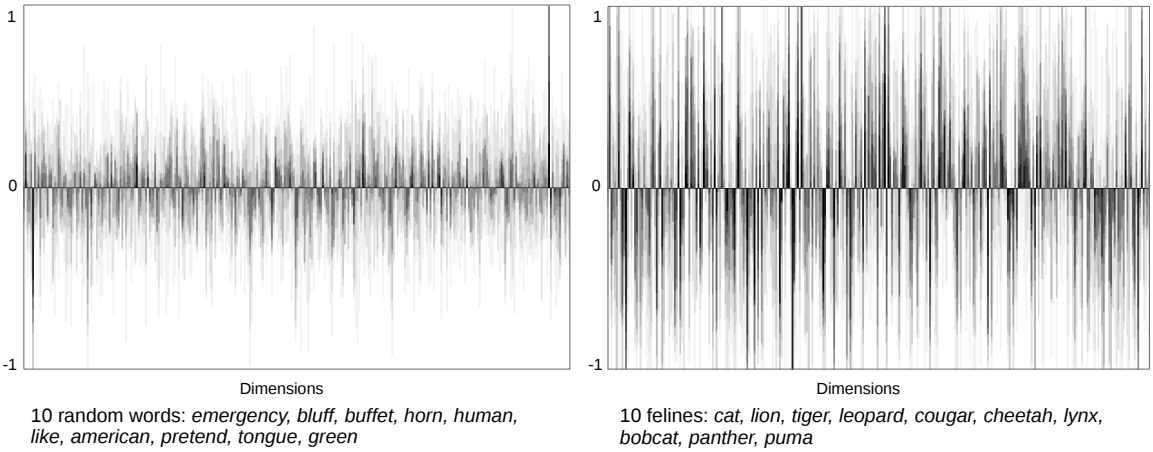


FIGURE 2.5: Heatmap histogram of 10 random words and 10 co-hyponyms in GloVe (© 2016 ACL)

Figure 2.5 (Gladkova & Drozd, 2016) is a simple visualization of the overlay of dimensions for 10 random words and 10 co-hyponyms in 300-dimensional GloVe(Pennington et al., 2014) embedding[9]. The columns of this diagram correspond to the values of 300 dimensions. The more words have similar values on a given dimension, the darker is the corresponding column.

Figure 2.5 shows that the randomly selected words do not have much in common distributionally: the overlay on most dimensions remains close to zero. However, for felines many features are shared – although it is not clear that they would all be interpretable. We could hypothesize about them ("animal"? "nounhood"? "catness"?), but GloVe clearly shows more feline features than we could find in dictionaries or elicit from human subjects. Perhaps there is a dimension (or a group of dimensions) created by the co-occurrences with words like *jump*, *stretch*, *hunt*, and *purr* - some "feline behavior" category not to be found in any linguistic resource.

Thus the way meaning is represented in word embeddings is both tempting for its cognitive plausibility, and puzzling in the sense that we do not really know yet how to work with this. There are proposals for mapping individual dimensions onto the familar linguistic features; for example, Tsvetkov, Faruqui, Ling, Lample, and Dyer (2015) propose to align dimensions of word embeddings with WordNet supersenses. But inherently any such attempt will be an attempt to convert the fuzzy and continuous into the known (from theoretical linguistics) and discrete. It would be both limited, since it necessarily involves information loss, and limiting, since it essentially brings the new model down to the old ones. So perhaps a more productive approach would be to focus on finding new ways to work with distributed representations.

[9]Unless specified otherwise, the examples cited in this section are derived from 2 word embeddings: GloVe (Pennington et al., 2014) and SVD, trained at 300 dimensions, window size 10. GloVe parameters: 100 iterations, x_{max}= 100, a = 3/4. The SVD (Singular Vector Decomposition) model was built with Pointwise Mutual Information (PMI), $a = 1$. The 5B web-corpus combines Wikipedia (1.8B tokens), Araneum Anglicum Maius (1.2B) (Benko, 2014) and ukWaC (2B) (Baroni, Bernardini, Ferraresi, & Zanchetta, 2009).

Of course, the simile between vector meaning representations in distributional semantics and patterns of neural activation also has its limits. We do not know at this point how exactly the human brain encodes meaning, but it is probably not the same matrix factorization process that is used in word embeddings. Also, needless to say, human brains and word embedding models have very different inputs (a web-corpus, even if it is very big, versus years of speech and perceptual experience). Thus the current word embeddings cannot be reasonably expected to arrive at exactly the same conceptualization schemes that humans have.

However, NLP keeps creating increasingly sophisticated models, and it is now possible to "ground" them in visual data in addition to texts (Bergsma & Goebel, 2011; Shutova, Tandon, & de Melo, 2015). There is also ongoing research on "plugging" distributional representations into a system for grammatical/logical parsing that would act as compositionality module (Lewis & Steedman, 2014), or combining multiple layers of neural networks in an end-to-end dialogue system, where some of these layers would be processing the output from others (Sukhbaatar, Weston, & Fergus, 2015). But success in any such endeavor would require better "distributional linguistics" than what we have at the moment.

2.5 Summary

In this chapter I provided an overview of FS in cognitive linguistics, from its early days into its current implementation in FrameNet projects. FrameNets are still unique among cognitive linguistic theories in its combining cognitive plausibility with relative NLP-friendliness (although lexicographic focus of FrameNets does not necessarily meet all the demands of the NLP community). Most importantly, its main focus - script-like structures in human experience - have immediate practical value in many applications, which should stimulate its faster development.

Frames have played an important role in the early days of AI, and currently FrameNets are often used for such NLP applications as inferencing and reconstructing event chains, although these applications no longer aim at being a general common-reasoning system. However, the current applications are still limited both practically and theoretically. Further success of FrameNets depends on whether it will be able to adapt to the new demands of these fields.

On the other hand, the development of computer architecture made it possible for distributional semantics to offer vector space models that do capture a lot of linguistic information. However, the theoretical framework and practical tools for working with distributed meaning representations require much future collaboration between linguists and computer scientists. This project aims to build one of the many bridges necessary for the two communities, by showing in what ways DS and FS can be mutually beneficial.

Chapter 3 "Event structure for a multilingual FrameNet" is not included in the abridged thesis due to planned publication elsewhere.

Chapter 4

Discovering verb classes in untagged corpora[1]

In section 2.3 I briefly described a range of NLP applications that make use of FrameNets, and are potentially useful for them. The current proposals that are relevant for frame-based lexicography include:

- extension of FrameNet through dictionaries such as WordNet (Johansson & Nugues, 2007);
- automatic or semi-automatic mapping of English frames to vocabularies of other languages (Fung & Chen, 2004; Hartmann, Gurevych, & Lap, 2013; Tonelli, 2010);
- induction of new frames (Green, Dorr, & Resnik, 2004);
- automatic labeling of semantic roles (Coppola, Gangemi, Gliozzo, Picca, & Presutti, 2009; Das et al., 2010; Kshirsagar et al., 2015), also cross-lingually (Johannsen, Alonso, & Søgaard, 2015, 81.6).

In addition to these frame-specific tasks, there are other, more general, NLP applications that could also be of considerable use in lexicography. For example, existing word sense disambiguation tools could be adapted to grouping corpus concordances by types of context, which would ease picking examples of use for different word senses and getting some idea about their distribution. It goes without saying that the output of such applications will be noisy – but it would still ease exploratory research.

[1] A large part of the work presented in this section was done in collaboration with Dr. Drozd of Global Scientific Information and Computing Center in Tokyo Institute of Technology. The contributions of the author consist in designing experiments, preparing data sets, evaluation and analysis of resulting data. Dr. Drozd provided early access to the in-development infrastructure for high performance natural language processing tools (work done in the scope of JSR CREST EBD project), mainly co-occurrence extraction infrastructure from very large corpora (Drozd & Matsuoka, 2014), helped to adjust it to the needs of the current research project, and run the machine-learning experiments. He is also the author of the LRCos technique employed in the second case study (Drozd & Matsuoka, 2016). Our joint research is presented in the following publications: Drozd et al. (2015a, © 2015 IEEE, DOI 10.1109/DSDIS.2015.30), Drozd, Gladkova, and Matsuoka (2015b, © 2015 ACM, DOI 10.1145/2835857.2835858), Gladkova et al. (2016, CC BY 4.0, ACL), Drozd, Gladkova, and Matsuoka (2016, CC BY 4.0, ACL), and Gladkova and Drozd (2016, CC BY 4.0, ACL).

Since this thesis argues in favor of closer collaboration between FS and DS, it is important to show the FS community the recent advances in issues that are considered more "core" FS. Many techniques proposed for exracting semantic relations (Lepage & Goh, 2009; Pantel & Pennacchiotti, 2006; Szumlanski & Gomez, 2010) could be adapted for discovery of FEs. However, in this dissertation I focused on event structure and therefore – on more abstract, morphosyntactic phenomena. The reason for this choice is that the current generation of FrameNets tend to focus on the lexical frames, and there is much work to be done on the abstract frames that could ensure consistency of the lexical frames with respect to event structure. Furthermore, aspectual classes and inchoativity/causativity are also semantic phenomena, and in the construction grammar framework they should be accounted for in FS terms, the same way as lexical constructions.

This chapter explores how DS could help FBCL in ensuring consistency of frame-to-frame relations in the database. In chapter 3 I argued that a multilingual FrameNet would require a more systematic and detailed representation of event structure. Such a model relies on systematic relations between lexical frames and the abstract event structure representation, which would enable inferencing and connecting events into event chains. This means that we need to be able to automatically induce large paradigmatic word classes that should all have the same relations to top-level frames, since establishing such relations manually would require too much effort.

For instance, the event model proposed in chapter 3 includes (1) the distinction between verbs that profile the finishing point of an event and those that do not, and (2) the distinction between causative and inchoative verbs. Accordingly, this chapter presents two case studies: automatic discovery of Russian imperfective/perfective verbs, and Japanese transitive/intransitive verbs.

4.1 Case study 1: imperfective/perfective verbs in Russian

4.1.1 Existing proposals for automatic induction of aspectual classes

It has long been shown that part-of-speech classes can be induced automatically from distributional data by clustering (Schütze, 1993). However, what would be needed for a multilingual FrameNet is more granular semantico-syntactic relations that correspond to aspectual distinctions.

There are many studies showing that very fine distinctions between classes of verbs can be made on the basis of annotated corpora. Rooth, Riezler, Prescher, Carroll, and Beil (1999) obtained 35 semantic classes partially overlapping with Levin's classes by estimation-maximization clustering of 1.3 million verb-noun pairs from BNC. Stevenson, Merlo, Kariaeva, and Whitehouse (1999) automatically induced 3 lexical classes of English verbs (unergatives, unaccusatives and object-drops) from a 65-million word corpus, on the basis of information about diatheses alternations. They experimented

with both decision-tree-based and neural-networks-based methods to achieve 55-78% accuracy.

Siegel and McKeown (2000) classified English verbs into 4 classes of states and events on the basis of 14 parser-coded grammatical context features in 2 copora (fiction and medical). 3 supervised learning methods were compared: decision tree, genetic programming's function trees, and log-linear regression (93.9%, 91.2% and 86.7% accuracy respectively).

A series of experiments by Schulte Im Walde explored a wide range of verb classification techniques. 153 English verbs were grouped into 30 Levin's classes with unsupervised hierarchical clustering, with an accuracy of 61%. The verbs were described by distributions over subcategorization frames, extracted from syntactic parses and combined with WordNet classes as selectional preferences for the frame arguments (Im Walde, 2000). The same subcategorization frame approach was applied to k-means-based unsupervised clustering of German verbs on a 35-million word corpus. A follow-up study indicated that the clustering works better with the same data and parameters when applied to a smaller number of verbs and classes (Im Walde, 2006).

A further experiment (Im Walde, Hying, Scheible, & Schmid, 2008) explored verb classification on the basis of their selectional preferences. Similarly to Rooth et al. (1999), the authors used the soft-clustering approach with the Expectation-Maximisation (EM) algorithm; but it is combined with the Minimum Description Length principle to induce WordNet-based selectional preferences for arguments within subcategorisation frames. This experiment ran on 5 million tuples (verbs and their arguments) extracted from BNC which yielded 'semantically interpretable' clusters.

Impressive results were also obtained with unsupervised clustering of Finnish verbs based on a self-organizing map (Lagus & Airola, 2005). 600 verbs were successfully grouped into semantically interpretable classes on the basis of their behavior in a syntactically parsed corpus containing 13.6-million tokens; the authors experiment with different context widths and conclude that they do not significantly affect the performance of the model when there is enough data.

The successful model by Siegel and McKeown (2000) was later extended by Friedrich and Palmer (2014), who use the arguments and modifiers of verbs to predict their being stative, dynamic, or both stative and dynamic. Among the most recent proposals in the same vein is the work by Falk and Martin (2016) who attempt to tackle aspectual variability in French on the basis of features encoded in the valency lexicon.

All of these projects use comparatively small corpora (the biggest is the 100-million-word BNC used by Im Walde et al. (2008)), and they all rely on linguistic features of verbs encoded in morphosyntactic annotation, or on noun-verb tuples extracted from a corpus, or some external linguistic resource. To our knowledge, no attempt has been made to induce aspectual classes from an unparsed corpus. However, a study on verb class disambiguation using untagged corpora (J. Li & Brew, 2007) shows that although absence of hand-tagged data decreases performance of the disambiguator, it still achieves comparable performance (average accuracy of 58-64% versus 64-69% for the classifier trained with hand-tagged data).

Thus the challenge in this chapter is to show that fine-grained semantico-morphological word classes can be induced even from "raw" text data (i.e. plain text corpora that only underwent tokenization), without recourse to syntactic parsers, lemmatizers or part-of-speech taggers. There are two reasons for this approach. First, induction of fine-grained semantico-syntactic classes from purely distributional data would lend strong support to distributional hypothesis, which enables a wide array of distributional, knowledge-poor methods in language processing. Second, tools such as part-of-speech taggers or syntactic parsers vary in accuracy and may not be available for a given language.

For the above reasons, this thesis explores the possibilities of bag-of-words word embeddings that are created on the basis of raw text corpora. However, I am not at all suggesting that symbolic, pattern-based approaches should be abandoned. In some cases distributional and symbolic approaches yield even better results together (see section 4.2.4.2). In other cases the distributional approaches cannot yet compete with rule-based tools. For example, there are currently no highly accurate and purely distributional syntactic parsers – making the existing rule-based parsers the obvious choice for projects that simply need a tool to work on some other linguistic phenomenon (as I will do in discussing selectional preferences in chapter 6).

The first case I will consider is imperfective-perfective distinction in Russian verbs. The reason for this choice is that Russian is a morphologically rich language, which presents a particular challenge for word embeddings and cosine similarity-based tasks. Basically the problem is that with more morphological forms per word the distributional semantic space becomes populated with more highly related words, which makes the margin of error very thin. Also with more word forms per word there are fewer contexts in which each of them is observed, which means that the vectors for individual word forms are going to be less informative. This effect could halve the accuracy of a system that performs well on English (Drozd et al., 2016).

4.1.2 Methodology: machine learning on distributed representations

4.1.2.1 Classification with machine learning

"Machine learning" is an umbrella term for algorithms that can discover patterns in data and make predictions on new data. This can be done with some labeled data being provided ("supervised machine learning"), without any labeled data ("unsupervised machine learning"), or with some combination of the two ("semi-supervised machine learning"). For example, a machine learning algorithm could be taught to predict whether an email is spam or not based on such information as the sender, subject keywords, certain textual features, etc. In the supervised paradigm, the algorithm would be "told" that several emails are spam, and then it would attempt to discover other emails similar to the ones it "knows" to be spam by some shared features.

Machine learning is widely used in a variety of tasks including clustering (e.g. automatic classification of emails into folders) and regression modeling (e.g. discovering

the relationship between variables such as house size and price). One of the most frequently used applications of machine learning is classification. In a supervised classification paradigm, the algorithm is provided with a set of classes and examples of items belonging to these classes. Then it attempts to attribute new items to these classes by their similarity to known items.

In this chapter, the task is to automatically discover verbs in a large untagged corpus of Russian while distinguishing between their two aspectual classes. We will accomplish this with supervised learning on the basis of a manually constructed training dataset (compiled from Russian dictionaries and grammatical reference materials).

The algorithm that will be used in this section (and also in section 4.2 and chapter 6) is logistic regression, one of the most popular machine learning algorithms. Logistic regression is often used for classification, because it can predict the probability of the item being in one class or another based on its features. It is possible to do binary or multinomial classification, i.e. classification into 2 or more classes. The predictions are made via logistic function. The overall formula for calculating the probability that x belongs to a given class is as follows (James, Witten, Hastie, & Tibshirani, 2013, p.134):

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 x}} \tag{4.1}$$

In this formula, e is the Euler's constant, and β_0 and β_1 are regression coefficients that essentially delineate the classes. The circumflexes ($\hat{p}$) indicate that these are estimated values, found through a process of maximum likelihood estimation to best fit the training data (i.e. the data points for which the correct classes are known).

The benefit of using supervised learning with word vectors as opposed to knowledge-based methods of working with corpora is primarily that we can leverage latent distributional information. The vector space models such as count-based SVD models or GloVe represent vocabulary differently from dictionaries: each word is characterized in terms of hundreds of distributional features simultaneously. If distributional hypothesis is correct, linguistic phenomena such as verb classes or selectional preferences are encoded somewhere among these dimensions. If so, then supervised learning algorithms such as logistic regression should be able to discover these linguistic structures despite noise, i.e. hundreds of distriutional features that are irrelevant for a given phenomenon.

It could be argued that conceptually, supervised machine learning is not so different from using a parser or tagger, since in both cases there is an external, non-distributional source of knowledge. However, there is a big difference in the extent of such external knowledge: to provide only examples of the target classes (which the classifier will then learn on the basis of purely distributional features), or to transform the whole corpus, essentially providing more information for each word in it.

For NLP tasks, the difference in the amount of handcrafted resources that are necessary can make a big difference. In many cases it is much more feasible to construct a dataset for supervised learning than obtaining a high-quality annotated corpus. For

instance, many approaches for aspectual verb classification reviewed in section 4.1.1 relied on English treebanks - but such resources may be unavailable for other languages in sufficient quantity. For example, Russian has a lot of syntactic homonymy, and the Russian National Corpus currently has only 516,852 disambiguated[2] sentences – less than 6 million tokens.

There is also a crucial difference at the *level* at which the external knowledge is used. In case of aspectual verb classification, most prior work reviewed in section 4.1.1 used parsers to provide their systems with knowledge of all syntactic structures in which a given verb participated. In doing so, the researcher essentially decides *a priori* what distributional information is relevant, and should be used for verb classification. The proposed approach delegates this function to machine learning, and enables the classification on the basis of numerous, seemingly unrelated criteria that are not known beforehand. This approach could lead to unexpected discoveries, as opposed to entirely theory-driven testing of one hypothesis at a time.

4.1.2.2 Corpora and models

In scope of this project we experimented with 3 corpora listed in Table 4.1.

Table 4.1: Corpora used for automatic classification of imperfective/ perfective verbs in Russian (© 2015 IEEE)

Corpus	Tokens	Unique tokens
Russian Wikipedia	274M	3.5M
Araneum Russicum Maius	1.2B	4.6M
Self-published fiction corpus	1.2B	5.6M
Combined corpus	2.6B	10M

The Wikipedia corpus consists of the Russian Wikipedia dump from 2015.06.03[3]. Araneum Russicum Maius 14.04 is compiled by SpiderLing web crawler; the texts were stripped of html markup and deduplicated (Benko, 2014). The corpus is distributed with morphological annotation by TreeTagger, which we did not use. In all three corpora we replaced all occurrences of 'ë' with 'e': although the correct spellings of many words include letter 'ë', it is commonly misspelled as 'e', which leads to hundreds of alternative spellings.

The self-published fiction corpus (SPFC) is assembled from texts published at `http://www.proza.ru`. This is the biggest Russian portal for aspiring writers; it now has over 5 million texts by over 220 thousand authors, the earliest publication date being 2002. Self-published fiction is a particularly interesting genre for lexicographic goals, since new fiction by a variety of writers is likely to contain many novel words that could not be obtained from existing dictionaries.

[2] `http://www.ruscorpora.ru/corpora-stat.html/`

[3] Plain text was extracted with `wikiextractor` tool v.2.55, (`https://github.com/attardi/wikiextractor`).

As described in section 2.4, the distributed representations of words can be built "explicitly", on the base of co-occurrence counts from corpora, or "implicitly" with methods based on neural nets. This project uses an "explicit" or count-based model, since the two approaches were shown to be mathematically similar (O. Levy & Goldberg, 2014b), but count-based models have the advantage of having more interpretable parameters.

