Eötvös Loránd University
Faculty of Humanities
PhD dissertation

MÁRTON MAKRAI

SYMBOLIC AND DISTRIBUTED WORD REPRESENTATIONS

CHAPTERS ON LEXICAL RELATIONS AND CROSS-LINGUAL METHODS

DOI: 10.15476/ELTE.2023.350

Doctoral School of Linguistics
Head: Gábor Tolcsvai Nagy MHAS

Theoretical Linguistics PhD Programme
Head: Zoltán Bánréti CSc

Members of the Committee:
Miklós Törkenczy DSc, chair
Attila Novák PhD, opponent
Tibor Szécsényi PhD, opponent
Beáta Gyuris PhD
Veronika Vincze PhD

Supervisor:
András Kornai DSc

Budapest, December 2023

SUPERVISOR:
András Kornai DSc

to the memory of my godfather Karcsi,
who never stopped urging me
to write this thesis

ABSTRACT

How do the relations used in semantic networks appear in (static) word embeddings? This dissertation is organized around this question. In other words, how can the cognitive structure and lexical relations between concepts be read out from the models trained based on co-occurrences? In addition to lexical relations proper, several chapters deal with argument structure. Another feature of the dissertation is that the tool of linear translation between word embeddings of different languages is used, in addition to its original goals (translation itself and the estimation of the quality of translation pairs) for measuring the *precision* of multi-sense word embeddings. Our theoretical framework is the `4lang` semantic network.

Ja, és fel a fejjel, sursum corda, shit happens.

— (András Kornai)

ACKNOWLEDGEMENTS

For constant help and encouragement is the standard acknowledgment that you should use – told me Andás Kornai when I was writing my mathematics MSc thesis that he advised. I met him in fall 2006, when he taught a Mathematical Linguistics class at the Budapest University of Technology and Economics. I heard about *deep cases* and *kārakas* sooner than about *thematic roles*. In the next several years, he taught me the kind of exploratory, data-driven approach to (computational) linguistics in a master and disciple fashion. While I was a full-time employee, he made it possible for me to attend the courses of the theoretical linguistics programme, some online courses, and conferences. I would like to thank his generosity.

I'm grateful for the motivating work environment and the exciting discussions we had in the Human Language Technologies research group of SZTAKI (the Hungarian Institute for Computer Science and Control). I thank Judit Ács for reminding me during a test presentation of my project that this will be the first point when László Kálmán raises his eyebrows; Gábor Borbély for discussions, either on computation, linguistics, or personal; Dani Lévai for speaking to me even after I spilled my cappuccino on his belongings; I'm grateful that I can call myself an *ismerős* 'acquaintance; familiar (to); =POSS know =REL[person]'[1] of Dávid Nemeskey, with whom we shared an interest in the linguistic content of `4lang` deep cases as well. I thank Kata Pajkossy for advice back then; Gábor Recski for doing what he could to keep the `4lang` project focused, and Attila Zséder, who, besides telling me many best practices in programming, had a great role in that the environment felt much more than a working group.

Most of how I think about (general) linguistics originates from András Komlósy, KomA, and László Kálmán. KomA's influence on how I approach grammatical functions and thematic roles is directly reflected in the thesis (Chapter 5). Kálmán's Socratic/provocative teaching method makes it a task on its own to select what I learned from him from among my whole way of thinking. I also thank Kata Naszádi for conversations that clarified my view of deep cases, Matyi Lagos for exciting discussions on noncounting languages and *below*, and Péter Rebrus for telling me what *defense* in *PhD defense* means.

1 This was the word he used to refer to Bori, now his wife, weeks before their getting engaged.

I would like to thank Tamás Váradi for enabling me to research this topic of distributional word representations full-time over nearly seven years, and Gábor Prószéky for instructive remarks on `4lang`. I'm grateful to Dávid Halász for his open look; Balázs Indig for a memorable bottle of Tatra Tea; Marci Miháltz and Iván Mittelholcz for honest personal feedback; Bence Nyéki for discussions that gave back my faith in representation learning; Csaba Oravecz for ideas, advice, and especially for his humor; Bálint Sass and Ági Kalivoda for much help and deep dialogues of comfort; Eszter Simon and Vera Arató for pieces of advice I took and I didn't regret; and Noémi Vadász for her realism.

The period of the thesis includes my work at the Institute of Cognitive Neuroscience and Psychology. I thank László Balázs who gave examples of being constructive in many kinds of situations, and Bea Ehmann for her enthusiasm at my first conference presentation (Kornai and Makrai 2013).

I'm grateful to Gábor Berend for sharing with me important machine learning insights, and for his patience and humor in more and less successful projects. I thank György Szaszák for his soft but focused leadership in our Hungarian summarization project (Makrai et al. 2022), and Gyuri Orosz for encouragement proper.

The average number of people who read a PhD thesis all the way through is 1.6 (Contestabile 2016). In my case, more than 1 out of those 1.6 must have been Attila Novák, who understood every detail, sometimes better than me. I thank him for his thorough critique. I'm grateful to Tibor Szécsényi who helped me to rethink my thesis on a higher, more conceptual level. I thank János Csirik for playing his impartial role at my pre-defense. I thank Kinga Gárdai, Zoltán Bánréti, Miklós Törkenczy, Beáta Gyuris, Cili Molnár, Veronika Vincze, and Gábor Alberti for their various kinds of work in my PhD process and beyond.

I thank my family and, following Sass (2011), all "those who prayed for me, and Them, who listened to these prayers" (translation mine).

CONTENTS

ACRONYMS

ACH the axiom-concept hypergraph, see Section 2.2.6

ACL The Association for Computational Linguistics

ACT Actions in CD, see Section 2.2.3

AGT Agent, verbal role, see Chapter 5

AI Artificial Intelligence, now technically synonymous to ML. See Section 2.2.4 for an overview of its history

ALS Alternating Least Squares algorithm, see Section 6.3

AMR Abstract Meaning Representation, see Section 2.4.8

AT The locative case of static location, see Chapter 5

BERT A deep LM architecture, the most famous one, see Section 4.3

BPE Bite-pair encoding, mentioned in Sections 4.2.11.2 and 4.3.2

CAUSE A binary predicate used both by Jackendoff (Section 2.3.5), and in `4lang` (Section 7.3)

CCG Combinatorial Categorical Grammar, mentioned in Sections 2.4.8 and 2.4.10

CD Conceptual Dependencies, see Section 2.2.3

CED The Collins-COBUILD dictionary (Sinclair 1987), mentioned in Section 8.2

CG Conceptual Graphs, see Section 2.2.5

CPD Canonical Polyadic Decomposition, see Section 6.3

CS Conceptual Structures, see Section 2.3.5

DAG Directed Acyclic Graph, mentioned in Section 7.1

DAT Dative, verbal role, see Chapter 5

DO Direct Object

DST Dialogue State Tracking, see Section 4.2.10

EFNILEX A computational lexicographic project of the European Federation of National Institutions for Language