The window size was set to 2 (both left and right contexts, i.e. before and after the target word), since this context size showed best results in our pilot studies. The sparse co-occurrence matrix contained PMI-weighted corpus frequencies, which helps to deal with bias by total corpus frequency. It was then reduced to 1000 dimensions with SVD technique. Only words occurring at least 10 times were included in the model vocabulary. The same procedure was applied to the combined corpus, which was a simple concatenation of the other three corpora.

4.1.2.3 Dataset

Our task could be formulated as a binary classification task: deciding whether a given verb is perfective or imperfective. However, in addition to discovering verbs, we also need to filter out words of other grammatical categories. Thus our classifier distinguishes between three[4] classes - perfective verbs, imperfective verbs, and "other" lexical units. For each of these classes we will need to provide hand-picked examples, on which the classifier will be trained.

As mentioned in section 3.2.1, the basic aspectual distinction in Russian is between perfective and imperfective verbs. Most verbs have both versions. They can be formed with such morphological patterns as (a) perfectivization of imperfective verbs with a prefix (e.g. *delat'* - *sdelat'* "to be doing - to have done"), (b) imperfectivization of a prefixed or root perfective verb with a suffix (*zabyt'* - *zabyvat'*, "to forget - to be forgetting"), and (c) full or partial suppletion (*brat'* - *vzyat'* "to be taking - to have taken") (Shvedova et al., 1980, p. 583). Imperfectivization is very productive, and it is possible to form imperfective verbs for most perfective verbs denoting change of state. Since there are numerous morphological patterns, it is not feasible to apply a method based on orthographic patterns, such as proposed by (Soricut & Och, 2015)[5].

The accuracy of supervised classifiers relies on the quality of training data. The dataset this study uses was created on the basis of several Russian dictionaries (Hagen, 2014; Lyashevskaya & Sharov, 2009; Yevgenjeva, 1999) and grammatical reference materials (Shvedova et al., 1980).

We proceeded from a list of words in their dictionary forms which were taken from frequency lists compiled on the basis of RNC (Lyashevskaya & Sharov, 2009). We

[4]We also conducted pilot studies with classifying all words into just 2 classes (perfective and imperfective verbs), and with separate classes for every part of speech. With the same classifier parameters and the same data set, both these approaches yielded more noisy results than classification with 2 target classes and 1 "noise" class.

[5]However, the method proposed by Soricut and Och (2015) would be helpful for automatic induction of smaller Russian aspectual subclasses that are formed by regular affixation patterns, several of which were discussed in section 3.2.1

aimed to include 100 examples for each morphological group that the classifier should be able to recognize[6]: 100 perfective verbs, 100 imperfective verbs, and up to 100 words for 11 other Russian morphological categories in the "other" class. Since some classes (e.g. conjunctions and particles) generally have fewer than 100 lemmas, in the end the "other" category contained 858 words (pronouns, adjectives, adverbs, numerals, nouns, prepositions, particles, interjections, and conjunctions).

Next, the lists of words in all 3 classes were expanded with all morphological forms for each word, taken from a widely used self-published morphological dictionary (Hagen, 2014). Adjectives, nouns, pronouns, numerals, and verbs have rich paradigms in Russian; this is especially true of verbs which in our dataset are merged with their transgressive and present participle forms, which inherit the aspectual class of the source word. Because of these forms and their own morphological forms, the overall quantity of word forms in the training data is comparable across the three classes: 3995 perfective verbs; 4890 imperfective verbs; 4751 "other" words (despite the fact that the verb classes have only 100 lemmas, and the "other" class contains over 800 lemmas). The reason why perfective verbs have fewer word forms than imperfective verbs is the asymmetry in their paradigms; for example, perfective verbs lack the present tense.

The lemmas for all word classes in this dataset are listed in the appendix A.

4.1.3 Evaluation

4.1.3.1 Effect of word frequency and corpus genre

The Achilles' heel of distributional methods is that they require a lot of occurrences for all target words. According to Zipf's law, even in a very large corpus most words inevitably occur only a few times, and thus they are unlikely to yield informative representations. This means that the accuracy of classification depends on whether the target words are frequent enough in a given corpus.

The frequencies of words in our dataset in the three corpora are presented in Figure 4.1. As expected, the bigger corpora (Araneum and SPFC) contain more words that are included in the dataset than does Wikipedia. SPFC and Araneum are overall comparable, although SPFC has a slight advantage in almost all frequency ranges. Most words in the dataset have frequency above 100 in these corpora. Both SPFC and Araneum should therefore yield informative vectors for roughly 13 out of 14 thousand words in the dataset.

However, in addition to frequency, another important factor is the quality and genre of the corpus. Web-corpora such as Araneum are large and relatively easy to collect by crawling the web, but they often contain commercials, meaningless advertisements,

[6]Another pilot study estimated performance with 200 frequent verbs in each class in the training dataset, but we found no improvement over 100 verbs. The optimal amount of training data depends on the task, but it often requires even fewer than 100 positive examples. In a subsequent experiment on analogy-based identification of Russian morphological paradigms we showed that the performance of classifier saturates at roughly 30 examples (Drozd et al., 2016); see also chapter 6 for effect of training set size in "leaning" selectional preferences.

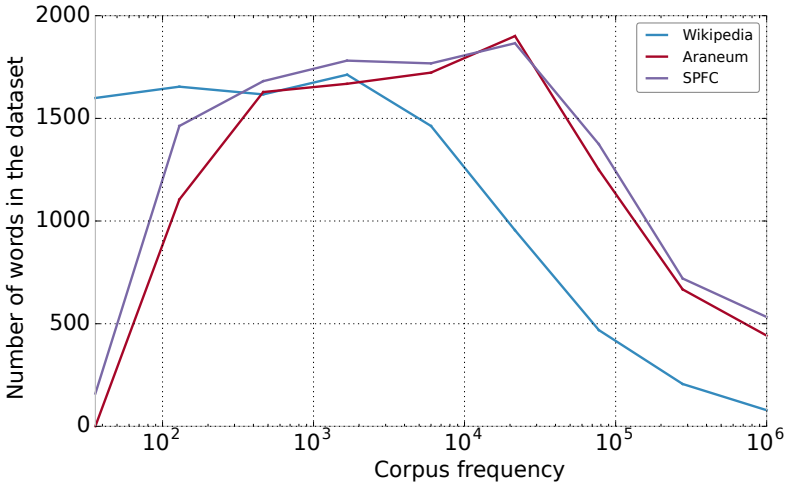


FIGURE 4.1: Frequency of words in the training dataset in corpora

duplicate texts, etc. Also, the quality of the text itself is often inferior (incorrect spelling, grammatical mistakes, incomplete sentences), which leads to a more noisy distributed representation. Wikipedia, while it is of higher quality than an average web corpus, also often includes lists of phrases rather than complete sentences due to its genre. In addition, web-corpora have to be "cleaned up" (determining languages and encodings of individual pages, deduplicating text, removing html markup and irrelevant elements such as forms and banners, etc). In long text-processing pipelines each of these steps can potentially introduce even more noise, but with corpora of such scale it is no longer feasible to check all the resulting texts manually.

Figure 4.2 shows the results of a pilot test with word vectors from all three corpora, and also their combination. The words in the dataset were classified with logistic regression[7], as described in section 4.1.2.1.

Since the labeled dataset is not very big, it is important to make sure that the results are reliable and can be reproduced with a different dataset. For this we performed 10-fold cross-validation[8]: a procedure in which the labeled dataset is randomly split in 10 parts, 1 of which is used for testing, and 9 – for training. The procedure is repeated 10 times, so that each part of the split dataset is once used for evaluation, and the average accuracy is computed. This procedure is repeated for each corpus and their combination. If a given word was not present in a given corpus, it was not included in the test.

The data in Figure 4.2 shows that the classification is the least successful on the word vectors created from the Wikipedia corpus. This could be attributed to the smaller corpus because the two other corpora (Araneum and SPFC) are comparable in size.

[7]We have also experimented with k-Nearest Neighbors (k-NN) algorithm. This is one of the simplest machine learning algorithms; it uses instance-based approach, basically comparing every item to be classified with every item in the training set to discover which elements of the training set are the most similar. In our test the performance of k-NN algorithm was inferior to that of logistic regression by 5-6%, and it was also much slower. Therefore for subsequent tests we relied on logistic regression.

[8]Cross-validation was performed with the `cross_val_score` function of `skikit` module, `http://scikit-learn.org/stable/modules/cross_validation.html`

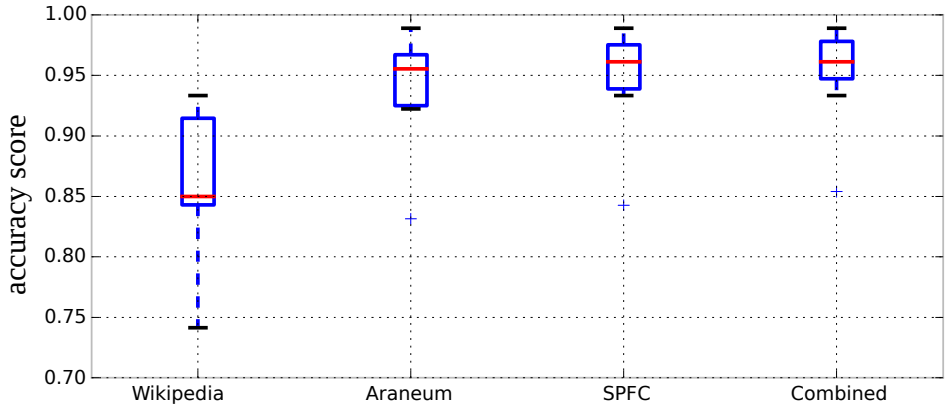


FIGURE 4.2: Average accuracy of detecting imperfective/perfective verbs from the annotated data set on 10-fold cross-validation
(© 2015 IEEE)

However, SPFC yields slightly better results than Araneum, and the performance of the combined corpus is not significantly better than performance of SPFC alone. We interpret this as evidence of higher quality of fiction texts as compared to the noisier web-crawl data, and also as evidence for existence of a certain threshold (for a given genre and quality level of texts, and for a given task), after which considerable increase in corpus size does not yield considerable gains in accuracy. Thus for the subsequent tests we will use only SPFC.

4.1.3.2 Performance on corpus data

In this step we evaluate the possibility of accurately detecting verbs in the untagged corpus data while simultaneously classifying them into perfective and imperfective verbs. Based on results of our pilot tests, we use vector space built from our best-performing SPFC corpus (1.2 billion words, no lemmatization, morphological or syntactic parsing). The entire labeled dataset described in section 4.1.2.3 is used as training data. After training, the classifier makes a decision for each of the remaining words in the corpus vocabulary, and outputs the class to which this word is attributed, together with the probability of this decision being correct.

For evaluation we randomly chose 100 words in each of 3 classes in 3 ranges of confidence (0.9 - 1.0, 0.8 - 0.9, 0.7 - 0.8), excluding verbs in the training set, and manually checked the accuracy of predictions. Table 4.2 presents the results of this evaluation.

The precision reported in Table 4.2 is calculated as "true positives / (true positives + false positives)". Consider the case of imperfective verbs. Given a sample of 300 words (100 predicted perfectives, imperfectives and "others"), we consider how many imperfectives in this sample were classified correctly (true positives), and how many were classified as imperfective verbs incorrectly (false positives). The classifier achieves over 90% precision in the confidence range 0.9 - 1.0 for both aspectual

Table 4.2: Accuracy of detecting imperfective/perfective Russian verbs in untagged corpora (© 2015 IEEE)

(a) Imperfective verbs

Confidence	0.9 - 1	0.8 - 0.9	0.7 - 0.8
Precision	0.9	0.59	0.34
Recall	0.98	0.93	0.89

(b) Perfective verbs

Confidence	0.9 - 1	0.8 - 0.9	0.7 - 0.8
Precision	0.91	0.75	0.67
Recall	0.99	0.89	0.89

classes. This result is comparable to the previous work which used small, syntactically annotated corpora.

Table 4.2 also reports recall, calculated as "true positives / (true positives + false negatives)". For example, in the case of imperfective verbs, false negatives are the imperfective verbs that were incorrectly attributed to some other class. Thus recall indicates for a given sample the ratio of the verbs of a given class that were classified correctly to the verbs of the target class that were found in the sample. In this experiment, recall stays above 80% even in the lower ranges of confidence, which indicates that we are not missing verbs.

The analysis of mistakes shows that lower precision values for lower ranges of confidence are explained by "other" words being increasingly misclassified as verbs. This could be improved by adjusting the training set. However, it is interesting that the cases of verbs classified into the wrong verb class are very rare. For example, in the random sample of 300 words in the confidence range 0.7 - 0.8 there were 31 wrong predictions for perfective and 62 for imperfective verbs (non-verbs classified as verbs), but only 6 cases of verbs attributed to the wrong verb class.

4.1.3.3 Discovering new verbs

In this experiment we check the accuracy of classifier as a tool for discovering novel lexemes not yet included in dictionaries, while simultaneously classifying them into perfective and imperfective verbs. This application of distributional semantics is of immediate interest to lexicography.

However, it is not easy to automatically select lists of new words that are not yet included in dictionaries by just checking the corpus word list against dictionaries. First, a lot of words not present in dictionaries are likely to be "noise" data, such as artifacts of incorrect encoding, brand names, etc. Second, dictionaries only list the infinitives of verbs, but a novel word could be used in a corpus in a different morphological form. The former problem is partially resolved with filtering out words that contain non-cyrillic characters, but the second would require a more sophisticated morphological module than is currently available for Russian. Thus, in this evaluation, we are limited to dictionary forms of verbs.

To get an estimate of how much linguistically interesting material can be obtained in this way, we selected words that the classifier trained on the entire dataset recognized as verbs at the confidence interval of 0.9 - 1.0. This procedure yielded 49326 imperfective and 28208 perfective verbs. It is possible to do the same to all of the classifier output, although there would be more noise: as novel words typically have low frequencies, their vectors can be expected to be built with less information and thus be more difficult to classify correctly.

These lists are still far from clean: there are proper nouns, misspelled words, words from close cyrillic-based languages such as Ukrainian, words accidentally split in parts while typing, etc. Random samples from such data contain about 40-60% noise. Further filters can be applied to deal with it, such as othography checking. However, sometimes in fiction words are also misspelled or hyphenated incorrectly for stylistic reasons, so any such filtering would depend on the goals of a particular linguistic project. For example, SPFC corpus has tokens of *strojs'* - a typical army command to make a line. This verb is less frequently used in forms other than imperative, and this particular usage is a deliberate misspelling which reflects a likely pronunciation by a sergeant.

Transliteration	Class	Translation
schekotnut'	perf.	"to give someone a bit of a tickle"
ofanatet'	perf.	"to become obsessive fanatic of something"
skreativit'	perf.	"to use some artistic skill on some task, as opposed to creating "real", more traditional art"
ostogramit'sya	perf.	"to drink a hundred milliliters [of vodka]" (a double shot in the US, quadruple shot in the UK)
istselovyvat'	perf.	"to be in the process of fully covering someone/something with kisses"
skhrumkat'	perf.	"to eat something entirely, munching away at it like a rabbit eating a carrot"
povoyevat'	perf.	"to casually go to war, to have a bit of a war"'
otsmyslit'	perf.	"to consider thoroughly and interpret, literally "to mean [smth] out"
podhvalivat'	imperf.	"to only contribute flattery during conversation"
istinstvovat'	imperf.	"to speak like you are delivering the revealed wisdom to the world"
gurmanit'	imperf.	"to go gourmet, to have fine food"
otatarit'	imperf.	"to make somebody/something Tatar"
otgavkivat'sya'	imperf.	"to quarrel", literally "to be barking off at somebody"

Table 4.3: Examples of new imperfective and perfective verbs in Russian self-published fiction (© 2015 IEEE)

For our purposes we used a simple filter of verb dictionary forms by three frequent endings (-ить, -ать, -ться), which left us with 3729 putative imperfective and 2269 perfective verbs not registered in the hunspell dictionary ("Russian Hunspell Dictionary [GNU Lesser GPL]," 2013) and 4-volume Russian academic dictionary (Yevgenjeva, 1999). We estimated accuracy by manually checking 50 verbs classified as perfective and imperfective at confidence range above 0.9. For both perfective and

imperfective verbs, 45 out of 50 were classified correctly, which is consistent with our general classifier performance evaluation (Table 4.2). However, we must note that a considerable portion of these verbs (23 out of 100) were misspellings and dialectal spellings rather than novel verbs.

Still, such lists provide a wealth of linguistically interesting data, some examples of which are provided in Table 4.3. Some of these verbs are more widely used than the others, but all are easy for native speakers to produce and understand when occasion calls for such complex concepts.

Thus case study 1 showed that it is possible to use word embeddings derived from "raw" text corpora to learn high-quality representations of aspectual classes, even in a morphologically rich language like Russian. In FBCL context this technique would be useful for establishing systematic frame-to-frame relations between lexical and abstract frames that serve as aspectual templates. Furthermore, this technique is generally useful for lexicography, since with a high-quality corpus it enables quick and large-scale discovery of novel lexemes of a particular class.

4.2 Case study 2: intransitive/transitive verbs in Japanese

This section extends the study on Russian imperfective/perfective verbs in several directions. First, I consider Japanese, a typologically different language, and a different linguistic relation: intransitive vs transitive verbs (e.g. *ochiru* "to drop" (self-motion), and *otosu* "to drop" (something)).

Second, this experiment also poses a more challenging task: automatic discovery not only of classes of verbs, but also of pairs of verbs holding the same relation. If the distinction between causative and inchoative frames is to be set via inheritance from the top-level event structure, as I proposed in section 3.3, then it would be helpful to be able to not only automatically identify large groups of verbs, but also to search for corresponding verbs that differ by event structure, but that should belong to the same scenario.

4.2.1 Word analogies in distributional semantics: success stories and limitations

As described in section 2.4, in distributional semantics words are represented as vectors, and thus any relations between word vectors need to be determined mathematically. In geometrical terms, the most intuitive measure of "relatedness" is how close the vectors are in space, usually measured as cosine of the angle between them. This creates many possibilities for comparing words by distance – but, unfortunately, cosine similarity alone does not provide any means to distinguish between different types of linguistic relations (as will be discussed in section 7.2).

One of the possibilities to pick a specific relation is through proportional analogies of the *a:b::c:d* kind. Analogies have also been long rejected in generative linguistics as an explanation for language acquisition through discovery, although now they are making a comeback (Itkonen, 2005, p.67-75), and there are several theoretical issues that will be discussed in section 4.2.4.1. But for now, let us consider them as a tool that may be imperfect, but that has long and successfully used in NLP for a variety of semantic tasks, including word sense disambiguation (Federici, Montemagni, & Pirrelli, 1997), detecting different semantic relations such as synonymy and antonymy (P. D. Turney, 2008), and ConceptNet relations and selectional preferences (Herdağdelen & Baroni, 2009).

In the current NLP discussion, analogical reasoning is understood as any "linguistic regularities" between pairs of words that correspond to regular mathematical relations. They are not limited to semantic phenomena, and just as well used for morphological analysis (Lavallée & Langlais, 2010; Soricut & Och, 2015). There are also proposals that utilize analogies as a single mechanism for detection of both morphological and semantic features (Lepage & Goh, 2009).

The current analogy boom in NLP started when Mikolov, Sutskever, et al. (2013, p. 4) showed that several linguistic relations can be modeled in word embeddings as linear offset between word vectors. For example, the result of calculation $\overrightarrow{madrid}$ - $\overrightarrow{spain}$ + $\overrightarrow{france}$ gives a point in vector space that is close to $\overrightarrow{paris}$[9].

Figure 4.3 provides a visualization of such relations for countries and capitals in the same GloVe model that I had used for illustration in section 2.4. In this visualization, multi-dimensional vectors are reduced to two dimensions with PCA (principal components analysis) technique. The resulting pairs of countries and capitals are mostly aligned.

This is an impressive result, given that it was obtained from merely a large volume of "raw" text data. It lends strong support to the idea of the close relationship between word distributions and their meanings.

Mikolov referred to this phenomenon as "linguistic regularities", a term now used to refer to any "similarities between pairs of words" (O. Levy et al., 2014). However, his study was intended as a demonstration of the possibility of discovering pairwise relations, and not of the extent to which it is possible. The datasets that were used to test how well word analogies can be solved with word embeddings only included a certain type of relations (semantic-only: SAT (P. Turney, Littman, Bigham, & Shnayder, 2003), SemEval2012-Task2 (Jurgens, Turney, Mohammad, & Holyoak, 2012), morphology-only: MSR (Mikolov, Yih, & Zweig, 2013)). The tasks in the SAT and SemEval2012-Task datasets were also different: solving a multiple-choice test in the former, and determining prototypicality of a relation in the latter.