FCA Formal Concept Analysis, see Section 7.1

FE Frame Element, see Section 2.4.4

FOR One of the argument cases used for relational nouns in `4lang`, see Chapter 5

FROM The locative case of Source, see Chapter 5

GL The Generative Lexicon, see Section 2.3.7

GLUE A multi-task benchmark for English, see Chapter 1

GMB Groningen Meaning Bank, see Section 2.4.8

GPT Generative Pre-trained Transformer (Radford et al. 2018)

HAS The binary relation of possession in `4lang`

HDBScan A hierarchical density-based clustering algorithom (McInnes, Healy, and Astels 2017)

HLBL Static word embeddings by Mnih and G. E. Hinton (2009)

HNC The Hungarian National Corpus (Oravecz, Váradi, and Sass 2014). We use it in Sections 7.4.2, 7.5 and 8.3

HPSG Head-Driven Phrase Structure Grammar

HS Hierarchical softmax. We experiment with it in Section 7.4.2

KB Knowledge Base, see Section 2.2.9

KR Knowledge Representation, mentioned in Chapter 3

KS14 A benchmark by Kartsaklis and Sadrzadeh (2014), see Section 6.4.1

LDA Latent Dirichlet Allocation, see Section 4.1.3

LDOCE The Longman Dictionary of Contemporary English (Section 2.4.1)

LDV Longman Defining Vocabulary, see Section 3.2

LFG Lexical Functional Grammar

LM Language Model, see Section 4.2

LREC Intl Conference on Language Resources and Evaluation

LSA Latent Semantic Analysis, see Section 4.1.3

LSTM Long short-term memory, one of the major neural network architectures (Hochreiter and Schmidhuber 1997)

ML Machine learning

MLM Masked language modeling, see Section 4.3.2

MLP Multi-layer perceptron

MRD Machine Readable Dictionary, see Section 2.4.1

MSE Multi-sense word embedding (Section 8.1) or Mean squared error (Section 7.5.6)

MSZNY *Magyar Számítógépes Nyelvészeti Konferencia*, the Hungarian NLP conference

MT Machine Translation

NGD Normalized Google Distance, see Section 4.1.5

NLP Natural Language Processing

NMT Neural Machine Translation, see Section 4.3.4

NN Nearest Neighbor, not to be confused with Neural Networks

NP Noun phrase, a concept in structuralist syntax

NSM Natural Semantic Metalanguage, see Section 2.3.3

NSP Next sentence prediction, one of the pre-training tasks for BERT, see Section 4.3.2

OBL Oblique, verbal role, see Chapter 5

OSub Open Subtitles Corpus, see Sections 7.4.3.2 and 8.1

PAT Patient, verbal role, see Chapter 5

PCA Principlal Component Analysis, see Section 4.1.1

PDT The Prague Dependency Treebank

POS Part-of-speech

POSS Possessive, one of the argument cases used for relational nouns in `4lang`, see Chapter 5

(P)PMI (Positive) Pointwise Mutual Information, see Sections 4.1.2 and 6.2

PP Prepositional phrase, a concept in structuralist syntax

PTM Pre-trained model

REL The contentless argument relation in `4lang`, see Chapter 5

RNN Recurrent Neural Network, one of the major neural network architectures

rNN Reverse Nearest Neighbor, see Section 8.1

RNNS2S RNN sequence-to-sequence model

SAT Boolean satisfiability problem, see https://en.wikipedia.org/wiki/Boolean_satisfiability_problem

SEL Sense enumeration lexicons, see Section 2.3.7

SENNA Static word embeddings by Collobert et al. (2011)

SGNS Skip-gram with negative sampling, see Section 4.2.5

SIF Smooth IDF (where IDF is Inverse Document Frequency)

SOTA State-of-the-art

SRL Semantic Role Labeling, see Chapters 2 and 5

SRT Semantic representations are abbreviated this way in the paper discussed in Section 2.4.10

SVD Singular Value Decomposition, see Section 4.1.3

SVO Subject, verb, and object (Section 6.4.2), especially in this order (Section 4.3.2)

TLC The Teachable Language Comprehender (Quillian 1969)

TO The locative case of Goal, see Chapter 5

UCCA Universal Conceptual Cognitive Annotation, see Sections 2.4.8 and 2.4.10

UD Universal Dependencies, see Section 2.4.9

UMAP A manifold approximation method for dimension reduction (McInnes et al. 2018)

uSIF A variant of SIF. We mention it in Section 4.3.6

VP Verb phrase, a concept in structuralist syntax

VSM Vector space models, introduced in Chapter 4

WMT The main conference on machine translation

WSD Word Sense Disambiguation, see Section 8.1

WSI Word Sense Induction, see Section 8.1

XLNET A deep language model by Yang et al. (2019)

Innumerable roads lead to "knowledge,"
and we try to explore many of them.

— Findler (1979)

1 INTRODUCTION

Computational representations of word meaning can be categorized as *symbolic or distributional*, the main examples for the two families being semantic networks and neural networks respectively. Both reflect the structuralist tradition of defining the meaning of words based on their relations to each other, let the relation be conceptual or the probability of co-occurrence. In semantic networks, the full meaning of any concept is *the whole semantic network* as entered from the concept node (Collins and Loftus 1975). Distributional models have achieved better results for decades, but it also has been a problem since the beginnings to interpret how these models actually work. This thesis contributes to this question by analyzing how various relations are represented in static word embeddings. How can the cognitive structure and the lexical relations between concepts be read out from the models trained on co-occurrences?

The relations we investigate range from those frequently discussed in theoretical linguistics like argument relations (e.g. the difference between a word as a subject or as an object), those that are the backbone of lexical databases (lexical relations, i.e. binary relations which are part of the lexical meanings[1] of words like hypernymy, antonymy, or causation); relations that are important from a practical point of view like translation, and analogical relations (e.g. man is to woman as king is to what), where the targeted relations can be any conceptually real one. Word ambiguity can also be called a relation between different lexemes with the same form, or different uses of the same lexeme, and this topic will also be discussed.

The phenomena targeted by alternative approaches to semantics are diverse, ranging from compositionality and the syntax-semantics interface through logical aspects of meaning, to the relation between linguistic meaning and conceptual phenomena. In this classification, our

1 We understand *lexical meaning* as context-independent conceptual meaning. Semantic networks are among the formalisms to represent this. The problem of separating world knowledge from linguistic knowledge will be discussed in Section 3.4.

interest involves both the compositionality of lexical meaning, and the syntax-semantics interface.

This thesis offers computational linguistics research submitted to a theoretical linguistics programme, while we follow a mathematical way of thinking along with an interest in psycho-linguistics.

Accordingly, we gladly reach for all kinds of models and tools and consider them to describe the same thing: the lexical representation of words is inseparable from their conceptual-semantic network.

1.1 4LANG AND ITS ROLE IN THE THESIS

Before drafting the organization of this thesis, we should note that some of our contributions are related to `4lang`, a theory and formalism for representing the semantics of natural language, which has been published along with partial implementation in many research papers and two books. `4lang` will be introduced in detail in Chapter 3, but for the purposes of this road-map of the thesis, let us sketch the `4lang` approach to the process of defining words by each other. `4lang` does not have a pre-defined set of primitives of definition, but we use the *definition graph*, the graph whose nodes are words, and there is an edge *dog* → *faithful* whenever *faithful* is used in the definition of *dog*. This graph is used for computing the *defining vocabulary*, the set of word which suffice to define the rest.

The more important `4lang`-related contributions of this thesis (Sections 3.3, 7.2 and 7.3) take derivatives of `4lang` – the definition graph, or a word embedding created from the graph – as input. Besides, the author of this thesis had a great role in the manual creation of a set of core definitions for `4lang`, but our claims related to this part of the work will focus on to the problem of thematic roles (Chapter 5). Nevertheless, we would like to help the reader to put the `4lang` *theory* in a greater context as well. The papers and the books introducing the theoretical background of `4lang` have assumed that the reader is familiar with a great bulk of literature covering early semantic networks and artificial intelligence, cognitive semantics, and early semantic resources. To make the thesis self-contained, we offer a detailed introduction to this part on the literature as well (Chapter 2).

1.2 ROADMAP

The thesis is organized in two parts: background (Part i) and main contributions (Part ii). Both parts discuss symbolic representations first, followed by distributional ones. Specifically, the background part includes a chapter on symbolic representations (Chapter 2), `4lang` (Chapter 3), and distributed word representations (Chapter 4) each. The main contributions investigate lexical relations in a very broad sense: besides lexical relations proper (hypernymy, antonymy, and causality),

we include thematic and syntactic relations along with other context-independent relations between words like word analogies, translation, and different types of ambiguity. On the distributional side, as we will see in Section 7.2, for a set of male and female words, such as ⟨king, queen⟩, ⟨actor, actress⟩, etc., the difference between the embedding vectors of words in each pair represents the meaning component of gender. In our understanding, these systematic vector differences, computationally represented by the so called vector offsets, correspond to semantic features or *lexical relations* familiar from hierarchical symbolic lexicons and semantic networks.

The first two foreground chapters investigate *verbs and their arguments*. Chapter 5 discusses deep cases in `4lang`, i.e. placeholders of arguments in the meaning representations of predicates, categorized by semanto-syntactic properties of the argument. Our discussion has been based both on theoretical principles, and on our experience in creating a formulaic meaning representation of each item in the defining vocabulary. Our main question is what inventory of deep cases (categories) is needed for the formulaic definition of each word in the defining vocabulary of a multilingual and radically monosemic semantic formalism.

Still on verb arguments, but moving from the symbolic treatment of thematic roles to the distributional representation of „syntactic roles" (i.e. grammatical functions), Chapter 6 investigates the use of different automatic association scores and tensor decomposition methods in the modeling of subject-verb-object triples. The context of this line of research is collocation extraction.

The remaining two chapters are motivated by the question whether relations which intuitively exist, and have been recorded by human labor can also be detected in data-driven distributional representations, more specifically, static word embeddings (word representations obtained with shallow neural networks).

Chapter 7 investigates several lexical relations proper along with word analogies and translation. Going back to Aristotle, word definition begins with specifying a superordinate concept (e.g. a *dog* is an *animal*), also called the *genus* (Section 3.1.2). The word for the superordinate concept is the hypernym, and hypernymy is the topic of our Section 7.1. Which putative semantic features like the already mentioned GENDER are captured by vector space models? What is the geometry of causality like?

The *distributional hypothesis* (Z. S. Harris 1954) says that a word can be described/represented based on how frequently it cooccurs with every other word. More specifically, the distributional *inclusion* hypothesis (Weeds and Weir 2003; Chang et al. 2018) says that hypernymy can be modeled based on that if *animal* is a hypernym of *dog*, *animal* will be grammatical in every context where *dog* is. It is less clear whether *animal* will appear in every context *at least as frequently* as *dog* does. We test the hypothesis with the tools of sparse coding.

Sparse vectors are vectors most of whose coordinates are zero, and non-zero coordinates ideally correspond to interpretable properties. It varies with models whether interpretability follows from the construction of the vectors, or the interpretation needs to be inferred from some latent structure. Even in the latter case, sparse representations tend to be more interpretable than less restricted ones. As far as sparse attributes (i.e. non-zero coordinates in *sparse word representations*) correspond to contexts, if follows from the distributional inclusion hypothesis discussed above that hypernymy should boil down to pointwise comparison. Section 7.1 tests this idea in hypernymy discovery.

Antonymy places words in contrast to those with an opposite meaning (e.g. *peace* ↔ *war*). Section 7.2) analyses this relation. Section 7.3 investigates causality, which has great importance in philosophy, theoretical linguistics, and psychology, while in computational linguistics it remains a bit exotic.

Analogical question like *man* : *woman* :: *king* : ? (*man* is to *woman* what *king* is to what?) have been one of the main evaluation paradigms for static word embeddings. We investigate which morphological and semantic regularities are represented by linear relations in word embeddings of Hungarian, a language with rich morphology and „free word order" (i.e. the order of the main constituents of the sentence is relatively free).

An important application of static word embeddings has been based on that vector spaces of difference languages share their structure to the extent that *word translation* can be formalized as a linear mapping (in the linear algebraic sense). More chapters of this thesis apply this method for various goals. Besides its original goals – translation itself, Section 7.4.2, and the quality estimation of translational pairs, Section 7.5 – we use it to measure the precision of multi-sense word embeddings as the detectors of word ambiguity (Chapter 8). We test whether the methods first published for better-resourced languages also work in medium-resourced languages such as Hungarian, Slovenian, and Lithuanian.

Still within the translation context, we extend linear mapping to triangulation, a.k.a. pivot based lexical induction: we test whether linear mapping can provide a smoother score for triangulated word translations than previous methods. Intuitively, smoothness means that some kind of extra noise in the triangulation (more precisely in pivot-counting) is eliminated by linear translation.

Our last chapter is concerned with one of the greatest problems in lexical semantics: *word ambiguity* and, more specifically, homonymy and polysemy. Static word embeddings, our main tools in the last two chapters, represent each word form with a single linear algebraic vector. This implies that a *crane* will be a thing which lifts blocks of concrete at some times, and takes care of its chicks at others. (This example is by Gábor Prószéky.)

It can be argued, especially from the engineering point of view, that the problem has been solved by contextualized word representations (CWRs, Section 4.3) provided by deep language models. However, the computational linguist still remains interested in the categorical distinction whether a word is homonymous, polysemous, or unambiguous. Coenen et al. (2019) show that the English BERT model, the most popular contextualized model, maps the word form *die* at different regions of the semantic space based on whether it is the German article, the game tool, or the verb, see Section 4.3.3. The former is an artifact of the corpus creation, and the latter two are cases of homonymy. However, Coenen et al. also show that within the verb, BERT also represents how many people die, in a scale-like fashion. This is a shade of the meaning of the sentence where semantics traditionally draws no distinction within the meaning of the predicate: the difference is solely attributed to the argument. There is active research on extracting discrete senses from CWRs.