Mikolov, Chen, Corrado, and Dean (2013) developed a new dataset that came to be known as the Google analogy test. It contains 9 morphological and 5 semantic categories, with 20-70 unique word pairs per category which are combined in all possible ways to yield 8,869 semantic and 10,675 syntactic questions. This set became

[9] All vocabulary in vector space models is typically lowercased.

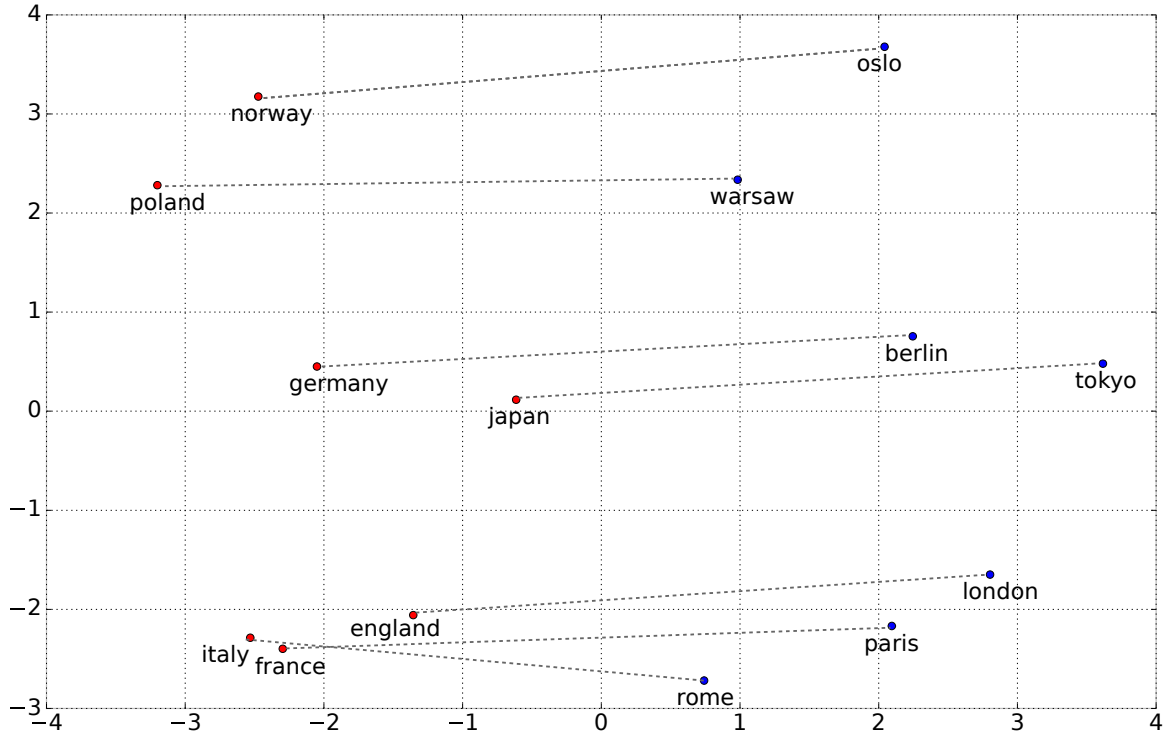


FIGURE 4.3: Linear relations between countries and capitals in GloVe

the de-facto standard for evaluating word embeddings, the assumption being that a greater number of linguistic relations corresponding to linear relations between word vectors is an indicator of the quality of the model.

However, the Google test set certainly does not show the full range of linguistic relations, and it is extremely unbalanced. With an unbalanced set, and potentially high variation in performance for different relations, it is important to consider results for individual relations, and not only the average of the whole test, as it is usually done. However, O. Levy et al. (2014) were among very few researchers who did that. They found that accuracy varied between 10.5% and 99.4%. Furthermore, merely looking at the structure of the test reveals that much success in the semantic part can be attributed to the fact that two out of five semantic categories explored the same *capital:country* relation and together constitute 56.7% of all semantic questions. This shows that a model may be more successful with some relations but not others, and more comprehensive tests are needed to show what it can and cannot do.

To investigate the extent of the "linguistic regularity" phenomenon I developed BATS, a bigger, balanced dataset (Gladkova et al., 2016) that included 10 relations for each of four types: inflectional and derivational morphology, and lexicographic and encyclopedic semantics. Each relation is represented with 50 unique word pairs, which yields 2480 questions (99,200 in total). A major feature of BATS that makes it distinct from MSR and Google test sets is that morphological categories are sampled to reduce homonymy. For example, for verb present tense the Google set includes pairs like *walk:walks*, which could be both verbs and nouns. It is impossible to completely eliminate homonymy, as a big corpus will have some creative uses for almost any word, but I excluded all words attributed to more than one part-of-speech

in WordNet (Fellbaum, 1998).

Evaluation on a count-based and neural-net-based models revealed that both of them show the same pattern of "easy" and "difficult" relations, with most of the new test set being difficult. The GloVe model that boasted 80.4% accuracy on the Google test set, achieved only 28.5% on BATS. Inflectional morphology and a part of encyclopedic relations achieved over 50% accuracy, but derivational morphology and lexicographic relations were very unsuccessful, most of them not achieving even 15%.

I interpreted these results as indicating that success of word analogies in vector space models depends on target words being frequent collocates (as are countries and capitals, particularly in the Wikipedia corpus that most studies use), or sharing many contexts. For example, the most successful lexicographic relation was binary antonymy: words in pairs like *up* and *down* both collocate with, e.g., *go* and *be*.

However, the fact that a simple bag-of-words model trained on "raw" text word embedding was able to achieve even 2% accuracy when the task was to find the correct answer out of all vocabulary which contained over 300,000 words (i.e. the random baseline would score 0.000003%), is still impressive. Furthermore, the margin of error was very thin. For example, in the person:occupation category the nearest neighbor of the hypothetical answer vector was *Depp:screenwriter* with similarity 0.36, which only slightly beat the correct answer *Depp:actor* (0.35). This indicates that word embeddings do in fact encode a lot of linguistic information, and better results should be achievable with more sophisticated methods.

4.2.2 Methodology: word analogies as a lexicographic tool for discovering pairwise relations

4.2.2.1 Pair-based vs set-based methods

As mentioned above, Mikolov, Chen, et al. (2013) were the first to demonstrate the possibility of capturing relations between words as the offset of their vectors. Given an analogy $a{:}a' :: b{:}b'$ (a is to a' as b is to b'), the answer to the question "a is to a' as b is to ?" is represented by hidden vector b', calculated as:

$$b' = argmax_{\ d \in V}(cos(b', b - a + a')) \tag{4.2}$$

Here V is the vocabulary excluding word vectors a,a' and b and *cos* is the cosine similarity distance, which is currently the de-facto standard way of measuring distance between word vectors:

$$cos(u, v) = \frac{u \cdot v}{||u||||v||} \tag{4.3}$$

This method will be referred to as **3CosAdd**. Vylomova, Rimmel, Cohn, and Baldwin (2016) use it for learning word analogies with spectral clustering and Support Vector Machines (SVM).

An alternative pair-based method was introduced by O. Levy et al. (2014) who propose to calculate the hidden vector as

$$b' = argmax_{b' \in V}(cos(b' - b, a' - a)). \tag{4.4}$$

They report that this method produces more accurate results for some categories. Its essence is that it accounts for $b' - b$ and $a' - a$ to share the same direction and discards lengths of these vectors. However, tests on BATS showed it to be consistently inferior to the other methods (Drozd et al., 2016), and it will not be considered in this study.

Linzen (2016a) reports results of experiments with 6 more functions, including reversing the relation, returning simply the nearest neighbor of the b' word, and the word most similar to both a' and b'. None of these functions outperformed 3CosAdd consistently. Reversal was beneficial for some relations, but it is only applicable to symmetrical one-on-one relations. Crucially, when the words a, a' and b are not excluded from the set of possible candidates, the performance drops to zeroes, and the closest neighbors of singular nouns tend to be their plural forms (detected with 70% accuracy as the nearest neighbors of the b word).

The vector offset approach relies on a single pair of words, which makes it sensitive to noise and word idiosyncrasies, such as differences in polysemy networks. Consider the above *king*:*queen* example: depending on the corpus, there may be more differences in their vectors than just masculinity/femininity. *queen* is also a musical group, and therefore appears in many contexts in which *king* does not appear.

The alternative is to learn the relation from a set of example pairs. The "naive" baseline would be a simple average of the offset between every pair of vectors in the training set:

$$b' = argmax\ _{b' \in V}(cos(b', b + avg\_offset)) \tag{4.5}$$

Method (4.5) will be referred to as **3CosAvg**. In this formula a and b represent words from source and target classes, and *avg_offset* is defined as:

$$avg\_offset = \frac{\sum_{i=0}^{m} a'_i}{m} - \frac{\sum_{i=0}^{n} a_i}{n} \tag{4.6}$$

Last but not the least, Drozd and Matsuoka (2016) proposed an alternative approach to discovering linguistic relations with analogies which I will refer to as **LRCos**. This method defines the analogy task in a different way to the above approaches. Suppose that we have a set of word pairs for which we know that the same relation holds between them, such as shown in Table 4.4:

Source	**Target**
France	Paris
Japan	Tokyo
China	Beijing

Table 4.4: Example analogy pairs set: capitals

In this set the right-hand-side and left-hand-side elements represent coherent groups of words - in this example, "countries" and "capitals". We shall refer to the left-hand-side of such analogies as the "source class", and to the right-hand-side - as "target class". Then the question "what is related to *France* as *Tokyo* is related to *Japan*?" can be reformulated as "what word belongs to the same class as *Tokyo* and is the closest to *France*?".

To answer this question Drozd and Matsuoka (2016) suggested training a binary classifier (in this case, logistic regression described in section 4.1.2.1) to predict if the word belongs to the target class. The source words along with random samples from the dictionary are used as negative training samples for the target words. Then the probability of a word being the correct answer for a given analogy problem is calculated by combining (in this study, multiplying) the probability of this word belonging to the target class, and its similarity with the vector a measured using angular distance.

$$b' = argmax_{\ b' \in V}(P_{\ (b' \in target\_class)} * cos(b', b)) \tag{4.7}$$

Theoretically combining similarity to the target class with proximity to a source vector enables further optimization through different weighting schemes, but so far our tests did not show significant gains over simple multiplication (Drozd et al., 2016).

4.2.2.2 Experiment set-up: corpus, models, and the dataset

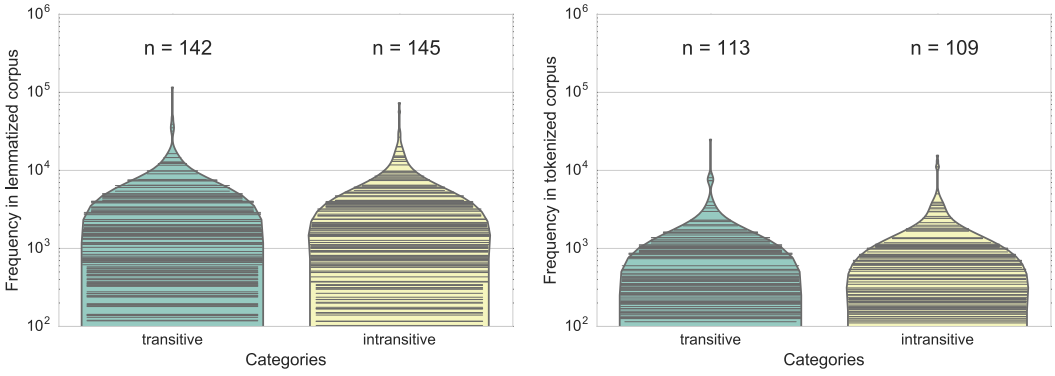
This experiment uses lemmatized and non-lemmatized versions of BCCWJ (Maekawa, 2008), the Balanced Corpus of Contemporary Written Japanese. This corpus is much smaller than the corpora used in the Russian case study in section 4.1, and lemmatization could considerably improve results by increasing amount of information that was available for constructing individual word vectors. However, with a sufficiently large corpus of Japanese it should be possible to achieve the same results with text that was only tokenized (although that in itself is not a trivial task for Japanese).

The VSM was based on Singular Value Decomposition, with Pointwise Mutual Information (PMI), parameter $a = 0.7$, 1000 dimensions. The lemmatized corpus contained 83,637,171 tokens, amounting to 2,483,231 unique words. After filtering out words occurring less than 100 times, the total vocabulary size constituted 33,007 words. For the tokenization-only corpus the total vocabulary constituted 83,887,562 words with 2,702,345 unique words, 35,940 of which occurred over 100 times.

Using this corpus, 6 models were built: window sizes 2-4 for both lemmatized and non-lemmatized version. This work was performed with the co-occurrence extraction kernel by Drozd et al. (2015b).

The dataset[10] for testing the accuracy of detecting Japanese transitive and intransitive verbs contained 159 verb pairs in the dictionary form, written with the Chinese character rather than hiragana[11]. The full test in shown in Appendix B.

The distribution of these verbs in the lemmatized and tokenized versions of the corpus is represented in violin plots in Figure 4.4. Violin plots show the distribution of the data and its probability density. The black horizontal lines correspond to data points, and darker areas indicate where more data points are concentrated. Although the shapes of the two plots are similar on the log-scale, we can see that lemmatized corpus has far more verbs that are more frequent (between 1000 - 10,000 occurrences), which theoretically should make the corresponding word vectors more informative.



(A) Lemmatized BCCWJ (B) Non-lemmatized BCCWJ

FIGURE 4.4: Transitive/intransitive verb dataset: frequency distribution in BCCWJ

Figure 4.4 shows also that the lemmatized corpus contains considerably more verbs from the dataset: 142 vs 113 transitive, 145 vs 109 intransitive. Because the test set for the tokenization-only model is smaller, these results will be less reliable, but it is still informative to see how the SVD model performs in this setting where it had less training data.

4.2.3 Evaluation

In this experiment the three methods described in section 4.2.2.1 (3CosAdd, 3CosAvg, and LRCos) were compared by accuracy with which they can detect the missing member of an intransitive/transitive verb pair.

In case of the "classic" pair-based **3CosAdd** method (4.2), all verb pairs from the testing set that were present in the corpus were combined in all possible ways to form "questions", except for repeating of the same pair. As mentioned above, the "linguistic regularities" approach assumes that proportional analogies of the a:a' :: b:b' kind can be solved as $b' \approx b$ - a + a' ($\overrightarrow{king}$ - $\overrightarrow{man}$ + $\overrightarrow{woman}$ $\approx$?$\overrightarrow{queen}$). To

[10]Japanese language learner's verb list from `http://d.hatena.ne.jp/Pulin/20150214/1423890973`.

[11]For example, 温まる: 温める, 集まる: 集める, etc.). However, the hiragana variants (あつめる, あたためる, etc.) were also included as a secondary option to be used in the test.

transfer this approach to the Japanese pairs of transitive/intransitive verbs, consider a set of word vectors $\overrightarrow{agaru} : \overrightarrow{ageru}$, $\overrightarrow{atsumaru} : \overrightarrow{atsumeru}$, $\overrightarrow{atatamaru} : \overrightarrow{atatameru}$, From this data we can form six questions:

1. $\overrightarrow{atatamaru}$ - $\overrightarrow{agaru}$ + $\overrightarrow{ageru} \approx ?$ ($\overrightarrow{atatameru}$)
2. $\overrightarrow{atatamaru}$ - $\overrightarrow{atsumaru}$ + $\overrightarrow{atsumeru} \approx ?$ ($\overrightarrow{atatameru}$)
3. $\overrightarrow{atsumaru}$ - $\overrightarrow{atatamaru}$ + $\overrightarrow{atatameru} \approx ?$ ($\overrightarrow{atsumeru}$)
4. $\overrightarrow{atsumaru}$ - $\overrightarrow{agaru}$ + $\overrightarrow{ageru} \approx ?$ ($\overrightarrow{atsumeru}$)
5. $\overrightarrow{agaru}$ - $\overrightarrow{atsumaru}$ + $\overrightarrow{atsumeru} \approx ?$ ($\overrightarrow{ageru}$)
6. $\overrightarrow{agaru}$ - $\overrightarrow{atatamaru}$ + $\overrightarrow{atatameru} \approx ?$ ($\overrightarrow{ageru}$)

Since in Japanese these verbs can be spelled with either Chinese characters (+hiragana endings) or simply hiragana, the procedure in this experiment is as follows: the verbs that form the question have to be present in the corpus vocabulary with the Chinese character spelling. But for the answer verb it was acceptable to be present in the corpus with only the hiragana spelling, since it is an acceptable correct answer[12]. This introduced an irregularity in forming the questions for 3CosAdd method: while there are 134 target verb pairs with Chinese character spellings present in the lemmatized corpus, the result was 18,360 analogy questions, and for the non-lemmatized corpus, there were 7832 questions.

The task for all methods is to find the correct answer (the missing verb) out of the whole vocabulary in the corpus, i.e. 33,007 words for the lemmatized corpus and 35,940 words for the non-lemmatized corpus. In case of 3CosAdd, for each of the automatically generated questions the vector addition/subtraction is performed according to the formula (4.2). The resulting vector is not a vector that is actually present in the vocabulary, so the answer is found by finding the existing word vector that is the closest to the hypothetical one.

The two set-based methods, 3CosAvg and LRCos, were evaluated in the so-called exclude-1 scheme. Given a set of 134 pairs present in the corpus, 1 of them is excluded, and 133 is used for obtaining the "rule" of transfer (this part differs by the method). Then the excluded pair becomes the question, and the learned "rule" is used to try to derive the answer. As shown in Figure 4.4, for the lemmatized corpus 142 questions were formed in the transitive:intransitive setting, and 145 pairs in the intransitive:transitive setting. For the tokenized corpus, it was 113 and 109, respectively.

For **3CosAvg** method (4.5), the baseline for the pair-based approach, the "rule" was learned by averaging the difference between all pairs of word vectors except two. Thus instead of $\overrightarrow{agaru}$ - $\overrightarrow{atsumaru}$ + $\overrightarrow{atsumeru} \approx ?$ $\overrightarrow{atsumeru}$, the $\overrightarrow{agaru}$ vector was added to the average of the difference between all transitive:intransitive verb pairs, and not to the difference between a specific pair.

[12]If more than one Chinese character spelling was possible, the first one was used to generate the questions, but any spelling was acceptable as the answer.

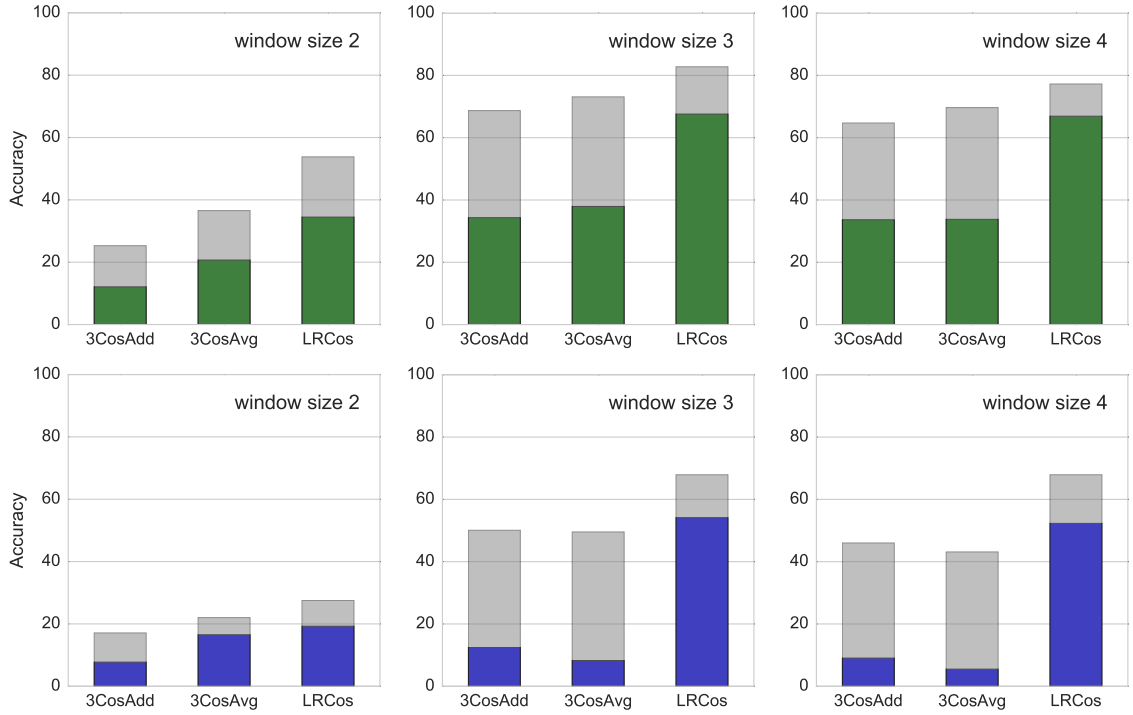
Finally, for the **LRCos** method (4.7), all available verb pairs except one are used to learn the representation of the target class. All examples of intransitive verbs are used as positive examples for training the logistic regression classifier (see section 4.1.2.1), and the set of negative examples is formed by combining the available examples of transitive verbs and the same amount of random words from the corpus vocabulary. For all word vectors in the vocabulary their similarity to the source vector and their probability of belonging to the correct target class is obtained. These two numbers are multiplied, and the best ranking vector is considered to be the right answer.