Multi-sense (static) word embeddings (MSEs) represent the different senses of an ambiguous word with different vectors. This means that they offer an answer to the question how many senses each word has, but they over-disambiguate: some vectors are redundant or simply contain noise. Chapter 8 offers a method in the linear translation setting to measure the precision of MSEs as detectors of word ambiguity. More precisely, we compute two measures, the first of which is trivial: sense vectors should be translated by the linear mapping correctly. If different senses of the same word get mapped to different words in the target language, these translations are evidence (in the theoretical linguist's sense) for the ambiguity of the source word. The second figure of merit is the ratio of putatively ambiguous words whose different senses are mapped to different words by the linear model.

The table of contents at the beginning of the dissertation goes down to sections. There is also a mini table of contents at the beginning of each chapter. These tables go one step deeper, to subsections.

Part I

BACKGROUND

The first three chapters of the thesis give the background in word representations.

Chapter 2 introduces symbolic word representations, which encode the meaning of words in an explicit form such as semantic networks or lexical definitions. Symbolic representations have been used in natural language processing since the earliest days of the field and are still used today.

Some of our contributions are related to `4lang`, a theory and formalism for representing the semantics of natural language. The more important `4lang`-related contributions of this thesis (Sections 3.3, 7.2 and 7.3) take derivatives of `4lang` — the definition graph, or a word embedding created from the graph — as input. Besides, the author of this thesis had a great role in the manual creation of a set of core definitions for `4lang`, but our claims related to this part of the work will focus on to the problem of argument structure (Chapter 5). Chapter 3 introduces `4lang` itself.

Chapter 4 gives a relatively complete account of distributional word representations, which capture the meaning of words based on their distributional patterns in a large text corpus. Distributional representations have become increasingly popular in recent years due to their effectiveness in a wide range of natural language processing tasks. This chapter provides an overview of the different types of distributional representations, including count-based methods and neural network-based methods.

Definition and word meaning need not have anything to do with grammaticalization or grammatical behavior. This is a fairly uninteresting claim about the relation between language and thought.

— *Pustejovsky (1995)*

2

SYMBOLIC REPRESENTATIONS

Some contributions of this thesis, Sections 3.3, 7.2 and 7.3 and especially Chapter 5, are related to the `4lang` theory and formalism for representing the semantics of natural language, which has been developed in the Human Language Technologies Research Group Budapest. The papers and the books introducing the theoretical background of `4lang` have assumed that the reader is familiar with a great bulk of literature on early semantic networks, artificial intelligence, cognitive semantics, and early semantic resources. In this first background chapter we would like to help the reader to navigate in this greater context of symbolic meaning representation. We provide an in-depth exploration of semantic networks and lexical resources, two critical tools for computational lexical semantics in the symbolic approach. We aim to offer a comprehensive overview of the evolution of semantic networks and lexical resources, from early developments to the most recent state-of-the-art approaches.

2.1 ROADMAP OF THE CHAPTER

The chapter is divided to three sections: one on early semantic networks, cognitive semantics, and modern lexical semantics each. The relevance of the former two is that they were very instructive for `4lang`. We also need to reflect on the relation of `4lang` to modern resources, especially regarding to the argument label system.

2.1.1 *Roadmap: Early semantic networks (2.2)*

One of the main contributions of this thesis (Chapter 5) proposes a set of verb argument roles in the `4lang` semantic network. Other important `4lang`-related contributions of this thesis (Sections 3.3, 7.2 and 7.3) take derivatives of `4lang`, especially the definition graph and a word embedding created from the graph as input. As a background for these contributions, Sections 2.2.1 and 2.2.2 give the basics of semantic networks, while Section 2.2.6 introduces considerations about the so called definition graph. The three sections in between review works that are more closely related to early artificial intelligence than to linguistics and that have had a strong impact on `4lang`.

THE TEACHABLE LANGUAGE COMPREHENDER We start the chapter with The Teachable Language Comprehender (Section 2.2.1), arguably the most seminal work on *semantic networks*. It is particularly instructive that Quillian (1969) emphasizes the *recursive* nature of the network, which corresponds to the similar nature of word definition, which becomes the focus of this thesis in Section 3.3. When writing the manual definitions, we built on the tradition of Aristotle's *genus* and *differentia specifica*, which is why it appears so many times

in the introductory chapters (besides Section 2.2.1, in Sections 2.2.8 and 3.1.2).

Though I have no related claim, my way of thinking was also influenced by what our group, the Human Language Technology (HLT) group at SZTAKI, thought and implemented about *pieces of information from different sources* (Nemeskey et al. 2013): the meaning of the words *(„global")*; what we know about one entity in the given situation *(„active")*; as well as a naive theory of a semantic field (e.g. we respect the things which are above us). Quillian is the forefather in this area as well. This section is where *attribute-value matrices* are mentioned for the first time in the dissertation (the other one is Section 2.2.4), which we also used in Nemeskey et al. (2013). The group also took the concept of *inheritance* from Quillian (Recski 2016b).

SPREADING ACTIVATION In Chapter 5, we propose meaning definitions for the defining vocabulary of `4lang`, and categorize the placeholders of the representations of the arguments within the definition of a predicate in a thematic role fashion. In our theory, this system of linking is complemented with a spreading activation mechanism for selectional preferences. Section 2.2.2 summarizes Collins and Loftus (1975)'s detailed treatment of the latter device. Besides, I take from this article the important sentence that *„the full meaning of any concept is the whole network as entered from the concept node."* Here is the first mention of that there can be several link types in a semantic network. One of the main features of `4lang` is that there are only three types of arrows. In this area, we usually refer to Woods (1975), discussed in Section 2.2.4.

ELEVEN VERB-TYPES The greatest added value of `4lang`'s manual definitions, compared to the definitions extracted from monolingual dictionaries, probably lies in the fact that ditransitive (i.e. three-participant) verbs are represented by binary lexical relations (predicates). This is also discussed in one of the earliest `4lang` articles (Kornai 2012). Specifically, among the deep cases proposed in Chapter 5 – in addition to the basic principles that, for example, even nouns can have multiple cases – the dative is probably the most interesting. The two verb classes exemplified by *give* and *say* are usually attributed to Schank (1972) as PTRANS and MTRANS (Section 2.2.3). Gábor Prószéky, in his referee report on Recski (2016b), wrote that Schank is particularly relevant to `4lang`, since both systems "put the considerations of the conceptual world to the fore, sometimes combining elements that are different from a linguistic point of view, and treat them uniformly." (translation mine)

Perhaps the most flourishing example of the proliferation of link types in computational linguistics is also Schank, who not only differentiates the lines and the heads of the arrows with solutions that

push the boundaries of the printing technology of the time, but also implicitly suggests that horizontal and vertical arrows are different.

WHAT'S IN A LINK? Woods (1975) pointed out that *„Links have been used to represent many different levels, e.g. implementation pointers, logical relations, semantic relations (e.g. "cases"), and arbitrary conceptual and linguistic relations."* While in the field of syntactic analysis, `4lang` a builds on the most common formalism (Section 3.7), and thus is in line with the theoretical linguistics tradition, we do not separate *pragmatics* (inferences) from semantics. This important theoretical question is discussed here for the first time in the dissertation.

CONCEPTUAL GRAPHS J. Sowa (1976) places our topic in the broadest context beyond computational linguistics proper: knowledge representation and logic. The latter is particularly important for a dissertation on semantics.

THE NAIVE PHYSICS MANIFESTO Hayes (1979) provides a very detailed theory for the definition of word meaning, what predicates to introduce and how to anchor their meaning. Moreover his examination is based on an axiom-concept graph similar to our definition graph (Section 1.1), which is very instructive for Section 3.3. Our research group follows a simpler principle both in maul and in dictionary-based automatic (Recski 2016a, 2018; Recski, Borbély, and Bolevácz 2016; Ács, Nemeskey, and Recski 2017) vocabulary reduction: we usually define rare words using more frequent ones.

Besides, it is the first time – the other one is Section 2.3.4 – that the abstract/naive locative approach is mentioned, which is an important feature of the manual `4lang` definitions. *„To really capture the notion of 'above', you probably have to go into analogies to do with e.g. interpersonal status: Judge's seats are raised; Heaven is high, Hell is low; to express submission, lower yourself, etc."* Hobbs (2008, summarized in our Section 2.2.7) begins to implement the program outlined by Hayes *(abstract core theories of commonsense knowledge)*.

KL-ONE The Aristotelian *genus* and *differentia specifica* mentioned in connection with Quillian appear in Brachman and Levesque (1985) as *super-concepts and local restrictions*.

CYC Cyc aimed to construct a comprehensive knowledge base of common sense knowledge. One of the eternal topics of semantics is which words can be defined as a conjunction of properties. The fist appearance of this question in this thesis is Section 2.2.9. Besides, `before` and `after` that `4lang` reinvents also come from the partial event slots in Cyc, 'before', 'during' and 'after'.

2.1.2 *Roadmap: Cognitive semantics (2.3)*

Section 2.3 introduces Katz and Fodor (1963)'s seminal paper on semantic features along with a line of semantic research that Kornai (2010a) describes as "the less formally stated, but often strikingly insightful work in linguistic semantics" exemplified by the work of Wierzbicka (1985, Section 2.3.3), Lakoff, Fauconnier, Langacker (1987), Talmy (1988, Section 2.3.4), Jackendoff (1990, Section 2.3.5), and others "often broadly grouped together as 'cognitively inspired' ". (References to sections in the present thesis added.) In Baroni and Lenci (2010)'s reflection, cognitive science and linguistics typically represent concepts as clusters of properties (see our Section 2.3.5): noun properties known as qualia roles (Section 2.3.7), verb selectional preferences and argument alternations (Section 2.3.6), event types, and "topical" relatedness between words, e.g. the relation between *dog* and *fidelity*.

SEMANTIC MARKERS AND DISTINGUISHERS Tibor Szécsényi (personal communication) asked whether the lexical representation of words includes their conceptual/semantic network. Section 3.4 will discuss the philosophically grounded delineation of lexical and world knowledge, but here we note that the approach of the present thesis is more closely related to that of data science, which uses all kinds of models and tools, and considers them to describe the same thing. Katz and Fodor (1963)'s seminal article discusses this common target of description for the case of word meaning. Just as Quillian is the father of the formal side of semantic nets, the linguistic information stored in these representations goes back to Katz and Fodor. The paper proposes a universal set of semantic markers and distinguishers similar to `4lang` concepts. The dissertation is about whether these kind of semantic features also appear in static word embeddings.

CASE GRAMMAR By calling our thematic placeholders of arguments *deep case*s, we strongly committed ourselves to Fillmore (1968). Although I mention Gruber (1965) and Ostler (1979) in the dissertation without further discussion, their influence is also indisputable. The theory is now part of the university curriculum with concepts like semantic roles (Agt, Pat, Dat, Loc), linking, alternations with permanent roles, semantic type (e.g. live), case frames, and linguistic tests (Vendler 1967).

NATURAL SEMANTIC METALANGUAGE Tibor Szécsényi's pre-opponent report summarizes my background chapter on symbolic systems as „how words are related to other words, *more precisely* the concepts denoted by words to the concepts denoted by other words" (emphasis mine). Indeed, although in principle we deal with concepts, our data is about words, so the two are practically synonymous for us. This princi-

ple is is part of Wierzbicka's notion of *natural semantic metalanguage* (NSM), and it can be briefly stated as „there is no (separate) metasemantics", the symbols (predicates/terms/etc.) are drawn from among the words of the object language, and their meaning is the meaning of the corresponding word itself. In particular, every semantically primitive meaning can be expressed by a word, morpheme, or fixed phrase in every language.

In `4lang`, this is clearly true for unaries (e.g. *person, move*) and more or less also for binaries (e.g. *at, cause, -er*). Goddard and Wierzbicka (1994, 1.1.5) suggests to derive the NSM for each language separately, while `4lang` denotes concepts intended to be *language-independent* with English words. In the dissertation, we cite an example of the danger of the opposite approach from the authors of Cyc (which is otherwise discussed in Section 2.2.9): *„what a human can read into* `laysEggsInWater(x)`. Wierzbicka (1972) (more precisely, my section is based on Goddard and Wierzbicka (1994)) of course discuss(es) what consequences this has for *primitives and the syntax of definitions*. What we have already seen with Schank, and is also adopted by `4lang`, arises again: words (morphemes, etc.) can have the same meaning even if the part of speech, the scope of use, or the polysemy pattern is different.

FORCE DYNAMICS IN LANGUAGE AND COGNITION The naive worldview already mentioned in relation to Hayes, which is also intended to be captured by the manual `4lang` definitions, is explained in the greatest detail by Talmy (1988): there is a parallel between the way we talk about physical and psychosocial things. Force dynamics is one of the basic conceptual categories that languages use to structure and organize meaning. Naive physics (unlike scientific physics) is asymmetric: motion and rest, strong and weak. Naive time and space are segmented – again only in opposition to the scientific theory.