All three methods were tested on both lemmatized and tokenized-only versions of the corpus in both directions: searching for the transitive verb when the intransitive verb is known, and vice versa. The results of these experiments are presented in Figure 4.5. Two accuracy ranges are reported, Top-1 and Top-5. Top-1 condition is shown with green color for the lemmatized corpus and blue – for the tokenized corpus; this means that the nearest neighbor of the hypothetical answer vector was a correct answer to the analogy problem. Top-5 condition, shown with gray color for both corpora, indicates the accuracy when a correct answer is found among the top five nearest neighbors of the hypothetical answer vector.

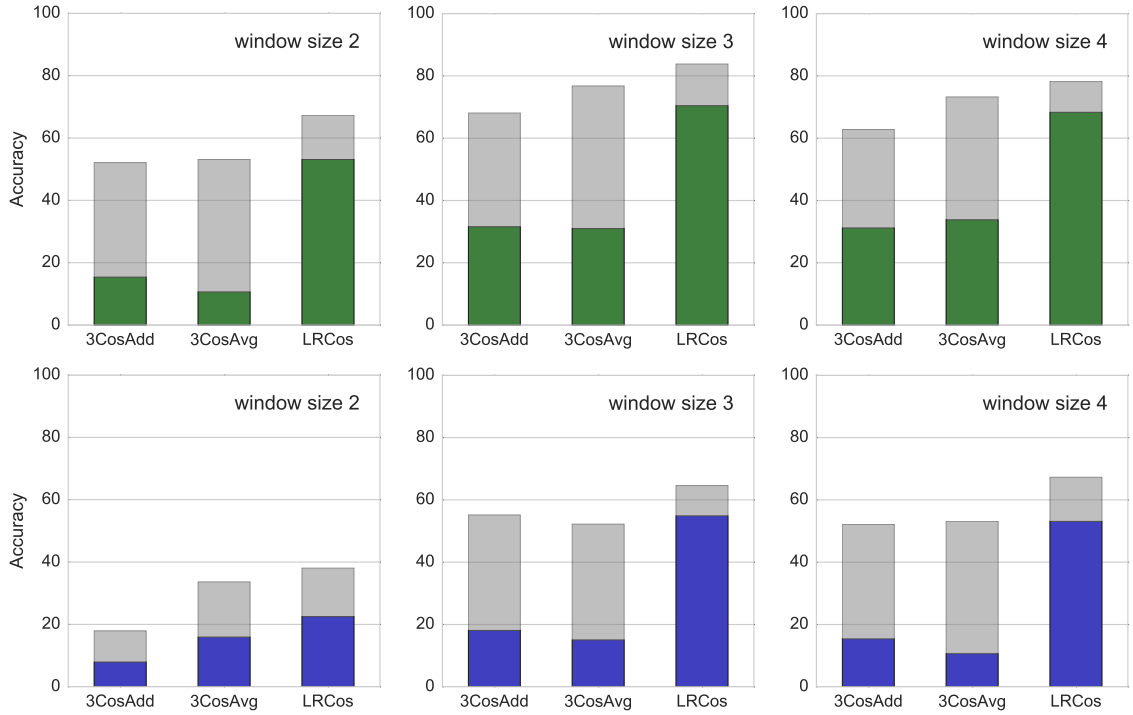
Figure 4.5 demonstrates the following facts:

- The lemmatized corpus offers consistently higher accuracy than the tokenized corpus, despite 30% more verbs in the test. I attribute this to higher quality of vectors that were built from more data for each verb.
- The overall pattern of results is similar in the transitive-intransitive and intransitive-transitive directions (although not identical), which suggests a high degree of symmetry in how Japanese transitive/intransitive verbs are encoded in the vector space.
- In all settings LRCos method outperforms 3CosAdd and 3CosAvg, consistent with findings for English (Drozd et al., 2016).
- Window size 3 seems to be the most beneficial for the SVD embedding in this task, consistent with the English data for many other categories (Gladkova et al., 2016). For the sake of comparison, these prior results for 3CosAdd and LRCos on the 40 linguistic relations in BATS dataset are shown in Figure 4.7.
- The margin of error in solving analogies with word embeddings is small, especially when we look at data for the window size 3. While 3CosAdd and 3CosAvg methods mostly fail to detect the correct answer as the nearest neighbor of the hypothetical vector, they come quite close: in most cases the accuracy is nearly doubled when we consider top-5 nearest neighbors.

The above results support the idea that DS could be useful as a tool for establishing frame-to-frame relations relevant to event structure, or checking the consistency of the database in this respect. However, while accuracy in the range of 60% is impressive from the academic point of view (once again, the choice is made from over 30,000 options, making use only of the distributional information encoded in the word vectors), from the point of view of practical lexicographic applications we could wish for further improvement.



(A) Transitive-intransitive verbs



(B) Intransitive-transitive verbs

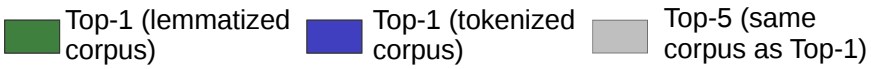
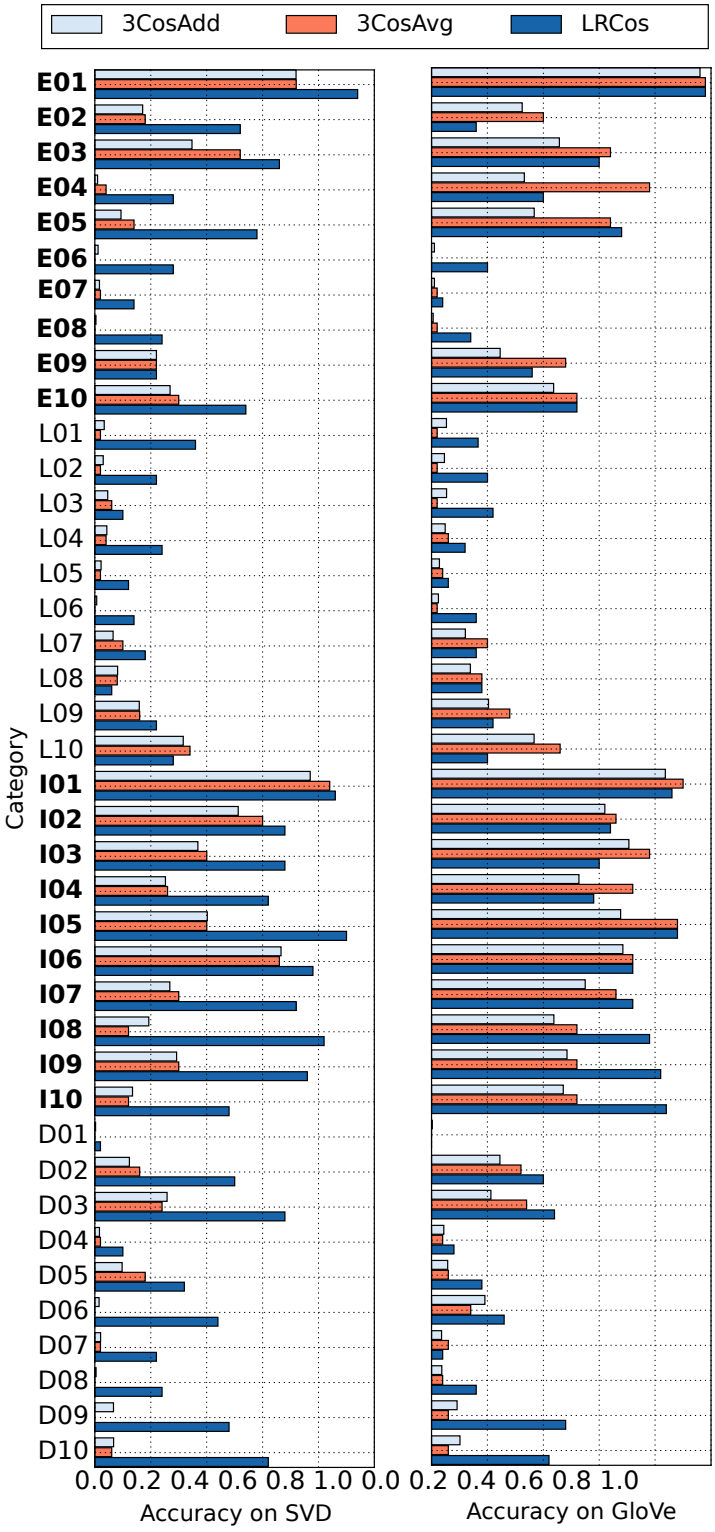


FIGURE 4.5: Accuracy of solving word analogies with Japanese intransitive/transitive verbs: 3CosAdd, 3CosAvg, and LRCos



Encyclopedic relations

E01: geography: capitals (*Athens:Greece*)
E02: geography: languages (Peru:*Spanish*)
E03: geography: UK counties (*York:Yorkshire*)
E04: people: nationality (*Lincoln:American*)
E05: people: occupation (*Lincoln:president*)
E06: animals: the young (*cat:kitten*)
E07: animals: sounds (*dog:bark*)
E08: animals: shelter (*fox:den*)
E09: thing:color (*blood:red*)
E10: male:female (*actor:actress*)

Lexicographic relations

L01: hypernyms: animals (*turtle:reptile*)
L02: hypernyms: miscellaneous (peach:*fruit*)
L03: hyponyms: miscellaneous (*color:white*)
L04: meronyms: substance (*sea:water*)
L05: meronyms: member (*player:team*)
L06: meronyms: part-whole (*car:engine*)
L07: synonyms: intensity (*cry:scream*)
L08: synonyms: exact (*sofa:couch*)
L09: antonyms: gradable (*clean:dirty*)
L10: antonyms: opposites (*up:down*)

Inflectional Morphology

I01: noun sg:pl (regular) (*student:students*)
I02: noun sg:pl (irregular) (*wife:wives*)
I03: adjective: comparative (*strong:stronger*)
I04: adjective: superlative (*strong:strongest*)
I05: infinitive: 3Ps.Sg (*follow:follows*)
I06: infinitive: participle (*follow:following*)
I07: infinitive: past (*follow:followed*)
I08: participle: 3Ps.Sg (*following:follows*)
I09: participle: past (*following:followed*)
I10: 3Ps.Sg : past (*follows:followed*)

Derivational Morphology

D01: noun+less (home*:homeless*)
D02: un+adjective (*able:unable*)
D03: adjective+ly (*usual:usually*)
D04: over+adjective (*used:overused*)
D05: adjective+ness (*mad:madness*)
D06: re+verb (*create:recreate*)
D07: verb+able (*edit:editable*)
D08: verb+er (*bake:baker*)
D09: verb+tion (*continue:continuation*)
D10: verb+ment (*argue:argument*)

FIGURE 4.6: Performance of 3CosAdd, 3CosAvg, and LRCos methods on 40 morphological and semantic relations in English (Drozd, Gladkova, & Matsuoka, 2016)

*The GloVe and SVD models were trained on a 5B non-lemmatized web-corpus.

The fact that Top-5 results for all methods are better than Top-1 results suggests that it should be possible to improve the accuracy if we provide the algorithm with additional information for ranking the candidate answers, and the remainder of this chapter offers suggestions on how this could be done.

4.2.4 Refining the algorithms

4.2.4.1 Pair-based or set-based?

To reiterate, LRCos performs considerably better in the task of identifying pairs of Japanese intransitive/transitive verbs (Figure 4.5). This is hardly a surprising result, given that LRCos makes use of more information than 3CosAdd (and even 3CosAvg, which merely averages offsets in a set rather than using them for classification).

However, Figure 4.5 also shows that, although 3CosAdd and 3CosAvg perform worse than LRCos, they do generally "head" in the correct direction: their Top-5 results are comparable with what LRCos achieves in the Top-1 setting – and that without recourse to machine learning and multiple pairs. This brings up the question of whether both methods are in principle capturing a comparable range of semantic information, and which one would be more promising for further development. A data-poor method such as 3CosAdd would be much easier to scale up to real-world applications.

Unfortunately, at present it seems that there is little to be done to further improve 3CosAdd, and machine learning techniques look more promising, despite their complexity and reliance for more data. This section discusses both practical and theoretical reasons for this conclusion.

Among the former, a big weakness of 3CosAdd is that its "classic" implementation relies on excluding from the pool of candidate answers the three source vectors. Obviously, this would lead the method to fail if the expected answer was among the source vectors[13]. It would also create problems with frequent misspellings or alternative spellings: indeed, many mistakes of LRCos in our case of Japanese intransitive/transitive verbs are attributed to an alternative spelling of the b vector that was not filtered out. For example, for $\overrightarrow{atsumeru} : \overrightarrow{atsumaru} :: \overrightarrow{ageru} : ?\overrightarrow{agaru}$ 3CosAdd would return $\overrightarrow{ageru}$ written with hiragana, since only the Chinese character was filtered out. Thus, to apply 3CosAdd on a large scale for, e.g., automatically inducing morphological paradigms, we would first need to gather all data on the spelling variation – which would be at least as laboursome as gathering training data for LRCos.

But even for completely unique word pairs without alternative spellings the exclusion of the source vectors in 3CosAdd is masking a problem. The very formula $king - man + woman$ creates the impression that the semantic features of these three source vectors can be differentiated and re-combined. However, in fact we are relying on the cosine similarity to the product of this calculation, and, with vectors containing

[13]This explains the low performance on categories with non-unique target vectors, such as category E09 *thing:color* in BATS 4.6. Basically, the "classic" 3CosAdd can not solve analogies such as *sugar:white :: snow:?*

hundreds of dimensions that are all equally participating in the cosine similarity, we are not guaranteed any controlled semantic shifts.

Let us see what would really happen in case of Japanese intransitive/transitive verbs if the source vectors were not excluded (Table 4.5). In case of vectors built with window sizes 3 and 4, we can see that over 95% of analogy problems $b + (a' - a)$ land on vector b rather than b' (which led to a mistake in my initial 3CosAdd experiments). The remaining mistakes fall onto a', which has a more prominent place with window size 2. Thus in this case a more accurate representation of the 3CosAdd method would be $\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman} = \overrightarrow{king}$, not $\overrightarrow{queen}$.

TABLE 4.5: Relative frequencies of the source vectors returned as Top-1 answers by the "honest" version of 3CosAdd (Japanese intransitive/transitive verbs)

Corpus	Window	Condition	a	a'	b	b'
lemmatized	2	intransitive-transitive	0	20.35	79.65	0
lemmatized	2	transitive-intransitive	0	22.47	77.53	0
lemmatized	3	intransitive-transitive	0	1.58	98.42	0
lemmatized	3	transitive-intransitive	0	1.27	98.73	0
lemmatized	4	intransitive-transitive	0	1.67	98.33	0
lemmatized	4	transitive-intransitive	0	1.51	98.49	0
tokenized	2	intransitive-transitive	0	23.52	76.48	0
tokenized	2	transitive-intransitive	0	39.21	60.79	0
tokenized	3	intransitive-transitive	0	2.66	97.34	0
tokenized	3	transitive-intransitive	0	3.93	96.07	0
tokenized	4	intransitive-transitive	0	3.33	96.67	0
tokenized	4	transitive-intransitive	0	4.79	95.21	0

That the $b + (a' - a)$ would land on b is easy to explain if the vectors a' and a are so similar that subtracting them does not result in sufficiently large a change to b. In this case, the "classic" 3CosAdd would be yielding correct answers only in cases where b' is the nearest neighbor of b, disregarding vectors a and a' entirely. This phenomenon has been pointed out by Linzen (2016b) for English: analogies for singular:plural nouns in English are solved with 70% accuracy by simply taking the nearest neighbor of the b vector, irrespective of $a' - a$. In the current case, the embeddings built from the lemmatized BCCWJ corpus have 36 out of 142 intransitive verbs (in the Chinese character spelling) as the nearest neighbors of the transitive verbs (in either spelling), which is overall consistent with the 3CosAdd results reported in Table 4.5.

This behavior casts a serious doubt on the potential of vector arithmetic as a means to capture semantic shifts between words. 3CosAdd method often does work, but it works for a different reason than what we might gather from $king - man + woman$ formula. In the majority of cases its success results not from any composition of semantic features, but from the structure of the neighborhood of the vector b; i.e. it is due to the embeddding rather than the method.

It would be too simplistic to say that no "semantic shifts" are ever happening at all. In some cases they do indeed occur. The point I would like to make here is that

3CosAdd cannot guarantee it, because it depends too much on the cosine similarity. It cannot guarantee that the difference between vectors a' and a would be always restricted to just one semantic feature, or that it would be even consistent between different pairs that are supposed to hold the same relations. It is clear that this is not the case if we look at the results for different combinations of word pairs in Table 4.6.

TABLE 4.6: Impact of source pairs in 3CosAdd method: Japanese intransitive/transitive verbs

No	a	a'	b	b'	score	$argmax_{\ b' \in V}$ $(cos(b', b - a + a'))$
1	集まる	集める	乗る	乗せる	0.41	乗せる
2	改まる	改める	乗る	乗せる	0.33	乗せる
3	受かる	受ける	乗る	乗せる	0.32	乗せる
4	落ちる	落す	乗る	乗せる	0.38	のる
5	寄る	寄せる	乗る	乗せる	0.4	のる
6	通る	通す	乗る	乗せる	0.33	のる
7	狭まる	狭める	乗る	乗せる	0.36	乗れる
8	始まる	始める	乗る	乗せる	0.34	乗れる
9	明ける	明かす	乗る	乗せる	0.31	乗れる
10	渡る	渡す	乗る	乗せる	0.37	手渡す
11	定まる	定める	乗る	乗せる	0.4	規定する

In all examples in Table 4.6 the "source" vectors a and a' are combined with the same b vector ($\overrightarrow{noru}$), which should result in $\overrightarrow{noseru}$; but different "source" vectors are producing different results. In lines 1-3 we are indeed getting the expected answer, but in lines 4-6 and 7-9 we are getting different wrong vectors – which are all however close to $\overrightarrow{noru}$ (in lines 4-6 we are in fact getting its alternative spellings with hiragana). Lines 10-11, on the other hand, exemplify a rarer case: the difference between a and a' retains enough of a' to make the result its own close neighbor, despite the addition of b. This shows clearly that the outcome of 3CosAdd is not reliably predicted by composition of semantic features supposedly captured by linear vector offset.

Mathematically, the unreliability of 3CosAdd could be attributed to three factors: (1) difference between a' and a vectors can retain more than the target semantic feature, (2) some b vectors have "denser" neighorhoods, increasing the chance for a mistake, and (3) some b vectors are more distant from b', which would also increase the chance for a mistake. There is evidence to back all these factors.

Some examples demonstrating the first factor can be found in Table 4.6. It is clear that subtracting word vectors that hold the same linguistic relation does not necessarily yield the same result – and this is to be expected, as linguistic relations are not necessarily reflected in the same way in all the contexts of all the different words that underlie their distributional representations. As the lines 10 and 11 show (together with data from window size 2 in Table 4.5), we cannot even expect the difference between a and a' to not express enough of their shared core semantics to completely "dissolve" the b vector. It is particularly the case for intransitivity/transitivity in

Japanese: syntactically these verb categories are mostly marked with particles *wo* and *ga*, while their head nouns can remain the same (cp. *doa ga shimaru* ("the door closes") - *doa wo shimeru* ("to close the door")). This means that the distributional representation of this linguistic category would be expressed with very few dimensions and be hard to discern from noise.

The second factor would be the easiest to see with data from a morphologically rich language. Consider the singular:plural noun category in English that in our data achieves almost 80% accuracy with the "classic" 3CosAdd (Figure 4.6). The similarly built SVD embeddings built from a large Russian corpus (13.4B tokens) showed less than 30% accuracy on the pairs of Russian nominative case forms contrasted with dative, instrumental and prepositional case in singular and plural (Drozd et al., 2016). If the embeddings tend to put morphological forms of the same word in the same neighborhoods, having more forms per word would make such neighborhoods denser, and 3CosAdd would have a higher chance to land on a wrong word form.

The third factor needs further investigation, but it could very well explain the overall low performance of 3CosAdd on derivational morphological categories in English (Figure 4.6). While morphological forms of the same word are likely to share many contexts and thus be distributionally similar (e.g. both *students* and *student* would be found in the close context of *school*), the same is not necessarily the case for suffixes that change parts of speech of their stems (e.g. *mad:madness*, *edit:editable*). Simply put, the words in such pairs are distributionally less similar to each other than inflected forms of the same word, and thus 3CosAdd would need to "reach" further from b in the distributional space. This would increase the chance of an error, especially with the variation in $(a' - a)$.