CONCEPTUAL STRUCTURES `4lang` was influenced by Jackendoff (1972, 1983)'s theory in several ways. Following Jackendoff's example of the famous *kill: cause to die* (=AGT CAUSE [=PAT[DIE]]), we have *put: cause to (be) at*, (=AGT CAUSE [=PAT AT =TO]). In the all-caps CAUSE, which plays a key role in the elimination of ditransitives, the reader familiar with semantics will discover the primitive conceptual predicate of conceptual structures (CS), although there is some difference: Jackendoff needs these primitives primarily because of the so-called ontological categories. For him, the semantic type of the arguments of each predicate is strictly regulated. In `4lang`, on the other hand, there are no semantic types.

While in `4lang`, the thematic role of an argument does not need to be predictable from its CS position, I was greatly influenced by the feature of CS that the thematic roles correspond to the configurations of the conceptual tree: As Chomsky defines the subject as the NP of

the rule S → NP VP, in semantics for all verbs expressing PTRANS, we see the scheme =AGT CAUSE [=PAT AT =TO], and for MTRANS verbs, we see the scheme =AGT CAUSE [=DAT KNOW = PAT].

Jackendoff also mentions that each semantic field has its own specific inference patterns, which is the naive theory of the given field. CS is also the predecessor of `4lang` in the representation of argument fusion and selection constraints as *unification*.

ENGLISH VERB CLASSES AND ALTERNATIONS The basic principle of Levin (1993)'s verb classes is the formulation of distributional semantics for verbs: We can infer the meaning of the verb from the expression possibilities of the arguments (and adjuncts) and vice versa. This is most closely related to our Chapter 6, but it is related to all my theses, either because of the verbs or the distributional models.

THE GENERATIVE LEXICON My first discussion of polysemy is in Section 2.3.1, but in connection to the motivating question of my cross-lingual word sense induction project (Chapter 8), i.e. the types of polysemy, the generative lexicon – a rich and flexible representation of lexical semantics – cannot be avoided either.

Besides, Pustejovsky (1995)'s theory gives the deepest account of the lexical content. It represents four kind of structures: argument, event, qualia, and inheritance structure. The argument structure of `4lang` is discussed in Chapter 5, while Recski (2016b) discusses inheritance in much detail. Our event structure is very simple (`before, after` and unmarked). The qualia structure (the components, the shape, the purpose, and the creation of things) is described by the definitions themselves – no further constraints apply to this in `4lang`.

2.1.3 *Roadmap: Modern lexical resources (2.4)*

COMPUTATIONAL LEXICOGRAPHY FOR NLP Many useful resources have appeared in the last three decades for computational lexicography, but we start with Boguraev and Briscoe (1989)'s Chapter 1 on machine-readable dictionaries, which is still instructive today: the content of the entries; the defining vocabulary and the problems arising in practice (ambiguous defining words used in a different meaning than they should be used in definitions, adjectives, idioms); the grammar of the definitions. In the field of the latter, for example, we noticed that *or* is often used in the definitions not because disjunction is needed to represent the meaning of many words, but rather it serves to give both a narrower and a broader description of the same thing.

The greatest part of Section 2.4 presents standard resources still in use today. The broadest context of my work is frames (Section 2.4.2). WordNet (Section 2.4.3) is the bread and butter in computational semantics. Our main criticism of it is that it over-disambiguates. I gave

an example of this during the presentation of Makrai (2013): the six meanings of *stomach*. I also used it in connection with antonymy (Section 7.2) and causation (Section 7.3).

The good thing about verb resources is the same as that about standards: there are so many to choose from. The main difference is in the granularity of the argument labels: FrameNet (Section 2.4.4) uses verb-specific tags, PropBank (Section 2.4.6) in principle uses only two core arguments consistently, and VerbNet (Section 2.4.5) has a granularity between the two. It is important that PropBank is used by AMR (Section 2.4.8), the computational semantic formalism with the largest community, which, like `4lang`, represents linguistic meaning with rooted, directed, edge- and node-labeled graphs, and abstracts away from syntactic differences. `4lang` is between VerbNet and PropBank in terms of granularity.

In addition to granularity, the other feature to consider is language dependence, though the universality of a tool does not necessarily depend on how the developers intended it. AMR, for example, claims to be an English-specific framework, yet it is perhaps the most popular universally as well.

ConceptNet (Section 2.4.7) captures general knowledge about words, e.g. that the purpose of a net is to catch fish, which constitutes most of the manual definitions of `4lang`, and what Pustejovsky would call qualia structure. Since semantic analysis always preceded by syntactic analysis, I also consider it important to present the *Enhanced English Universal Dependencies* (Section 2.4.9) dwelling at the border of the two levels of language.

Although we touch on many issues in the overview of various semantic representations, in Section 2.4.10 we give a shallower but even broader draw of the aspects, summarizing Abend and Rappoport (2017) and Koller, Oepen, and Sun (2019). Finally, in Section 2.4.10.4, we also provide a short discussion of Minimal Recursion Semantics and event logic.

2.2 EARLY SEMANTIC NETWORKS

2.2.1 *The Teachable Language Comprehender*

Quillian proposed a spreading-activation theory of human semantic processing, and tried to implement it in computer simulations of memory search, comprehension, and priming. In the description of the memory of the seminal Teachable Language Comprehender (TLC), Quillian (1969) defines text comprehension as relating assertions made or implied in some text to information previously stored as part of the comprehender's general knowledge of the world. Assertions in the text and permanent world knowledge are represented in TLC by the same format. TLC aims to understand general English texts without specific

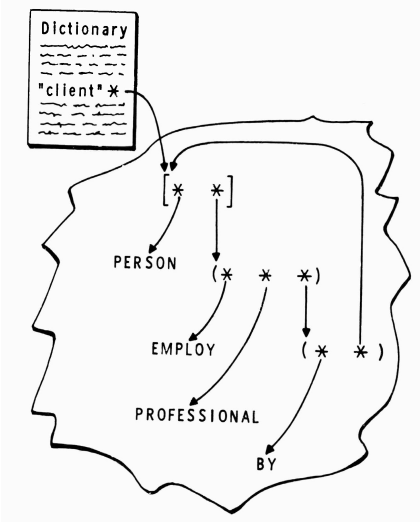

Figure 1: Associative links (Quillian 1968)

(mathematical or visual) reasoning rather than working in a restricted universe like SHRDLU (Winograd 1972). Here we describe the representation format of assertions, but not e.g. the syntactic component (consisting of so called form test) and the teaching protocol.

Figure 1 shows the memory unit representing *client*. The graph represents that a client (represented by the first tuple) is a person who employs a professional. The second tuple represents the (generic) employment event, whose patient is the professional. The third tuple modifies the previous one. This modification correspond to that the employment is done by the client.