All of the above are empirical observations that indicate problems with 3CosAdd as a method for inducing semantic shifts by analogy. To conclude the discussion, I will also list some obvious theoretical concerns. Consider once again the famous example $\overrightarrow{king}$ - $\overrightarrow{man}$ + $\overrightarrow{woman}$ = $\overrightarrow{queen}$. Semantically this can be interpreted in two ways:

- ($\overrightarrow{king}$ - $\overrightarrow{man}$) + $\overrightarrow{woman}$ = $\overrightarrow{queen}$. The difference $\overrightarrow{king}$ - $\overrightarrow{man}$ must encode the "royalty" feature that can be added to $\overrightarrow{woman}$.
- $\overrightarrow{king}$ + ($\overrightarrow{woman}$ - $\overrightarrow{man}$) = $\overrightarrow{queen}$. The difference between $\overrightarrow{woman}$ - $\overrightarrow{man}$ must encode the "femaleness" feature that can be added to $\overrightarrow{king}$.

While vector arithmetic operations are commutative, it is not clear that so are semantic operations. Moreover, in this particular case both $\overrightarrow{woman}$ - $\overrightarrow{man}$ and $\overrightarrow{king}$ - $\overrightarrow{man}$ are interpretable; but this does not necessarily hold for other relations: $\overrightarrow{agaru}$ - $\overrightarrow{shimaru}$ + $\overrightarrow{shimeru}$ is only meaningful for $\overrightarrow{agaru}$ + ($\overrightarrow{shimeru}$ - $\overrightarrow{shimaru}$). At this point we do not know whether the postulated semantic features must actually make sense, and, if so – which of the possible interpretations is relevant.

The commutativity problem is a part of the bigger problem of mathematical interpretation of semantics, which also includes the problem with symmetry of similarity judgements – upon which analogy is based. Logically *a is like b* is equivalent to *b is like a*, but humans do not necessarily agree with both statements to the same degree. This problem is far from being resolved, although connectionist models have

proposed a number of ways to simulate asymmetry through biases, saliency features, or structural alignment (Thomas & Mareschal, 1997, p.758). There are also hybrid symbolic-distributed models such as LISA (Hummel & Holyoak, 2003). Viability of these solutions in a large-scale semantic model remains an empirical question.

There is also the problem of impossible meaning combinations. As far as vectors are concerned, we can subtract $\overrightarrow{scarf}$ from $\overrightarrow{palace}$ and add $\overrightarrow{kangaroo}$, and some vector will be the closest to the result of that calculation. But in doing so, can we say that we are still modeling natural language semantics? In other words, 3CosAdd points at the major challenge for DS in general: reflecting the distinction between named, unnamed and non-existing parts of the mental lexicon. As far as unnamed entities go, a promising approach is to think of word vectors as "points" in distributional semantic space we need to start thinking of them as regions (Erk, 2009; Vilnis & McCallum, 2015) or densities (Jameel & Schockaert, 2016). Such an approach would bring distributional semantic space closer to conceptual space, which can also be populated to different degrees and "chunked" by languages in different ways. However, it will remain to be seen whether such models are indeed more cognitively plausible and accurate, and whether it would be possible to mathematically account for nonsensical vector combinations via degree of "populatedness" of a region in vector space.

Thirdly, in so far as 3CosAdd approach postulates "semantic features" that can be identified by vector subtraction and then "added" to other word vectors, it is subject to all the standard objections to componential semantic analysis: semantic features are not easy to define clearly, they may be defined unnecessarily (or not defined where they are needed), they apply only to portions of vocabulary, it imposes binary oppositions that are psycholinguistically unrealistic, and they are locked within symbolic representation of language, with no way to reach out to the real-world entities (Leech, 1981, pp.117-119). While mainstream linguistic semantics has long moved past componential analysis, it retains its intuitive appeal to many computer scientists.

The fourth problem is that, in so far as vectors are derived from corpus data, they are subject to limitations of the corpora, particularly in what concerns missing data (to be discussed in more detail in section 6.2.2). Simply put, observing a phenomenon in a corpus proves its existence in the language, but its absence does not necessarily imply its non-acceptability. Thus, when we look for the distributional difference between $\overrightarrow{man}$ and $\overrightarrow{woman}$, we do not necessarily observe the full range of the relevant difference, which perhaps could have helped to bring the 3CosAdd calculation closer to the desired outcome. Furthermore, in condensed vectors such as SVD or in neural word embeddings the features are also blended in a non-transparent way, which could further obscure the selection of the relevant features. This objection applies to any distributional method, but pair-based methods are at further disadvantage compared to set-based methods such as LRCos since they have to rely on individual word pairs and are more subject to word idiosyncrasies.

All the observations and arguments above suggest that, despite the intuitive appeal of 3CosAdd, there are inherent limitations to what semantic operations can be modeled with simple arithmetic over word vectors. Thus I would argue in favor of the set-based methods and adopting more complex algorithms to improve over simple vector

similarity.

4.2.4.2 Enhancing machine learning with text patterns

To reiterate, LRCos is different from 3CosAdd in that it relies on supervised machine learning to obtain the representation of the target class of words, given a set of examples that are known to hold the target linguistic relation (in English the method saturated at about 30 examples (Drozd et al., 2016)). This step dramatically improves accuracy of the method by enabling more precise search only among the words that are likely to be in the correct class. To illustrate this point, let us compare the Top-5 results for LRCos and 3CosAdd for several queries with the best-performing model (lemmatized corpus, window size 3). Tables 4.7-4.8 list the 5 nearest neighbors of the hypothetical answer vector for 5 random verbs.

Table 4.7: Nearest neighbors of the hypothetical answer vector: 3CosAdd method, lemmatized corpus.

(1) 上がる- 集まる+ 集める	(2) 温まる- 収まる+ 収める	(3) 定まる- 上がる+ 上げる	(4) 掛かる- 汚れる+ 汚す	(5) 閉まる- 集まる+ 集める
0.45 あがる	0.30 温める	0.42 規定する	0.45 かかる	0.42 閉める
0.43 上げる	0.27 あたためる	0.41 規定	0.31 掛ける	0.32 あける
0.42 下がる	0.27 暖める	0.38 あがる	0.30 かける	0.30 ロックする
0.32 下げる	0.26 冷える	0.35 下がる	0.27 掛る	0.30 しまる
0.28 低い	0.24 冷やす	0.34 上げる	0.24 経つ	0.29 閉じる

Table 4.8: Nearest neighbors of the hypothetical answer vector: LRCos method, lemmatized corpus.

(1) 上がる	(2) 温まる	(3) 定まる	(4) 掛かる	(5) 閉まる
0.44 上げる	0.42 温める	0.23 きめる	0.37 掛ける	0.40 閉める
0.32 下げる	0.38 暖める	0.22 見定める	0.34 かける	0.36 あける
0.21 引き上げる	0.34 あたためる	0.21 しめす	0.23 かかる	0.24 開ける
0.21 高める	0.31 冷やす	0.21 明示する	0.21 節約する	0.24 ノックする
0.19 押し上げる	0.28 ほぐす	0.20 決める	0.18 費やす	0.23 締める

The output of both models in Tables 4.7-4.8 make it clear that 3CosAdd and LRCos are not doing the same thing. While 3CosAdd is biased towards the neighborhoods of the b vector, LRCos is remarkably morphosemantically coherent: note that all of the candidate answers are of the correct (transitive) class, and semantically close to the source verb. This suggests that LRCos method has more potential as a lexicographic tool.

The advantages of LRCos over 3CosAdd become even more clear when we turn to the non-lemmatized corpus; the output for the same verbs is shown in Tables 4.9-4.10. If the above concerns about 3CosAdd's bias towards close neighbors of b are grounded, then we would expect it to have more difficulty with data that has more

morphological forms: basically lemmatization step is decreasing the morphological diversity and the amount of vectors that are very similar to each other. But since lemmatization takes time, is error-prone, and good lemmatizers are unavailable for many languages, it would be desirable to have a method that can work with "raw" textual data.

This prediction is born out: Table 4.9 shows that 3CosAdd has difficulty finding verbs in the correct form (the dictionary form). However, LRCos method (Table 4.10) deals with this problem gracefully: all of the nearest neighbors of the hypothetical answer vector are verbs in dictionary form, and of the correct class. Furthermore, the semantic coherence of the retrieved neighborhoods is also preserved.

Table 4.9: Nearest neighbors of the hypothetical answer vector: 3CosAdd method, non-lemmatized corpus.

(1) 上がる- 集まる+ 集める	(2) 温まる- 収まる+ 収める	(3) 定まる- 上がる+ 上げる	(4) 掛かる- 汚れる+ 汚す	(5) 閉まる- 集まる+ 集める
0.45 上がっ	0.23 癒す	0.21 定義する	0.40 かかる	0.39 閉める
0.44 上がり	0.23 温める	0.21 挙げる	0.37 かかり	0.33 開ける
0.44 上げる	0.23 冷やす	0.20 しめす	0.35 掛かり	0.31 開けよう
0.41 下がる	0.22 リラックスする	0.20 傾ける	0.34 かかっ	0.30 閉め
0.41 あがる	0.21 冷やさ	0.20 とどめる	0.34 かから	0.29 閉まっ

Table 4.10: Nearest neighbors of the hypothetical answer vector: LRCos method, non-lemmatized corpus.

(1) 上がる	(2) 温まる	(3) 定まる	(4) 掛かる	(5) 閉まる
0.49 上げる	0.28 温める	0.19 測定する	0.25 掛ける	0.26 閉める
0.35 下げる	0.19 壊す	0.19 推定する	0.24 かける	0.25 開ける
0.26 上げ	0.19 冷やす	0.19 とどめる	0.24 節約する	0.25 あける
0.24 引き下げる	0.18 ほぐす	0.18 定義する	0.20 費やす	0.19 下ろす
0.24 高める	0.17 癒す	0.18 判定する	0.17 貯める	0.17 ノックする

These results suggest that the classifier included in LRCos does have potential for improving performance on non-lemmatized corpora, and the method could be developed further for this purpose. However, while it outperforms 3CosAdd method by a large margin, it still achieves only 60% accuracy on the current dataset, and to be practically useful in lexicographic applications it would need to be further improved.

There are several directions which could yield further enhancements, including experiments with various word embedding models and their parameters, and also with more sophisticated machine learning algorithms. But from a practical point of view there is an easier, and readily available solution: combining the knowledge about linguistic relations that is derived from word distributions with clues that can be derived from their surface forms.

In the case of morphologically related groups of words, such as Japanese intransitive/transitive verbs, we already have an additional source of information. For

instance, in solving the problem $\overrightarrow{agaru}$ - $\overrightarrow{atatamaru}$ + $\overrightarrow{atatameru}$ = ? ($\overrightarrow{ageru}$, we already know that the answer should start with the same Chinese character as the first word in its pair ($\overrightarrow{agaru}$), and all the information we need is already contained in the test data. The beginning and ending sequences of characters in words have been demonstrated to be sufficient for building a fairly detailed morphology of a language (Soricut & Och, 2015).

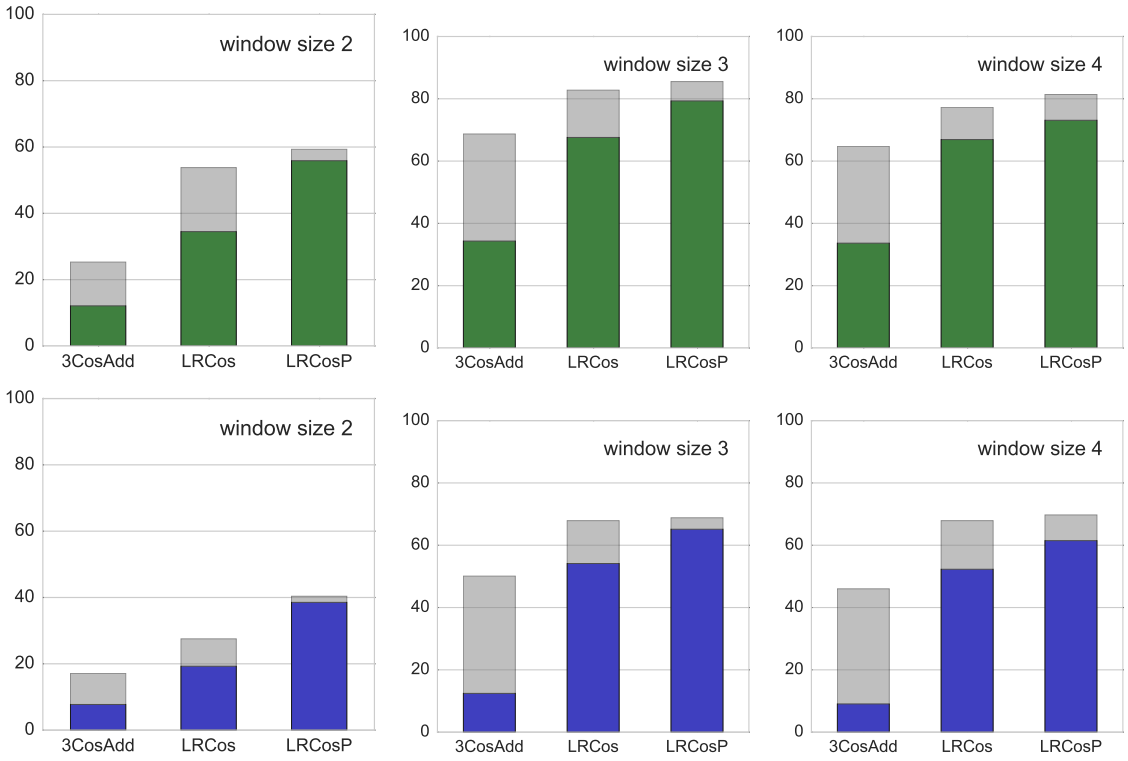
To explore this possibility, I implemented LRCosP, a simple modification of LRCos algorithm (Figure 4.7). Upon obtaining 5 nearest neighbors of the hypothetical answer vector according to (4.7), I check whether any of them shares the first character with the source word. If so, the score of this answer is multiplied by 2, which makes it more likely to be output as the candidate answer.

This procedure could be further refined (e.g. to avoid words that contain Chinese characters not contained in the source word, and/or to require that the answer ends in one of the endings of Japanese verbs in their dictionary forms, such as *-ru* or *-u*). But even in this crude form LRCosP considerably improves over LRCos. Figure 4.7 shows that in several conditions the Top-1 performance of LRCosP reaches the Top-5 performance of LRCos, achieving roughly 80% Top-1 accuracy with lemmatized corpus (window size 3) in both transitive-intransitive and intransitive-transitive settings.

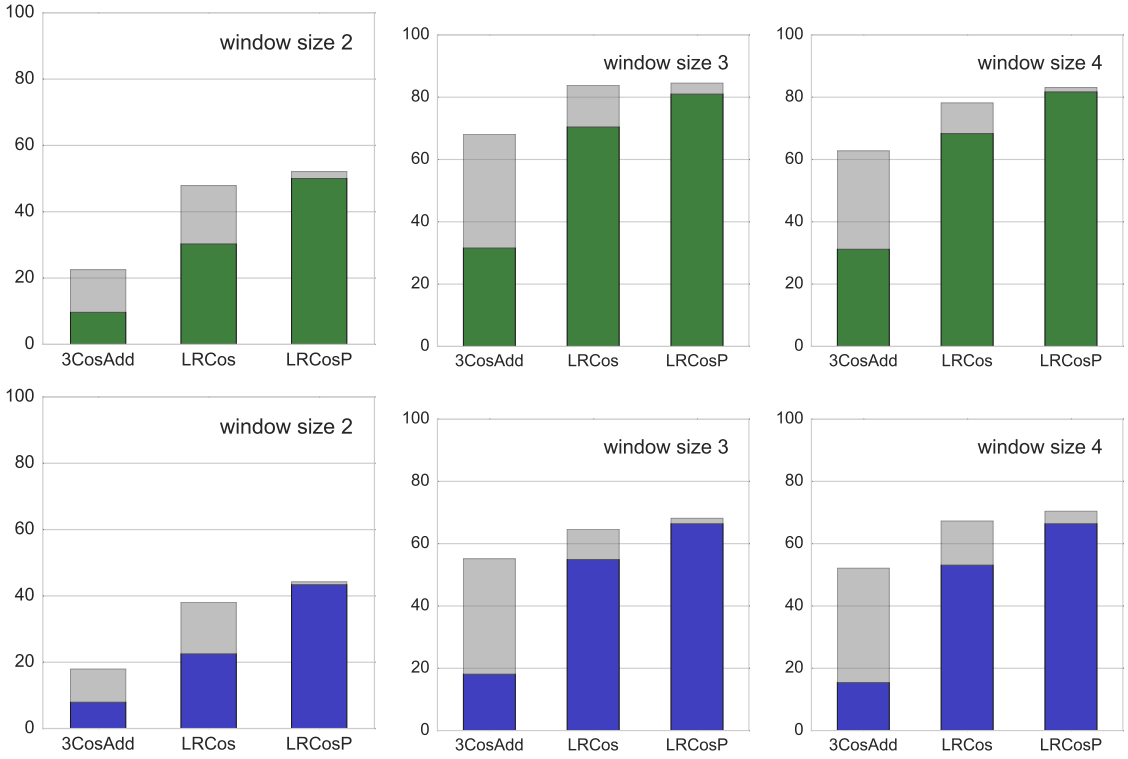
Combining pattern-based search with distributional information could be seen as "cheating". However, it is perhaps not fair to DS to expect that it would be able to accurately derive any linguistic paradigms merely from distributions: humans have many more sources of information about word usage than their co-occurrences, and they certainly are aware of similarities in the word form, which does not play any role in constructing a word-level embedding with SVD. There are also other, morphology-aware approaches to building word embeddings such as character-based embeddings.

Thus the current study brings up several important questions for the field in general. What is the place of morphology word-level distributional semantic models? Ideally, do we want a combination of character/word-level models? of character-level, word-level, and multimodal models? If so, what roles should they all play in modeling human linguistic knowledge?

This case study described a way of using DS methodology to solve a practical lexicographic problem (automatic discovery of verbs with certain semantic features relevant for event structure). But it also showed that in-depth investigation of a linguistic phenomenon can yield useful insights for NLP. It is not only FS that could benefit from the tools offered by DS, but also DS needs the tasks that FS could set, in order to make progress in its own goals. The biggest challenge in this field is the interpretation of distributed representations, as highly-dimensional space is impossible to visualize or interpret directly. It is through discovery of linguistic properties that align with the mathematical properties of the vector space that such interpretation can be made. In this particular case, a practical question that arose from linguistic inquiry highlighted major issues with pair-based methods of solving word analogies, posing several methodological questions for the field in general.



(A) Transitive-intransitive verbs



(B) Intransitive-transitive verbs

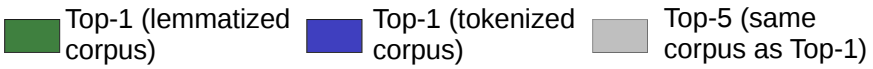


FIGURE 4.7: Accuracy of solving word analogies with Japanese intransitive/transitive verbs: 3CosAdd, LRCos and LRCosP

4.3 Summary

This chapter discussed one of the practical applications of distributional semantics to the needs of frame-based computational lexicography – the automatic discovery of semantico-morphological classes. This is essential not only for building a frame database, but also for ensuring consistency of frame-to-frame relations, which in the case of aspectual classes have to be established for large groups of words.

While current research on automatic induction of aspectual classes of verbs is based on syntactically parsed corpora, the first case study showed that such classification is also possible on the basis of raw text - even in a morphologically rich language like Russian that presents additional difficulties for distributional methods. Based on a relatively small dataset with a 100 "seed" words in each category, we were able to classify perfective and imperfective verbs with over 90% accuracy in the confidence interval 0.9-1.0.

Another lexicographic advantage of the proposed approach is that with a suitable corpus (which in our experiments turned out to be fiction) it is possible to discover a lot of newly derived words that are not yet attested in dictionaries, while simultaneously obtaining their aspectual classification.

The second case study focused on automatic discovery of pairs of words that have the same relation, namely Japanese intransitive/transitive verbs. I experimented with 3 methods of finding the missing members of pairs of such verbs, one of which was further developed to achieve 80% accuracy on my dataset. Thus the case study 2 also confirms the high potential of DS methods for lexicographic purposes.

Importantly, the case study 2 highlighted several problems with pair-based methods of solving analogies with word embeddings – their reliance on the target words being close neighbors of the source words, their inconsistencies and necessity of deeper linguistic grounding. This example shows that, while FS could obviously benefit from the practical achievements of DS, such collaboration would also be useful for DS in enabling more systematic explorations of the structure of distributed meaning representations.

This chapter discussed two ways to discover aspectual verb classes with DS tools, and showed two successful case studies in Russian in Japanese. However, no claims are made with regards to these mechanisms being applicable to solving any other semantic tasks (or even that the proposed mechanisms cannot be further improved). The very volume of current NLP literature suggests that different linguistic relations call for different solutions, and my own work on interaction between linguistic relations of different types, different word embeddings and different methods of solving analogies illustrates the same point. Nevertheless, it is hard to think of a linguistic problem to which NLP has not already offered several solutions.

Chapter 5 "Selectional preferences across languages", Chapter 6 "Classifying selectional preferences", Chapter 7 "Discussion: towards a unified semantics", and Chapter 8 "Conclusion" are not included in the abridged thesis due to planned publication elsewhere.

Appendix A

Dataset: Russian imperfective and perfective verbs

This appendix contains the lists of words used in case study 1 described in section 4.1. The procedure for their selection is described in section 4.1.2.3.

Due to the large size of the full dataset (3995 perfective verbs; 4890 imperfective verbs; 4751 "other" words), the lists below include only the lemmas of all words (up to 100 per morphological class). These lists were automatically expanded with word forms from a Russian morphological dictionary (Hagen, 2014). The present participle forms (e.g. *делать* ("do") - *делающий* ("doing")) and transgressive forms (e.g. *делать* ("do") - *сделав* ("having done")) are included in the verb datasets and obtained automatically from the morphological dictionary. Any duplicate forms were merged.

Imperfective verbs

быть, мочь, говорить, знать, хотеть, идти, иметь, видеть, думать, жить, делать, смотреть, работать, дать, понимать, сидеть, являться, любить, стоить, считать, казаться, писать, стоять, давать, помнить, ждать, находиться, оставаться, играть, лежать, следовать, читать, бывать, ходить, начинать, называть, хотеться, вести, бояться, происходить, существовать, становиться, слышать, использовать, пытаться, чувствовать, заниматься, продолжать, слушать, отвечать, рассказывать, представлять, брать, спать, помочь, приходить, просить, спрашивать, принимать, искать, произойти, выходить, пить, глядеть, ехать, начаться, верить, держать, позволять, уходить, собираться, относиться, требовать, получать, проходить, составлять, приходиться, стараться, уметь, проводить, нравиться, поехать, положить, входить, оказываться, показывать, касаться, действовать, кричать, предлагать, молчать, бежать, петь, вызывать, показаться, выглядеть, состоять, выступать, ставить, возникать

Perfective verbs

сказать, стать, сделать, понять, пойти, спросить, получить, оказаться, взять, прийти, остаться, выйти, начать, увидеть, найти, решить, пройти, принять, написать, подумать, уйти, посмотреть, вернуться, появиться, показать, поставить, смочь, заметить, представить, создать, узнать, приехать, рассказать, забыть, провести, вспомнить, открыть, привести, оставить, войти, назвать, успеть, предложить, подойти, удаться, умереть, сесть, случиться, установить, поднять, встать, бросить, объяснить, снять, услышать, связать, позволить, вызвать, отметить, заявить,

получиться, остановиться, убить, согласиться, сообщить, принести, служить, попросить, отдать, подняться, возникнуть, позвонить, выпить, обнаружить, почувствовать, передать, составить, определить, улыбнуться, уехать, родиться, отказаться, занять, добавить, произнести, направить, обратиться, выбрать, упасть, перестать, собрать, простить, перейти, исчезнуть, объявить, пригласить, закрыть, придумать, попробовать, достигнуть

"Other" words

- **Nouns:** год, человек, время, дело, жизнь, день, рука, раз, работа, слово, место, лицо, друг, глаз, вопрос, дом, сторона, страна, мир, случай, голова, ребенок, сила, конец, вид, система, часть, город, отношение, женщина, деньги, земля, машина, вода, отец, проблема, час, право, нога, решение, дверь, образ, история, власть, закон, война, бог, голос, тысяча, книга, возможность, результат, ночь, стол, имя, область, статья, число, компания, народ, жена, группа, развитие, процесс, суд, условие, средство, начало, свет, пора, путь, душа, уровень, форма, связь, минута, улица, вечер, качество, мысль, дорога, мать, действие, месяц, государство, язык, любовь, взгляд, мама, век, школа, цель, общество, деятельность, организация, президент, комната, порядок, момент, театр
- **Adjectives:** новый, большой, должен, последний, российский, русский, общий, высокий, хороший, главный, лучший, маленький, молодой, государственный, полный, советский, настоящий, старый, разный, нужный, белый, собственный, черный, основной, далекий, подобный, следующий, равный, живой, известный, военный, важный, великий, простой, огромный, политический, московский, готовый, красный, современный, социальный, ранний, особый, целый, плохой, сильный, скорый, крупный, внутренний, экономический, правый, федеральный, близкий, похожий, различный, необходимый, единственный, легкий, человеческий, международный, дорогой, небольшой, местный, бывший, американский, широкий, мировой, тяжелый, возможный, отдельный, средний, красивый, короткий, серьезный, интересный, добрый, национальный, длинный, страшный, прошлый, общественный, детский, единый, определенный, низкий, чужой, странный, чистый, поздний, специальный, научный, сложный, реальный, способный, малый, старший, личный, свободный, обычный, прекрасный
- **Adverbs:** еще, уже, очень, можно, надо, нет, тоже, более, конечно, также, вдруг, почти, сразу, хорошо, сегодня, совсем, вообще, больше, вместе, например, нужно, опять, снова, нельзя, особенно, рядом, назад, совершенно, значит, давно, действительно, наконец, часто, быстро, долго, правда, иногда, чуть, затем, слишком, вполне, далее, может, впрочем, наверное, пока, достаточно, менее, кстати, сначала, довольно, однажды, домой, скоро, наиболее, обычно, далеко, трудно, возможно, точно, весьма, легко, впервые, видно, немного, практически, необходимо, вовсе, рано, несмотря, сильно, кажется, известно, дома, завтра, видимо, мало, одновременно, тихо, недавно, вновь, вперед, вчера, полностью, плохо, постоянно, едва, ясно, обязательно, прямо, медленно, спокойно, вскоре, невозможно, примерно, гораздо, уж, неожиданно, наоборот, естественно
- **Pronouns:** так, как, где, там, потом, сейчас, тут, теперь, тогда, здесь, потому, всегда, почему, все, поэтому, никогда, куда, как-то, зачем, туда, откуда, сюда, столь, никак, когда, почему-то, где-то, иначе, что, когда-то, отсюда, оттуда, навсегда, нечего, куда-то, столько, никуда, чего, сколько, везде, оттого, отчего, когда-нибудь, нигде, где-нибудь, по-другому, всюду, по-своему, некогда, как-нибудь, откуда-то, ничуть, повсюду, вон, куда-нибудь, нисколько, тут-то, ничего, сколь, кое-как, так-то, зачем-то, много, кое-где, туда-сюда, отчего-то, что-то, сколько-нибудь, тогда-то, посему, этак, как-никак, когда-либо, отовсюду, почем, доселе, эдак, че, по-иному, всего, чего-то, сколько-то, по-нашему, потому-то, ниоткуда, по-моему, по-вашему, столько-то, там-то, , этот, который, свой, весь, тот, такой, его, один, сам, другой, наш, мой, ее, какой, их, самый, каждый, какой-то, ваш, некоторый, любой, никакой, всякий, твой, иной, многий, данный, сей, какой-нибудь, некий, остальной, прочий, чей, какой-либо, таков, каков, таковой, кой, чей-то, такой-то, этакий, кое-какой, немногий, тот-то, ихний, эдакий, каковой, оный, ничей, какой-никакой,

экий, чей-нибудь, , я, он, это, она, они, мы, то, что, ты, все, вы, себя, кто, ничто, что-то, никто, оно, кто-то, что-нибудь, многое, прочее, нечто, многие, кто-нибудь, другое, остальное, остальные, кое-что, некоторые, свое, что-либо, се, кое-кто, кто-либо, некто, че, немногие, немногое, шо, сие, то-то, всякое, што

- **Numerals:** два, три, пять, оба, четыре, десять, двадцать, шесть, сто, тридцать, семь, двое, сорок, восемь, пятьдесят, полтора, трое, пятнадцать, девять, двенадцать, двести, шестьдесят, триста, семьдесят, одиннадцать, четверо, восемьдесят, восемнадцать, девяносто, пятьсот, шестнадцать, семнадцать, четырнадцать, тринадцать, пятеро, четыреста, девятнадцать, девятьсот, шестеро, восемьсот, семеро, шестьсот, семьсот, первый, второй, один, третий, четвертый, пятый, шестой, седьмой, восьмой, девятый, десятый, двадцатый, тридцатый, двенадцатый, одиннадцатый, семнадцатый, шестидесятый, семидесятый, девятнадцатый, пятидесятый, сороковой, тринадцатый, восемнадцатый, девяностый, пятнадцатый, восьмидесятый, шестнадцатый, четырнадцатый, сотый

- **Prepositions:** в, на, с, по, к, из, у, за, от, о, для, до, при, во, со, под, после, без, через, об, перед, между, над, про, кроме, среди, из-за, ко, против, около, вместо, вокруг, прежде, возле, сквозь, ради, благодаря, согласно, из-под, вроде, спустя, мимо, вдоль, помимо, внутри, обо, насчет, вне, включая, путем, относительно, изо, посреди, подобно, накануне, вопреки, передо, свыше, вследствие, напротив, меж, посредством, поверх, ввиду, впереди, вблизи, надо, позади, выше, близ, безо, внутрь, поперек, сверх, пред, наподобие, ото, ниже, подле, вслед, взамен, посередине, средь, подо, превыше, сродни, исключая, сзади, минус

- **Conjunctions:** и, что, но, как, или, если, когда, чтобы, то, да, чем, хотя, ни, однако, пока, даже, ведь, поскольку, словно, либо, причем, будто, чтоб, зато, ибо, только, ли, хоть, же, раз, итак, иначе, точно, пусть, плюс, нежели, тем, дабы, коли, едва, ежели, притом, коль, благо, покуда, ан, иль, дак, кабы, яко, все-таки, ровно, соответственно, лишь

- **Interjections:** ах, о, господи, ой, увы, ага, ох, эх, угу, эй, а, э, а-а, ну-ка, ура, мм, ай, э-э, тьфу, блин, ей-богу, алло, здорово, ух, ха-ха-ха, стоп, ха, поди, ого, ишь, ну-ну, а-а-а, боже, хм, фу, ха-ха, м-м, гм, браво, здравствуйте, хо, м, здрасте, оп, о-о, буль-буль, пардон, марш, м-да, э-э-э, але

- **Particles:** не, же, только, бы, вот, даже, ну, ли, ни, да, просто, нет, лишь, именно, ведь, это, уж, все-таки, пусть, хоть, ж, всего, будто, разве, прямо, вроде, лучше, ладно, то, давай, мол, пожалуйста, хотя, еще, неужели, б, вон, давайте, хорошо, там, спасибо, якобы, де, себе, вообще-то, словно, угодно, точно, исключительно, аж, было, конечно, пускай, да-да, небось, таки, фон, как, нет-нет, неужто, ль, то-то, ван, все, во, авось, бишь, невесть, а, нибудь, так, ка, кое, ишь, дак, аминь, да-а, ага, бен, с, не-ет, вишь, на, ди, ровно, нате, прям, се, пока, ла, неизвестно, что, ну-с

Appendix B

Japanese intransitive/transitive verbs used in case study 2

上がる(あがる)　上げる(あげる)
温まる(あたたまる)　温める(あたためる)
集まる(あつまる)　集める(あつめる)
改まる(あらたまる)　改める(あらためる)
受かる(うかる)　受ける(うける)
埋まる(うまる)　埋める(うめる)
植わる(うわる)　植える(うえる)
収まる(おさまる)　収める(おさめる)
終わる(おわる)　終える(おえる)
掛かる(かかる)　掛ける(かける)
重なる(かさなる)　重ねる(かさねる)
固まる(かたまる)　固める(かためる)
被さる(かぶさる)　被せる(かぶせる)
絡まる(からまる)　絡める(からめる)
変わる(かわる)　変える(かえる)
決まる(きまる)　決める(きめる)
極まる(きわまる)　極める(きわめる)
加わる(くわわる)　加える(くわえる)
下がる(さがる)　下げる(さげる)
定まる(さだまる)　定める(さだめる)
静まる(しずまる)　静める(しずめる)
閉まる・締まる(しまる)　閉める・締める(しめる)
窄まる(すぼまる)　窄める(すぼめる)
据わる(すわる)　据える(すえる)
狭まる(せばまる)　狭める(せばめる)
染まる(そまる)　染める(そめる)
高まる(たかまる)　高める(たかめる)
助かる(たすかる)　助ける(たすける)
溜まる(たまる)　溜める(ためる)
縮まる(ちぢまる)　縮める(ちぢめる)
繋がる(つながる)　繋げる(つなげる)
詰まる(つまる)　詰める(つめる)
強まる(つよまる)　強める(つよめる)
連なる(つらなる)　連ねる(つらねる)
留まる(とどまる)　留める(とどめる)
止まる(とまる)　止める(とめる)
始まる(はじまる)　始める(はじめる)
広まる(ひろまる)　広める(ひろめる)
深まる(ふかまる)　深める(ふかめる)
曲がる(まがる)　曲げる(まげる)
混ざる(まざる)　混ぜる(まぜる)
纏まる(まとまる)　纏める(まとめる)
茹だる(ゆだる)　茹でる(ゆでる)
弱まる(よわまる)　弱める(よわめる)
赤らむ(あからむ)　赤らめる(あからめる)
開く(あく)　開ける(あける)
入る(はいる)　入れる(いれる)
浮かぶ(うかぶ)　浮かべる(うかべる)
傾く(かたむく)　傾ける(かたむける)
叶う(かなう)　叶える(かなえる)
絡む(からむ)　絡める(からめる)
苦しむ(くるしむ)　苦しめる(くるしめる)
沈む(しずむ)　沈める(しずめる)
従う(したがう)　従える(したがえる)

退く(しりぞく)　退ける(しりぞける)
竦む(すくむ)　竦める(すくめる)
進む(すすむ)　進める(すすめる)
育つ(そだつ)　育てる(そだてる)
揃う(そろう)　揃える(そろえる)
立つ(たつ)　立てる(たてる)
違う(ちがう)　違える(ちがえる)
縮む(ちぢむ)　縮める(ちぢめる)
付く(つく)　付ける(つける)
届く(とどく)　届ける(とどける)
向く(むく)　向ける(むける)
止む(やむ)　止める(やめる)
歪む(ゆがむ)　歪める(ゆがめる)
緩む(ゆるむ)　緩める(ゆるめる)
折れる(おれる)　折る(おる)
切れる(きれる)　切る(きる)
砕ける(くだける)　砕く(くだく)
裂ける(さける)　裂く(さく)
溶ける・解ける(とける)　溶く・解く(とく)
取れる(とれる)　取る(とる)
抜ける(ぬける)　抜く(ぬく)
捩れる(ねじれる)　捩じる(ねじる)
剥げる(はげる)　剥ぐ(はぐ)
弾ける(はじける)　弾く(はじく)
剥ける(むける)　剥く(むく)
焼ける(やける)　焼く(やく)
破れる(やぶれる)　破る(やぶる)
割れる(われる)　割る(わる)
明ける(あける)　明かす(あかす)
枯れる(かれる)　枯らす(からす)
焦げる(こげる)　焦がす(こがす)
冷める・醒める(さめる)　冷ます・醒ます(さます)
出る(でる)　出す(だす)
溶ける(とける)　溶かす(とかす)
慣れる(なれる)　慣らす(ならす)
逃げる(にげる)　逃がす(にがす)
漏れる(もれる)　漏らす(もらす)
揺れる(ゆれる)　揺らす(ゆらす)
癒える(いえる)　癒す(いやす)
肥える(こえる)　肥やす(こやす)
費える(ついえる)　費やす(ついやす)
生える(はえる)　生やす(はやす)
冷える(ひえる)　冷す(ひやす)
増える(ふえる)　増やす(ふやす)
燃える(もえる)　燃やす(もやす)
閉じる(とじる)　閉ざす(とざす)
伸びる(のびる)　伸ばす(のばす)
満ちる(みちる)　満たす(みたす)
起きる(おきる)　起こす(おこす)
落ちる(おちる)　落す(おとす)
降りる(おりる)　降ろす(おろす)
過ぎる(すぎる)　過ごす(すごす)
尽きる(つきる)　尽くす(つくす)
現れる(あらわれる)　現す(あらわす)
隠れる(かくれる)　隠す(かくす)
汚れる(けがれる)　汚す(けがす)
零れる(こぼれる)　零す(こぼす)
壊れる(こわれる)　壊す(こわす)
倒れる(たおれる)　倒す(たおす)
潰れる(つぶれる)　潰す(つぶす)
流れる(ながれる)　流す(ながす)
逃れる(のがれる)　逃す(のがす)
離れる(はなれる)　離す(はなす)
汚れる(よごれる)　汚す(よごす)
乗る(のる)　乗せる(のせる)
寄る(よる)　寄せる(よせる)
分かれる(わかれる)　分ける(わける)
消える(きえる)　消す(けす)
捕まる(つかまる)　捕まえる(つかまえる)
見える(みえる)　見る(みる)
揺れる(ゆれる)　揺する(ゆする)
篭る(こもる)　篭める(こめる)
聞える(きこえる)　聞く(きく)
移る(うつる)　移す(うつす)
返る・帰る(かえる)　返す・帰す(かえす)
下る(くだる)　下す(くだす)

転がる(ころがる)　転がす(ころがす)
通る(とおる)　通す(とおす)
灯る(ともる)　灯す(ともす)
治る(なおる)　治す(なおす)
残る(のこる)　残す(のこす)
回る(まわる)　回す(まわす)
戻る(もどる)　戻す(もどす)
渡る(わたる)　渡す(わたす)
動く(うごく)　動かす(うごかす)
驚く(おどろく)　驚かす(おどろかす)
乾く(かわく)　乾かす(かわかす)
散る(ちる)　散らす(ちらす)
照る(てる)　照らす(てらす)
退く(どく)　退かす(どかす)
轟く(とどろく)　轟かす(とどろかす)
飛ぶ(とぶ)　飛ばす(とばす)
泣く(なく)　泣かす(なかす)
鳴る(なる)　鳴らす(ならす)
減る(へる)　減らす(へらす)
惑う(まどう)　惑わす(まどわす)
迷う(まよう)　迷わす(まよわす)
刺さる(ささる)　刺す(さす)
挟まる(はさまる)　挟む(はさむ)
塞がる(ふさがる)　塞ぐ(ふさぐ)
跨がる(またがる)　跨ぐ(またぐ)
滅ぶ(ほろぶ)　滅ぼす(ほろぼす)
潤う(うるおう)　潤す(うるおす)
無くなる(なくなる)　無くす(なくす)
積もる(つもる)　積む(つむ)

Appendices C, D,E are not included in the abridged thesis due to planned publication elsewhere.

Bibliography

Aharon, R. B., Szpektor, I., & Dagan, I. (2010). Generating entailment rules from FrameNet. (pp. 241–246). Proceedings of the acl 2010 conference short papers. Association for Computational Linguistics.

Ajjanagadde, V. (1994). Unclear distinctions lead to unnecessary shortcomings: examining the rule vs fact, role vs filler, and type vs predicate distinctions from a connectionist representation and reasoning perspective. In *Proceedings of the AAAI* (Vol. 94).

Ameka, F. K. & Levinson, S. C. (2007). Introduction: The typology and semantics of locative predicates: posturals, positionals, and other beasts. *Linguistics*, *45(5)*, 847–871. doi:10.1515/LING.2007.025

Annesi, P. & Basili, R. (2010). Cross-lingual alignment of FrameNet annotations through Hidden Markov Models. In *Computational Linguistics and Intelligent Text Processing* (pp. 12–25). Springer.

Apresjan, J., Boguslavsky, I., Iomdin, B., Iomdin, L., Sannikov, A., & Sizov, V. (2006). A syntactically and semantically tagged corpus of Russian: state of the art and prospects. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)* (pp. 1378–1381). Genoa, Italy: ELRA.

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley Framenet project. In *Proceedings of the 17th international conference on Computational Linguistics* (Vol. 1, pp. 86–90). Association for Computational Linguistics.

Baroni, M., Bernardi, R., & Zamparelli, R. (2014). Frege in space: A program of compositional distributional semantics. *Linguistic Issues in Language Technology*, *9*.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, *43*(3), 209–226. doi:10.1007/s10579-009-9081-4

Baroni, M. & Lenci, A. (2011). How We BLESSed Distributional Semantic Evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics* (pp. 1–10). GEMS '11. Stroudsburg, PA, USA: Association for Computational Linguistics.