Information is encoded as either a *unit* or as a *property*. Units (square brackets) represent objects and events, while properties (parentheses) encode predications. Both brackets and parentheses are ordered lists of pointers (asterisks) to other units or properties. The first pointer in a unit leads to some other unit referred to as that unit's *superset*. The remaining elements, if any, point to properties. Similarly to what we see with semantic markers and distinguishers (Section 2.3.1), the superset and these properties are analogous to the Aristotelian genus and differentia specifica, respectively, with the development that memory units in TLC represent not only lexical items but also specific entities along with what is asserted about them at some point of text comprehension:

> [A] concept is always represented in our format by pointing to some generic unit, its superset, of which it can be considered a special instance, and then pointing to properties stating how that superset must be modified in order to constitute the concept intended. (Quillian 1968)

Properties are attribute-values pairs including traditional dimensions such as $(color, white)$ and dependency pairs (a theory by Tesniére

(1959), formalized by Hays (1964)) such as $(on, hill)$ or $(employed, professional)$. The first element points to the attribute and the second to the value. These two obligatory elements are followed optionally by any number of pointers to other properties. The semantic content of attribute-value pairs is exemplified using *young client* where correct comprehension involves to "supply the fact that this client's 'age' is being judged young, which is not explicit in the text".

The network is responsible for inheritance between concepts, the computation of semantic relatedness, disambiguation, and anaphora resolution with a mechanism that gave rise to the whole theory of *spreading activation* in computational linguistics, to which we turn now.

2.2.2 *Spreading activation*

Simply put, spreading activation is a heuristic variant of shortest path search or breadth-first search in edge-weighted semantic networks with the psychological motivation of modeling semantic memory search and priming. Nemeskey et al. (2013) report spreading activation experiments in the `4lang` framework. This subsection describes Collins and Loftus (1975)'s elaboration of Quillian's theory (shedding light on several misconceptions and offering additional assumptions). Collins and Loftus wanted to give an account of psycholinguistic experiments of their time. In their interpretation, "the full meaning of any concept is the whole network as entered from the concept node", quite reminiscent of the structuralist view on word meaning.

Collins and Loftus extend Quillian's theory of semantic memory search and semantic priming in order to deal with a number of psychological experiments. Priming is a phenomenon whereby exposure to one stimulus influences a response to a subsequent stimulus, without conscious guidance or intention. Earlier exposure to a word influences the response time to a later exposure. The resulting theory can also be considered as a model of human semantic processing in a computer. They argue that the adequacy of a psychological theory should not be measured solely by its ability to predict experimental data: a theory should *produce* the behavior that it purports to explain.

2.2.2.1 *Quillian's theory of semantic memory*

In their first section, Collins and Loftus try to correct a number of the common misconceptions of the original theory. While Quillian's theory was developed as a program for a digital computer, Collins and Loftus elaborate it in psychological terms. People's concepts contain indefinitely large amounts of information, less and less relevant in a specific situation. Concepts (particular senses of words or phrases) can be represented as a node in a network, with properties of the concept represented as labeled relational links. Collins and Loftus specify some properties of the links:

- Links usually go in both directions between two concepts.
- Links have a level of *criteriality*, which are numbers indicating how essential each link is to the meaning of the concept. The criterialities in two directions can be different.
- The full meaning of any concept is the whole network as entered from the concept node.
- There are the following kinds of links:
  - superordinate (IS-A) and subordinate links,
  - modifier links,
  - disjunctive sets of links,[1]
  - conjunctive sets of links, and
  - a residual class of links, which allowed the specification of any relationship where the relationship (usually a verb relationship) itself was a concept.
- Links could form paths of any length.

Priming affects links as well as nodes.

Spreading activation means that search in memory for concepts involves traversing in parallel (simulated in the computer by a breadth-first search) along the links from the node of each concept specified by the input words. The words might be parts of a sentence or stimuli in an experimental task. At each node reached in this process, an activation tag is left that specifies the starting node and the immediate predecessor. If the so called *intersection* node between the two nodes has been found, the *path* that led to the intersection can be reconstructed by following the tags back to both starting nodes. The path is finally evaluated to decide if it satisfies the constraints imposed by syntax and the context.

Collins and Loftus discuss common misinterpretations concerning Quillian's theory. The goal of this section is not to decide these questions, just to show what specific problems arise if one wants to apply spreading activation. The questions may be answered on empirical grounds. There is no difficulty for Quillian's theory in adapting to either solution to the problems below.

- There is a stronger and a weaker version of the cognitive economy principle: "all properties are stored only once in memory and must be retrieved through a series of inferences for all words except those that they most directly define", vs "every time one learns that X is a bird, one does not at that time store all the properties of birds with X in memory" (just possibly some subset of the properties).

1 `4lang`, the semantic network we will introduce in the next chapter, has no disjunction of edges.

- "All links are equal." In Quillian's original theory, there were criteriality tags on links, as we described earlier. Links were assumed to have different accessibility (i.e. strength or travel time). The accessibility of a property depends on how often a person thinks about or uses a property of a concept. Whether criteriality and accessibility are treated as identical or different is a complex issue.
- Memory search (to make a categorization judgment) proceeds from the instance to the category, and not the other way round. I.e. in a categorization task, response time is measured for a subject to decide whether or not a particular instance (e.g., *car*) is a member of one or more categories (e.g., *flower* or *vehicle*).
- "Search rate is slower in proportion to the number of paths that must be searched." vs "Independent parallel search is like a race where the speed of each runner is independent of the other runners" which was a common assumption in psychology.
- Other misconceptions concern whether the network is a rigid hierarchy, or whether the theory predicts it will always take less time to compare concepts that are close together in the semantic network.

2.2.2.2 *The extended theory*

In their next section, Collins and Loftus extend the theory with several assumptions to apply it to some psychological experiments (also transforming the theory from computer terms to quasi-neurological terms): Local Processing Assumptions, Global Assumptions About Memory Structure and Processing, and Assumptions About the Semantic Matching Process, i.e. the categorization tasks, which asks "Is X a Y?". This process occurs in many aspects of language processing, such as matching referents, assigning cases, and answering questions.

2.2.2.3 *Defining and characteristic features*

In the last section, Collins and Loftus deal with the aspects of semantic processing where the model of Smith (1974) is the major competitor to Quillian's theory. Smith represents concepts as bundles of semantic features of two kinds: defining and characteristic features. Defining features are those that an instance must have to be a member of the concept, and features can be more or less defining. Characteristic features are those that are commonly associated with the concept, but are not necessary for concept membership. (The latter correspond to Aristotle's *propria* – e.g. man is the only animal that can laugh – and `4lang`'s defaults.)

Categorization (decisions like "Is a car a flower?") consists of two stages. In Stage 1, all features are investigated, both characteristic and

defining. If the match is above a positive criterion, the subject answers "yes"; if it is below a negative criterion, the subject answers "no"; and if it is in-between, the subject makes a second comparison, which is based on just the defining features. If the instance has all the defining features of the category, the subject says "yes".

The distinction between defining and characteristic features has an inherent difficulty, pointed out "throughout the ages", that there is no feature that is absolutely necessary for any category.