Barsalou, L. W. (1983). Ad hoc categories. *Memory and cognition*, *11*(3), 211–227.

Barsalou, L. W. (2010, October). Grounded Cognition: Past, Present, and Future: Topics in Cognitive Science. *Topics in Cognitive Science*, *2*(4), 716–724. doi:10.1111/j.1756-8765.2010.01115.x

Basili, R., De Cao, D., Croce, D., Coppola, B., & Moschitti, A. (2009). Cross-language frame semantics transfer in bilingual corpora. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 332–345). Springer: Berlin, Heidelberg.

Benko, V. (2014). Aranea: Yet Another Family of (Comparable) Web Corpora. In P. Sojka, A. Horák, I. Kopeček, K. Pala, & Masarykova Univerzita (Eds.), *Text, speech, and dialogue: 17th international conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings* (8655, pp. 257–264). Lecture notes in computer science Lecture notes in artificial intelligence. Cham: Springer International Publishing Switzerland.

Bergsma, S. & Goebel, R. (2011). Using Visual Information to Predict Lexical Preference. In *RANLP* (pp. 399–405).

Bergsma, S., Lin, D., & Goebel, R. (2008). Discriminative learning of selectional preference from unlabeled text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 59–68). Association for Computational Linguistics.

Bethard, S., Derczynski, L., Pustejovsky, J., & Verhagen, M. (2015). Semeval-2015 task 6: Clinical tempeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Association for Computational Linguistics.*

Bicknell, K. & Dodge, E. (2004, December). *Image Schemas and Force-Dynamics in FrameNet* (No. TR-04-00X). International COmputer Science Institute.

Blacoe, W. & Lapata, M. (2012). A Comparison of Vector-based Representations for Semantic Composition. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 546–556). EMNLP-CoNLL '12. Stroudsburg, PA, USA: Association for Computational Linguistics.

Boas, H. C. (2009). Semantic frames as interlingual representations for multilingual lexical databases. In H. C. Boas (Ed.), *Multilingual FrameNets in computational lexicography: Methods and applications* (Vol. 200, pp. 59–100). Trends in linguistics. Studies and monographs. De Gruyter Mouton.

Boleda, G. & Erk, K. (2015). Distributional Semantic Features as Semantic Primitives - Or Not. In *2015 AAAI Spring Symposium Series.*

Brachman, R. J., McGuinness, D. L., Patel-Schneider, P. F., Resnick, L. A., & Borgida, A. (1991). Living with CLASSIC: When and how to use a KL-ONE-like language. *Principles of semantic networks*, *401456.*

Bruni, E., Tran, N.-K., & Baroni, M. (2014). Multimodal Distributional Semantics. *J. Artif. Intell. Res.(JAIR)*, *49*, 1–47.

Bruni, E., Uijlings, J., Baroni, M., & Sebe, N. (2012). Distributional semantics with eyes: using image analysis to improve computational representations of word meaning. In *Proceedings of the 20th ACM international conference on multimedia - MM'12* (pp. 1219–1228). doi:10.1145/2393347.2396422

Bullinaria, J. A. & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, *39*(3), 510–526.

Bullinaria, J. A. & Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. *Behavior research methods*, *44*(3), 890–907.

Burchardt, A., Erk, K., & Frank, A. (2005). A WordNet detour to FrameNet. *Sprachtechnologie, mobile Kommunikation und linguistische Resourcen*, *8*, 408–421.

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Padó, S., & Pinkal, M. (2006). The SALSA corpus: a German corpus resource for lexical semantics. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*.

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Padó, S., & Pinkal, M. (2009). Using FrameNet for the semantic analysis of German: annotation, representation, and automation. In H. C. Boas (Ed.), *Multilingual FrameNets in computational lexicography: methods and applications* (pp. 209–244). Berlin/New York: Mouton de Gruyter.

Calvo, H. & Gelbukh, A. (2004). Unsupervised Learning of Ontology-Linked Selectional Preferences. In A. Sanfeliu, J. F. M. Trinidad, & J. A. C. Ochoa (Eds.), *Progress in Pattern Recognition, Image Analysis and Applications* (3287, pp. 418–424). Lecture Notes in Computer Science. Springer: Berlin, Heidelberg. doi:10.1007/978-3-540-30463-0_52

Chambers, N. & Jurafsky, D. (2009). Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (Vol. 2, pp. 602–610). Association for Computational Linguistics.

Chambers, N. & Jurafsky, D. (2010). Improving the use of pseudo-words for evaluating selectional preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 445–453). Association for Computational Linguistics.

Chang, N., Gildea, D., & Narayanan, S. (1998). A dynamic model of aspectual composition. In *Proceedings of the 20th Annual Meeting of the Cognitive Science Society, Madison, Wisconsin. Mahwah, NJ: Lawrence Erlbaum Associates* (pp. 226–31).

Chang, N., Narayanan, S., & Petruck, M. R. (2002). Putting frames in perspective. In *Proceedings of the 19th international conference on Computational linguistics* (Vol. 1, pp. 1–7). Association for Computational Linguistics. doi:10.3115/1072228.1072384

Chen, B. & Fung, P. (2004). Automatic construction of an English-Chinese bilingual FrameNet. In *Proceedings of HLT-NAACL 2004: Short Papers* (pp. 29–32). Association for Computational Linguistics.

Chen, Y.-N., Wang, W. Y., & Rudnicky, A. I. (2014). Leveraging frame semantics and distributional semantics for unsupervised semantic slot induction in spoken dialogue systems. In *Spoken Language Technology Workshop (SLT), 2014 IEEE* (pp. 584–589). IEEE.

Chklovski, T. & Pantel, P. (2004). VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Vol. 4, pp. 33–40).

Church, K. W. & Hanks, P. (1990, March). Word Association Norms, Mutual Information, and Lexicography. *Comput. Linguist. 16*(1), 22–29.

Clarke, D. (2012, March 1). A context-theoretic framework for compositionality in distributional semantics. *Computational Linguistics*, *38*(1), 41–71. doi:10.1162/COLI_a_00084

Coppola, B., Gangemi, A., Gliozzo, A., Picca, D., & Presutti, V. (2009). Frame detection over the semantic web. In *European Semantic Web Conference* (pp. 126–142). Springer.

Croft, W. (1990). Possible verbs and the structure of events. In S. L. Tsochatzidēs (Ed.), *Meanings and prototypes: Studies in linguistic categorization* (pp. 48–73). London/New York: Routledge.

Croft, W. (2012). *Verbs: aspect and causal structure*. Oxford linguistics. Oxford [England]; New York: Oxford University Press.

Dagan, I., Lee, L., & Pereira, F. C. (1999). Similarity-based models of word cooccurrence probabilities. *Machine Learning*, *34*, 43–69.

Das, D., Schneider, N., Chen, D., & Smith, N. A. (2010). SEMAFOR 1.0: A probabilistic frame-semantic parser. *Language Technologies Institute, School of Computer Science, Carnegie Mellon University*.

Dodge, E. (2010). *Constructional and Conceptual Composition* (Ph.D. Dissertation, UC Berkeley, Department of Linguistics).

Dowty, D. R. (1986). The effects of aspectual class on the temporal structure of discourse: semantics or pragmatics? *Linguistics and philosophy*, *9*(1), 37–61.

Drozd, A., Gladkova, A., & Matsuoka, S. (2015a). Discovering Aspectual Classes of Russian Verbs in Untagged Large Corpora. In *Proceedings of 2015 IEEE International Conference on Data Science and Data Intensive Systems (DSDIS)* (pp. 61–68). doi:10.1109/DSDIS.2015.30

Drozd, A., Gladkova, A., & Matsuoka, S. (2015b). Python, Performance, and Natural Language Processing. In *Proceedings of the 5th Workshop on Python for High-Performance and Scientific Computing* (1:1–1:10). PyHPC '15. New York, NY, USA: ACM. doi:10.1145/2835857.2835858

Drozd, A., Gladkova, A., & Matsuoka, S. (2016). Word embeddings, analogies, and machine learning: beyond king - man + woman = queen. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 3519–3530). Osaka, Japan, December 11-17: ACL.

Drozd, A. & Matsuoka, S. (2014, September). HPC and Interactive Big Data Analytics: Case Study of Distributional Semantics. In *IPSJ SIG Notes* (Vol. 12, pp. 1–7). Naha: Information Processing Society of Japan (IPSJ).

Drozd, A. & Matsuoka, S. (2016). *Linguistic Regularities from Multiple Samples* (No. C-283). Department of Mathematical and Computing Sciences, Tokyo Institute of Technology.

Ellsworth, M. & Janin, A. (2007). Mutaphrase: Paraphrasing with framenet. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing* (pp. 143–150). Association for Computational Linguistics.

Erk, K. (2007). A simple, similarity-based model for selectional preferences. In *Proceedings of the 45th Annual Meeting on Association for Computational Linguistics* (Vol. 45, pp. 216–223).

Erk, K. (2009). Supporting inferences in semantic space: representing words as regions. In *Proceedings of the Eighth International Conference on Computational Semantics* (pp. 104–115). IWCS-8 '09. Stroudsburg, PA, USA: Association for Computational Linguistics.

Erk, K. (2010). What is word meaning, really?(and how can distributional models help us describe it?) In *Proceedings of the 2010 workshop on geometrical models of natural language semantics* (pp. 17–26). Association for Computational Linguistics.

Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, *6*(10), 635–653.

Erk, K. (2016). What do you know about an alligator when you know the company it keeps. *Semantics and Pragmatics*, *9*(17), 1–63. doi:10.3765/sp.9.17

Erk, K. & Pado, S. (2006). Shalmaneser–a toolchain for shallow semantic parsing. In *Proceedings of LREC* (Vol. 6).

Erk, K., Padó, S., & Padó, U. (2010). A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, *36*(4), 723–763.

Falk, I. & Martin, F. (2016, August). Automatic Identification of Aspectual Classes across Verbal Readings. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics* (pp. 12–22). Berlin, Germany: Association for Computational Linguistics.

Federici, S., Montemagni, S., & Pirrelli, V. (1997). Inferring semantic similarity from distributional evidence: an analogy-based approach to word sense disambiguation. In *Proceedings of the ACL/EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications* (pp. 90–97).

Feizabadi, P. S. & Padó, S. (2015). Combining Seemingly Incompatible Corpora for Implicit Semantic Role Labeling.

Feldman, J., Dodge, E., & Bryant, J. (2015). Embodied Construction Grammar. In B. Heine & H. Narrog (Eds.), *Oxford handbook of linguistic analysis* (2nd ed., pp. 121–145). Oxford handbooks in linguistics. Oxford University Press: New York.

Fellbaum, C. (Ed.). (1998). *Wordnet: An electronic lexical database*. MIT Press Cambridge. Language, speech, and communication series.

Fillmore, C. J. (1975). An alternative to checklist theories of meaning. In *Proceedings of the Annual Meeting of the Berkeley Linguistics Society* (Vol. 1, pp. 123–131).

Fillmore, C. J. (1977). Topics in lexical semantics. In R. W. Cole (Ed.), *Current issues in linguistic theory* (pp. 76–138). Indiana University Press.

Fillmore, C. J. (1979). On fluency. In C. J. Fillmore, D. Kempler, & W. S.-Y. Wang (Eds.), *Individual Differences in Language ability and Language behavior* (pp. 85–101). New York: Academic Press.

Fillmore, C. J. (1985). Frames and the semantics of understanding. *Quaderni di semantica*, *6*(2), 222–254.

Fillmore, C. J. (1992). Corpus linguistics or computer-aided armchair linguistics. In *Directions in corpus linguistics. Proceedings of Nobel Symposium* (Vol. 82, pp. 35–60).

Fillmore, C. J. (2007). Valency issues in FrameNet. In K. Götz-Votteler & T. Herbst (Eds.), *Valency: Theoretical, Descriptive and Cognitive Issues* (Vol. 187, pp. 129–60). Trends in linguistics. Studies and monographs. Walter de Gruyter.

Fillmore, C. J. & Atkins, B. T. S. (1994). Starting where the dictionaries stop: The challenge for computational lexicography. In B. T. Atkins & A. Zampolli (Eds.), *Computational Approaches to the Lexicon* (pp. 349–393).

Fillmore, C. J. & Atkins, B. T. (1992). Toward a frame-based lexicon: The semantics of RISK and its neighbors. *Frames, fields, and contrasts: New essays in semantic and lexical organization*, *103*.

Fillmore, C. J. & Baker, C. (2010). A frames approach to semantic analysis. In B. Heine & H. Narrog (Eds.), *The Oxford Handbook of Linguistic Analysis* (pp. 313–340). Oxford Handbooks in Linguistics. Oxford: Oxford University Press.

Fillmore, C. J., Lee-Goldman, R., & Rhomieux, R. (2012). The FrameNet constructicon. In H. C. Boas & I. A. Sag (Eds.), *Sign-based construction grammar* (no. 193, pp. 309–372). CSLI lecture notes. Stanford, Calif: CSLI Publications/Center for the Study of Language and Information.

Fillmore, C. J., Petruck, M. R. L., Ruppenhofer, J., & Wright, A. (2003, January 9). Framenet in Action: The Case of Attaching. *International Journal of Lexicography*, *16*(3), 297–332. doi:10.1093/ijl/16.3.297

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2002). Placing search in context: The concept revisited. In *ACM Transactions on Information Systems,* (Vol. 20(1), pp. 116–131). ACM.

Firth, J. R. (1957). A synopsis of linguistic theory 1930-55. *Studies in Linguistic Analysis (special volume of the Philological Society)*, *1952-59*, 1–32.

Friberg Heppin, K. & Gronostaj, M. T. (2014). Exploiting FrameNet for Swedish: Mismatch? *Constructions and Frames*, *6*(1), 52–72. doi:10.1075/cf.6.1.04hep

Friedrich, A. & Palmer, A. (2014). Automatic prediction of aspectual class of verbs in context. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL). Baltimore, USA* (pp. 517–523).

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Mikolov, T., et al. (2013). DeViSE: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems* (pp. 2121–2129).

Fung, P. & Chen, B. (2004). BiFrameNet: bilingual frame semantics resource construction by cross-lingual induction. In *Proceedings of the 20th international conference on Computational Linguistics* (p. 931). Association for Computational Linguistics.

Fürstenau, H. & Lapata, M. (2012). Semi-supervised semantic role labeling via structural alignment. *Computational Linguistics*, *38*(1), 135–171.

Gibbs, R. W. [R. W.] & Newman, J. (2002). Embodied standing and the psychological semantics of stand. In *The linguistics of sitting, lying and standing* (Vol. 51, pp. 387–400). Typological studies in language.

Gibbs, R. W. [Raymond W.], Beitel, D. A., Harrington, M., & Sanders, P. E. (1994, January 1). Taking a Stand on the Meanings of Stand: Bodily Experience as Motivation for Polysemy. *Journal of Semantics*, *11*(4), 231–251. doi:10.1093/jos/11.4.231

Gladkova, A. & Drozd, A. (2016). Intrinsic evaluations of word embeddngs: what can we do better? In *Proceedings of The 1st Workshop on Evaluating Vector Space Representations for NLP* (pp. 36–42). Berlin, Germany: Association for Computational Linguistics.

Gladkova, A., Drozd, A., & Matsuoka, S. (2016). Analogy-based detection of morphological and semantic relations with word embeddings: what works and what

doesn't. In *Proceedings of the NAACL-HLT SRW* (pp. 47–54). San Diego, California, June 12-17, 2016: Association for Computational Linguistics.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Golub, G. H. & Van Loan, C. F. (1996). *Matrix Computations (3rd Ed.)* Baltimore, MD, USA: Johns Hopkins University Press.

Green, R., Dorr, B. J., & Resnik, P. (2004). Inducing frame semantic verb classes from WordNet and LDOCE. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics* (pp. 375–382). Association for Computational Linguistics.

Hagen, M. (2014). Full paradigm: russian morphological dictionary (in russian).

Harris, Z. (1954). Distributional structure. *Word*, *10*(23), 146–162.

Hartmann, S., Gurevych, I., & Lap, U. K. P. (2013). FrameNet on the Way to Babel: Creating a Bilingual FrameNet Using Wiktionary as Interlingual Connection. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)* (Vol. 1, pp. 1363–1373). Sofia, Bulgaria: Association for Computational Linguistics.

Hasegawa, Y., Lee-Goldman, R., & Fillmore, C. J. (2014). On the universality of frames: Evidence from English-to-Japanese translation. *Constructions and Frames*, *6*(2), 170–201. doi:10.1075/cf.6.2.03has

Hasegawa, Y., Lee-Goldman, R., Ohara, K. H., Ellsworth, M., & Fillmore, C. J. (2012). *The Frames-and-Constructions Approach to Paraphrase*. ICCG7.

Hasegawa, Y. & Ohara, K. (2006, September). Interview with professor Charles J. Fillmore. (in Japanese). *The Rising generation*, *152*(6), 354–359.

Haspelmath, M. (1993). More on the typology of inchoative/causative verb alternations. *Causatives and transitivity*, *23*, 87.

Hayes, P. J. (1971). *The Frame Problem and Related Problems on Artificial Intelligence*. Stanford University.

Herdağdelen, A. & Baroni, M. (2009). BagPack: A General Framework to Represent Semantic Relations. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics* (pp. 33–40). GEMS '09. Stroudsburg, PA, USA: Association for Computational Linguistics.

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665–695.

Hindle, D. & Sag, I. (1975). Some more on anymore. In *Analyzing Variation in Language: Papers from the Second Colloquium on New Ways of Analyzing Variation* (pp. 89–110). Georgetown University Press.

Hu, B., Lu, Z., Li, H., & Chen, Q. (2014). Convolutional neural network architectures for matching natural language sentences. In *Advances in Neural Information Processing Systems* (pp. 2042–2050).

Hummel, J. E. & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, *110*(2), 220–264. doi:10.1037/0033-295X.110.2.220

Im Walde, S. S. (2000). Clustering verbs semantically according to their alternation behaviour. In *Proceedings of the 18th conference on Computational linguistics-Volume 2* (pp. 747–753). Association for Computational Linguistics.

Im Walde, S. S. (2006). Experiments on the automatic induction of German semantic verb classes. *Computational Linguistics*, *32*(2), 159–194.

Im Walde, S. S., Hying, C., Scheible, C., & Schmid, H. (2008). Combining EM Training and the MDL Principle for an Automatic Verb Classification Incorporating Selectional Preferences. In *ACL* (pp. 496–504). Citeseer.

Itkonen, E. (2005). *Analogy as structure and process: approaches in linguistic, cognitive psychology, and philosophy of science*. Human cognitive processing. Amsterdam ; Philadelphia: John Benjamins Pub. Co.

Jameel, S. & Schockaert, S. (2016). D-GloVe: A feasible least squares model for estimating word embedding densities. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 1849–1860). Osaka, Japan, December 11-17.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer Texts in Statistics. New York, NY: Springer New York.

Johannsen, A., Alonso, H. M., & Søgaard, A. (2015). Any-language frame-semantic parsing. *TARGET*, *82*, 80–.

Johansson, R. & Nugues, P. (2007). Using WordNet to extend FrameNet coverage. In *Building Frame Semantics Resources for Scandinavian and Baltic Languages* (pp. 27–30). Department of Computer Science, Lund University.

Jurgens, D. A., Turney, P. D., Mohammad, S. M., & Holyoak, K. J. (2012). Semeval-2012 task 2: Measuring degrees of relational similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)* (pp. 356–364). Montréal, Canada, June 7-8, 2012: Association for Computational Linguistics.

Kaisser, M. & Webber, B. (2007). Question Answering Based on Semantic Roles. In *Proceedings of the Workshop on Deep Linguistic Processing* (pp. 41–48). DeepLP '07. Stroudsburg, PA, USA: Association for Computational Linguistics.

Kallmeyer, L. & Osswald, R. (2014). Syntax-driven semantic frame composition in Lexicalized Tree Adjoining Grammars. *Journal of Language Modelling*, *1*(2), 267–330.

Katrenko, S. (2012). Could you make me a favour and do coffee, please? implications for automatic error correction in English and Dutch. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation* (pp. 49–53). Association for Computational Linguistics.

Kawahara, S. (2015). Comparing a forced-choice wug test and a naturalness rating test: an exploration using rendaku. *Language Sciences*, *48*, 42–47. doi:10.1016/j.langsci.2014.12.001

Keller, F. & Lapata, M. (2003). Using the web to obtain frequencies for unseen bigrams. *Computational linguistics*, *29*(3), 459–484.

Kiela, D., Hill, F., & Clark, S. (2015). Specializing Word Embeddings for Similarity or Relatedness. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP))* (pp. 2044–2048). Lisbon, Portugal, 17-21 September 2015: Association for Computational Linguistics.

Kilgarriff, A. (2014). The Sketch Engine: ten years on. *Lexicography*, 1–30.

Kim, J.-u., Hahm, Y., & Choi, K.-S. (n.d.). Korean FrameNet Expansion Based on Projection of Japanese FrameNet. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations* (pp. 175–179). Osaka, Japan, December 11-17: ACL.

Koontz-Garboden, A. (2005). On the typology of state/change of state alternations. In *Yearbook of Morphology 2005* (pp. 83–117). Springer.

Kshirsagar, M., Thomson, S., Schneider, N., Carbonell, J., Smith, N. A., & Dyer, C. (2015). Frame-Semantic Role Labeling with Heterogeneous Annotations. *people*, *3*, A0.

Kudo, T. (2005). Mecab: Yet another part-of-speech and morphological analyzer. http://mecab. sourceforge..

Lagus, K. & Airola, A. (2005). Semantic clustering of verbs. In *Acquisition and Representation of Word Meaning: Theoretical and computational perspectives, Linguistica Computazionale XXII-XXIII* (pp. 263–287). IEPI, Pisa-Roma.

Lakoff, G. (2014, February 18). Charles Fillmore, discoverer of frame semantics, dies in SF at 84: he figured out how framing works. http://www.huffingtonpost.com/george-lakoff/charles-fillmore-discover_b_4807590.html.

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

Lapesa, G. & Evert, S. (2014). A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. *Transactions of the Association for Computational Linguistics*, *2*, 531–545.

Lavallée, J.-F. & Langlais, P. (2010). Unsupervised morphological analysis by formal analogy. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments* (pp. 617–624). Springer.

Lazaridou, A., The Pham, N., & Baroni, M. (2016). The red one! On learning to refer to things based on discriminative properties. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 213–218). Berlin, Germany, August 11-12: ACL.

Leech, G. (1981). *Semantics: the study of meaning*. Harmondsworth: Penguin Books.

Lemmens, M. (2002). The semantic network of Dutch posture verbs. In J. Newman (Ed.), *The linguistics of sitting, standing and lying* (Vol. 51, pp. 103–140). Typological studies in language.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *From context to meaning: Distributional models of the lexicon in linguistics and cognitive science, special issue of the Italian Journal of Linguistics*, *20*(1), 1–31.

Lepage, Y. & Goh, C.-l. (2009). Towards automatic acquisition of linguistic features. In *Proceedings of the 17th Nordic Conference on Computational Linguistics (NODALIDA 2009), eds., Kristiina Jokinen and Eckard Bick* (pp. 118–125).

Levy, O. & Goldberg, Y. (2014a). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Vol. 2, pp. 302–308).

Levy, O. & Goldberg, Y. (2014b). Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems* (pp. 2177–2185).

Levy, O., Goldberg, Y., & Ramat-Gan, I. (2014). Linguistic regularities in sparse and explicit word representations. *CoNLL-2014*, 171–180.

Lewis, M. & Steedman, M. (2013). Combining distributional and logical semantics. *Transactions of the Association for Computational Linguistics*, *1*, 179–192.

Lewis, M. & Steedman, M. (2014). Combining formal and distributional models of temporal and intensional semantics. In *Proc. of ACL*.

Li, H. & Abe, N. (1998). Generalizing case frames using a thesaurus and the MDL principle. *Computational linguistics*, *24*(2), 217–244.

Li, J. & Brew, C. (2007). Disambiguating Levin verbs using untagged data. *Proceedings of RANLP 2007*.

Linzen, T. (2016a). Issues in evaluating semantic spaces using word analogies. In *Proceedings of the First Workshop on Evaluating Vector Space Representations for NLP*. Association for Computational Linguistics.

Linzen, T. (2016b). Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP* (pp. 13–18). Berlin, Germany: Association for Computational Linguistics.

Litkowski, K. (2010, July 15). CLR: Linking events and their participants in discourse using a comprehensive FrameNet dictionary. (pp. 300–303). Proceedings of the 5th international workshop on semantic evaluation. Association for Computational Linguistics.

Lönneker-Rodman, B. (2007). *Multilinguality and FrameNet* (No. TR-07-001). Berkeley: ICSI.

Lund, K. & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208. doi:10.3758/BF03204766

Lyashevskaya, O. & Sharov, S. (2009). *Frequency dictionary of modern russian language (in russian)*. Moscow: Azbukovnkik.

Madden, C. & Zwaan, R. (2003). How does verb aspect constrain event representations? *Memory & Cognition*, *31*(5), 663–672. doi:10.3758/BF03196106

Maekawa, K. (2008). Balanced Corpus of Contemporary Written Japanese. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP)* (pp. 101–102).

Mahesh, K., Nirenburg, S., & Beale, S. (1997). If you have it, flaunt it: Using full ontological knowledge for word sense disambiguation. In *Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation* (pp. 23–25).

Malaia, E., Gonzalez-Castillo, J., Weber-Fox, C., Talavage, T. M., & Wilbur, R. B. (2015). Neural Processing of Verbal Event Structure: Temporal and Functional Dissociation Between Telic and Atelic Verbs. In R. G. de Almeida & C. Manouilidou (Eds.), *Cognitive Science Perspectives on Verb Representation and Processing* (pp. 131–140). Springer International Publishing. doi:10.1007/978-3-319-10112-5_6

Matlock, T. (2011). The conceptual motivation of aspect. *Motivation in Grammar and the Lexicon*, 133–147.

Matsubayashi, Y., Okazaki, N., & 'ichi Tsujii, J. (2009). A comparative study on generalization of semantic roles in FrameNet. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint*

Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1 (pp. 19–27). Association for Computational Linguistics.

Matsumoto, D. & Kudoh, T. (1987). Cultural similarities and differences in the semantic dimensions of body postures. *Journal of Nonverbal Behavior*, *11*(3), 166–179. doi:10.1007/BF00990235

McCarthy, D., Venkatapathy, S., & Joshi, A. K. (2007). Detecting Compositionality of Verb-Object Combinations using Selectional Preferences. In *EMNLP-CoNLL* (pp. 369–379).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of International Conference on Learning Representations (ICLR)*. arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Mikolov, T., Yih, W.-t., & Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. In *HLT-NAACL* (pp. 746–751).

Minsky, M. (1974). *A framework for representing knowledge.* Massachusetts Institute of Technology Cambridge, MA, USA.

Newman, J. (2002a). A cross-linguistic overview of the posture verbs "sit", "stand", and "lie". In J. Newman (Ed.), *The linguistics of sitting, standing and lying* (Vol. 51, pp. 1–24). Amsterdam/Philadelphia: John Benjamins.

Newman, J. (Ed.). (2002b). *The linguistics of sitting, standing and lying.* Philadelphia, PA: J. Benjamins Pub. Typological studies in language.

Ohara, K. H. (2009). Frame-based contrastive lexical semantics in Japanese FrameNet: The case of risk and kakeru. In H. C. Boas (Ed.), *Multilingual FrameNets in computational lexicography: methods and applications* (200, pp. 163–182). Trends in Linguistics. Studies and Monographs [TiLSM]. New York: Mouton de Gruyter.

Ohara, K. H. (2013). Toward Constructicon Building for Japanese in Japanese FrameNet. *Revista Veredas*, *17*.

Ohara, K. H. (2014). Relating Frames and Constructions in Japanese FrameNet. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '14)* (pp. 2474–2477). European Language Resources Association (ELRA).

Ohara, K. H., Fujii, S., Ohori, T., Suzuki, R., Saito, H., & Ishizaki, S. (2004). The Japanese FrameNet project: An introduction. In *Proceedings of the Workshop on Building Lexical Resources from Semantically Annotated Corpora at LREC 2004*.

Osswald, R. & Van Valin Jr, R. D. (2014). FrameNet, Frame Structure, and the Syntax-Semantics Interface. In T. Gamerschlag, D. Gerland, R. Osswald, & W. Petersen (Eds.), *Frames and Concept Types* (94, pp. 125–156). Studies in Linguistics and Philosophy. Springer International Publishing.

Ovchinnikova, E., Montazeri, N., Alexandrov, T., Hobbs, J. R., McCord, M. C., & Mulkar-Mehta, R. (2014). Abductive reasoning with a large knowledge base for discourse processing. In *Computing Meaning* (pp. 107–127). Springer.

Padó, S. & Lapata, M. (2005). Cross-lingual bootstrapping of semantic lexicons: the case of Framenet. In *Proceedings of the national conference on artificial intelligence* (Vol. 20, pp. 1087–1092). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Padó, S., Padó, U., & Erk, K. (2007). Flexible, Corpus-Based Modelling of Human Plausibility Judgements. In *EMNLP-CoNLL* (pp. 400–409).

Pajusalu, R. (2001). The polysemy of seisma "to stand": multiple motivations for multiple meanings. *Papers in Estonian Cognitive Linguistics. Publications of the Department of General Linguistics*, *2*, 170–91.

Palmer, A. & Sporleder, C. (2010). Evaluating FrameNet-style semantic parsing: the role of coverage gaps in FrameNet. In *Proceedings of the 23rd international conference on computational linguistics: posters* (pp. 928–936). Association for Computational Linguistics.

Palmer, M., Bonial, C., & McCarthy, D. (2014). SemLink+: FrameNet, VerbNet and Event Ontologies. In *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore* (Vol. 1929, pp. 13–17).

Pantel, P. & Pennacchiotti, M. (2006). Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* (pp. 113–120). Association for Computational Linguistics.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)* (Vol. 12, pp. 1532–1543).

Piñón, C. (2001). Modelling the causative-inchoative alternation. *Linguistische Arbeitsberichte*, *76*, 273–293.

Pustejovsky, J. (1991). The syntax of event structure. *Cognition*, *41*(1), 47–81.

Rapp, R. (2003). Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of the Ninth Machine Translation Summit* (pp. 315–322).

Rappaport Hovav, M. & Levin, B. (1998). Building verb meanings. In M. Butt & W. Geuder (Eds.), *The projection of arguments: Lexical and compositional factors* (83, pp. 97–134). CSLI lecture notes. CSLI Publications, Center for the Study of Language and Information.

Rastogi, P. & Van Durme, B. (2014). Augmenting framenet via PPDB. In *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation* (pp. 1–5).

Rehbein, I., Ruppenhofer, J., & Pinkal, C. S. M. (2012). Adding nominal spice to SALSA–frame-semantic annotation of German nouns and verbs. In *Proceedings of the 11th Conference on Natural Language Processing (KONVENS' 12)* (pp. 89–97).

Resnik, P. (1996). Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, *61*(1), 127–159.

Riaz, M. & Girju, R. (2014). Recognizing causality in verb-noun pairs via noun and verb semantics. *EACL 2014*, 48.

Ritter, A., Etzioni, O. et al. (2010). A latent dirichlet allocation method for selectional preferences. In *Proceedings of the 48th Annual Meeting of the Association for*

Computational Linguistics (pp. 424–434). Association for Computational Linguistics.

Rooth, M., Riezler, S., Prescher, D., Carroll, G., & Beil, F. (1999). Inducing a semantically annotated lexicon via EM-based clustering. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 104–111). Association for Computational Linguistics.

Rüggeberg, C. S. & Sato, H. (2004). Spanish FrameNet and FrameSQL. In *VIII Jornadas de Lingüística* (pp. 105–116). Servicio de Publicaciones.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R., Johnson, C. R., & Scheffczyk, J. (2006). *FrameNet II: Extended theory and practice*. Berkeley, California: International Computer Science Institute.

Russian Hunspell Dictionary [GNU Lesser GPL]. (2013).

Rychlý, P. (2008). A Lexicographer-Friendly Association Score. In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008* (pp. 6–9). Masaryk University, Brno.

Sag, I. A., Boas, H. C., & Kay, P. (2012). Introducing sign-based construction grammar. In H. C. Boas & I. A. Sag (Eds.), *Sign-based construction grammar* (no. 193, pp. 1–30). CSLI lecture notes. Stanford, Calif: CSLI Publications/Center for the Study of Language and Information.

Santus, E., Lu, Q., Lenci, A., & Huang, C. (2014). Unsupervised antonym-synonym discrimination in vector space.

Santus, E., Yung, F., Lenci, A., & Huang, C.-R. (2015). EVALution 1.0: an Evolving Semantic Dataset for Training and Evaluation of Distributional Semantic Models. In *Proceedings of the 4th Workshop on Linked Data in Linguistics (LDL-2015)* (pp. 64–69).

Saurı, R., Goldberg, L., Verhagen, M., & Pustejovsky, J. (2009). *Annotating Events in English. TimeML Annotation Guidelines*. Brandeis University. Version TempEval-2010.

Schank, R. C. & Abelson, R. P. (1977). *Scripts, Plans, Goals, and Understanding: An Inquiry Into Human Knowledge Structures*. Psychology Press.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In *In Proceedings of the ACL SIGDAT-Workshop*. Citeseer.

Schmid, H. & Laws, F. (2008). Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1* (pp. 777–784). Association for Computational Linguistics.

Schmidt, T. (2006). Interfacing lexical and ontological information in a multilingual soccer FrameNet. In *Proc. of OntoLex* (pp. 75–81).

Schmidt, T. (2009, January 14). The Kicktionary −A Multilingual Lexical Resource of Football Language. In H. C. Boas (Ed.), *Trends in Linguistics. Studies and Monographs [TiLSM]* (pp. 101–134). New York: Mouton de Gruyter.

Schnabel, T., Labutov, I., Mimno, D., & Joachims, T. (2015a). Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015), Lisbon, Portugal* (pp. 298–307). Association for Computational Linguistics.

Schnabel, T., Labutov, I., Mimno, D., & Joachims, T. (2015b). Evaluation methods for unsupervised word embeddings. In *Proc. of EMNLP*.

Schönefeld, D. (2006). From conceptualization to linguistic expression: Where languages diversify. In S. T. Gries & A. Stefanowitsch (Eds.), *Corpora in Cognitive Linguistics. Corpus-Based Approaches to Syntax and Lexis* (pp. 297–344). Mouton de Gruyter.

Schuler, K. K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon*.

Schutze, C. T. (1996). *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. University of Chicago Press.

Schütze, H. (1993). Part-of-speech induction from scratch. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics* (pp. 251–258). Association for Computational Linguistics.

Séaghdha, D. O. (2010). Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 435–444). Association for Computational Linguistics.

Séaghdha, D. O. & Korhonen, A. (2012). Modelling selectional preferences in a lexical hierarchy. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation* (pp. 170–179). Association for Computational Linguistics.

Shutova, E., Tandon, N., & de Melo, G. (2015). Perceptually grounded selectional preferences. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 950–960.

Shvedova, N., Arutyunova, N., Bondarko, A., Ivanov, V., Lopatin, V., Ukuhanov, I., & Filin, F. (Eds.). (1980). *Russkaya grammatika: fonetika, fonologiya, udareniye, intonatsiya, slovoobrazovaniye, morfologiya*. Moscow: Nauka.

Siegel, E. V. & McKeown, K. R. (2000). Learning methods to combine linguistic indicators: Improving aspectual classification and revealing linguistic insights. *Computational Linguistics*, *26*(4), 595–628.

Socher, R., Huval, B., Manning, C. D., & Ng, A. Y. (2012). Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 1201–1211). Association for Computational Linguistics.

Soricut, R. & Och, F. (2015). Unsupervised morphology induction using word embeddings. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL* (pp. 1627–1637).

Spitkovsky, V. I., Alshawi, H., Chang, A. X., & Jurafsky, D. (2011). Unsupervised dependency parsing without gold part-of-speech tags. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1281–1290). Association for Computational Linguistics.

Spranger, M. & Loetzsch, M. (2009). The semantics of sit, stand, and lie embodied in robots. In *Proceedings of the 31th Annual Conference of the Cognitive Science Society (Cogsci09)* (pp. 2546–2552). Cognitive Science Society.

Sprouse, J., Schütze, C. T., & Almeida, D. (2013, September). A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010. *Lingua*, *134*, 219–248. doi:10.1016/j.lingua.2013.07.002

Stevenson, S., Merlo, P., Kariaeva, N., & Whitehouse, K. (1999). Supervised learning of lexical semantic verb classes using frequency distributions. *Proceedings of SigLex99: Standardizing Lexical Resources*, 15–22.

Sukhbaatar, S., Weston, J., Fergus, R., et al. (2015). End-to-end memory networks. In *Advances in neural information processing systems* (pp. 2440–2448).

Szumlanski, S. & Gomez, F. (2010). Automatically Acquiring a Semantic Network of Related Concepts. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (pp. 19–28). CIKM '10. New York, NY, USA: ACM. doi:10.1145/1871437.1871445

Tan, Y. & Hovy, E. (2013). Learning Fine-Grained Selectional Restrictions.

The British National Corpus, version 3 (BNC XML Edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium.

The Pham, N., Lazaridou, A., & Baroni, M. (2015). A Multitask Objective to Inject Lexical Contrast into Distributional Semantics. *Proceedings of ACL-IJCNLP, Beijing, China*, *2*, 21–26.

Thomas, M. S. & Mareschal, D. (1997). Connectionism and psychological notions of similarity. (pp. 757–762). 19th annual conference of the cognitive science society, 7 - 10 aug 1997. Stanford, USA.

Tonelli, S. (2010). Semi-automatic techniques for extending the FrameNet lexical database to new languages.

Tonelli, S. & Giuliano, C. (2009). Wikipedia as frame information repository. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1* (pp. 276–285). Association for Computational Linguistics.

Tsvetkov, Y., Faruqui, M., Ling, W., Lample, G., & Dyer, C. (2015). Evaluation of word vector representations by subspace alignment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 2049–2054). Lisbon, Portugal, 17-21 September 2015: Association for Computational Linguistics.

Turney, P. D. (2008). A uniform approach to analogies, synonyms, antonyms, and associations. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)* (pp. 905–912).

Turney, P. D., Pantel, P. et al. (2010). From frequency to meaning: vector space models of semantics. *Journal of artificial intelligence research*, *37*(1), 141–188.

Turney, P., Littman, M. L., Bigham, J., & Shnayder, V. (2003). Combining independent modules to solve multiple-choice synonym and analogy problems. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing* (pp. 482–489).

UzZaman, N., Llorens, H., Allen, J., Derczynski, L., Verhagen, M., & Pustejovsky, J. (2012). Tempeval-3: Evaluating events, time expressions, and temporal relations. *arXiv preprint arXiv:1206.5333*.

Van de Cruys, T. (2014). A Neural Network Approach to Selectional Preference Acquisition. In *EMNLP* (pp. 26–35).

Vecchi, E. M., Marelli, M., Zamparelli, R., & Baroni, M. (2015). Spicy adjectives and nominal donkeys: Capturing semantic deviance using compositionality in distributional spaces.

Viberg, Å. (2013, January 1). Posture verbs: A multilingual contrastive study. *Languages in Contrast*, *13*(2), 139–169. doi:10.1075/lic.13.2.02vib

Vilnis, L. & McCallum, A. (2015). Word Representations via Gaussian Embedding. In *International Conference on Learning Representations (ICLR)*. San Diego, California, May 7-9, 2015. arXiv: 1412.6623

Vylomova, E., Rimmel, L., Cohn, T., & Baldwin, T. (2016). Take and took, gaggle and goose, book and read: evaluating the utility of vector differences for lexical relation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1671–1682). Berlin, Germany: Association for Computational Linguistics.

Wang, H., McAllester, M., Bansal, K., & Gimpel, D. (2015). Machine comprehension with syntax, frames, and semantics. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, 408–413.

Wilensky, R. (1986). Knowledge representation—a critique and a proposal. *Experience, Memory and Reasoning*, 15–28.

Wilson, B. J. & Schakel, A. M. (2015). Controlled Experiments for Word Embeddings. *arXiv preprint arXiv:1510.02675*.

Yap, F. H., Chu, P. C. K., Yiu, E. S. M., Wong, S. F., Kwan, S. W. M., Matthews, S., ... Shirai, Y. (2009). Aspectual asymmetries in the mental representation of events: Role of lexical and grammatical aspect. *Memory & cognition*, *37*(5), 587–595.

Yevgenjeva, A. (Ed.). (1999). *Dictionary of the russian language (in russian)* (4th ed.). Moscow: Russkij yazyk, Poligrafresursy.

Zaliznyak, A. A. & Shmelev, A. (2000). *Vvedeniye v russkuyu aspektologiyu*. Yazyki russkoj kul'tury.

Zanzotto, F. M., Pennacchiotti, M., & Pazienza, M. T. (2006). Discovering asymmetric entailment relations between verbs using selectional preferences. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* (pp. 849–856). Association for Computational Linguistics.

Zapirain, B., Agirre, E., Màrquez, L., & Surdeanu, M. (2013). Selectional preferences for semantic role classification. *Computational Linguistics*, *39*(3), 631–663.