> There is for living things a biologists' taxonomy, which categorizes objects using properties that are not always those most apparent to the layman. Thus, there are arbitrary, technical definitions that are different from the layman's ill-defined concepts, but this is not true in most domains. There is no technical definition of a game, a vehicle, or a country that is generally accepted.
>
> ...
>
> The decision that a 'wren' is not a 'sparrow' would be made because they are mutually exclusive kinds of birds. They are both small songbirds, and it is hard to believe that many people know what the defining features of a sparrow are that a wren does not have. The fact that there are cases where people must use superordinate information to make correct categorization judgments makes it unlikely that they do not use such information in other cases. (Collins and Loftus 1975)

If categorization consists of comparing features between the instance and the category, then it should not matter whether the instance or category is presented first, but experimental data indicates that there is an asymmetry.

Another experiment that might show difficulties with the defining feature model is a categorization task of birds and animals on the one hand, and mammals and animals on the other. Deciding that bird names are in the category 'bird' is faster than that they are in the category 'animal,' whereas people are slower at deciding that mammal names are in the category 'mammal' than in the category 'animal'.

A final argument against defining features is that people have *incomplete knowledge* about the world: we often do not have stored particular superordinate links or criterial properties. Any realistic data base for a computer system will have this same kind of incomplete knowledge. The strongest criticism of the Smith (1974) model is that it breaks down when people lack knowledge about defining features. By viewing superordinate links as highly criterial properties, Quillian's extended theory encompasses a revised version of Smith's model as a special case of a more general procedure.

Levelt, Roelofs, and Meyer (1999) mention two other arguments against defining words as bundles of features. When a word's semantic

features are active, then the feature sets for all of its hypernyms or superordinates are active. Still, there is no evidence that speakers tend to produce hypernyms of intended targets. The other argument is the apparent lack of a semantic complexity effect: words with more complex feature sets are not harder to access than simple ones (measured in reaction time).

2.2.3 *Eleven verb-types*

Most of the preceding two sections concerned the formalism and the search heuristic implementing spreading activation. Now we turn to the semantic content of networks. The first model we investigate is Conceptual Dependency (CD, Schank (1972)).

CD was used by many computer programs of the time that understood English (MARGIE, the Script Applier Mechanism, and the Plan Applier Mechanism). From a linguistic point of view, CD is a meaning representation formalism which is inter-lingual, independent of paraphrase, and appropriate for drawing inferences.

In CD, the process of syntactic parsing is simultaneous with that of drawing some types of inferences. Schank (1973) distinguishes inference from logical deductions (i.e. those applied in automatic theorem proving). "The intent of inference-making is to 'fill out' a situation which is alluded by an utterance [and to tie] pieces of information together to determine such things as feasibility, causality and intent of the utterance." While deductions are highly directed from axioms to some well-defined goal, inferences "are generally made *to see what they can see*".

CD is a deep representation: the representation of a sentence including *buy a book* should include two actions of transfer (one whose object is the book and the other whose object is the price) and the (roles of) participants in these actions. Default arguments (e.g. the object of the verb *drink* is an alcoholic beverage) are also subsumed, though Schank notes that the presence of this default in many languages may be an artifact of shared culture, not that of the underlying (language-independent) concept. Semantic arguments are meant broadly, e.g. the representation of *hit* should include the instrument. Assertions in CD graphs have a measure of confidence attached to them.

We describe the formalism of CD in some more detail as it has been very influential. There are conceptual categories:

- concepts of things that produce a picture (PP) of a real world item in the mind of the hearer, usually expressed by (common or proper) nouns,
- actions (ACTs) that are mostly expressed by verbs, and
- attributes modifying the former two (PA and AA, respectively).

The possible dependencies between concepts are specified by conceptual (relation) rules. Links may be modified for tense. To formulate dependency rules, verbs are "mapped into a conceptual construction that may use one or more [...] *primitive ACTs* in certain specified relationships plus other objects and states". Probably the most famous of these fourteen primitive ACTs are the three types of transfer, transfer of *abstract* relations, e.g. ownership or control (ATRANS), that of *physical* objects (PTRANS), and that of information (*mental* transfer, MTRANS). These are related to the deep dative case `DAT` in `4lang`, see Section 5.4.2.2. In CD, there are four cases: OBJECTIVE, RECIPIENT, DIRECTIVE, and INSTRUMENTAL.

Schank (1973) also discusses inferences that are independent of the specific language. Understanding the sentence *John told Mary that he wants a book* involves the default inference that John wants the books for MTRANS (i.e. for reading), and hearers of this sentence make the inference so spontaneously that they do not even remember whether this ACT was explicitly stated. Another example of the many types of inferences discussed are those about the reasons for actions (motivations of agents). The base for such inferences is constituted by so called *belief patterns*, sequences of causally-related ACTs and states that are shared by many speakers within a culture.

2.2.4 *What's in a link?*

The history of artificial intelligence (AI, and, consequently, that of knowledge representation and connectionism) consists of summers and winters. Based on the Contents section of the Wikipedia page, this history can be summarized as follows:

- The birth of artificial intelligence (1952–1956)
- The golden years (1956–1974)
- The first AI winter (1974–1980)
- Boom (1980–1987): expert systems, knowledge, fifth generation computers, and connectionism
- Bust: the second AI winter (1987–1993)
- Application in industry and specific isolated problems (1993–2001)
- Deep learning, big data and artificial general intelligence (2000–present)

Hubert Dreyfus argued that human intelligence and expertise depend primarily on unconscious instincts rather than conscious symbolic manipulation. Early approaches to artificial common-sense reasoning may seem so naive to the contemporary reader that we are not surprised that winters (periods with disappearing enthusiasm and funding) came. The problems were made explicit by Woods (1975) dealing with the theoretical underpinnings of network representations and the semantics of the networks (nodes and links) themselves. He pointed out that despite the

many publications and demo systems, there was no theory of semantic networks, and existing networks were inadequate for the representations of many linguistic phenomena. Links were used to represent what Brachman and Levesque (1985) call many different *levels*, e.g. implementational pointers, logical relations, semantic relations (e.g. "cases"), and arbitrary conceptual and linguistic relations.

In section 2, Woods discusses what semantics is, and whether it can be separated or even distinguished from syntax on the one hand and inference or "thought" on the other. In his terms, linguistics renders disambiguated representations to sentences, while philosophy maps these to truth values. Retrieval and inference are not part of semantics, nor is pure disambiguation among syntactic parses, even if this is based on selectional restrictions and so-called semantic features. A system needs a separate semantic module for the justification calling it semantic.

The most characteristic notion in a semantic *network* is that of a link that may model human associations. Semantic representations need to be precise, formal, unambiguous, and logically adequate.[2] Woods discusses the problems of the existence of canonical forms, the connection between attribute-value matrices and networks, relations of more than two arguments (e.g. x is 'between y and z')[3], and most importantly the logical type of nodes. Woods' Section 4 discusses two problems that are difficult for AI, restrictive relative clauses, intensional entities (representations of entities without commitment to existence or distinctness), and quantification. Solutions to these problems in the `4lang` theory are offered in Kornai (2023), though they have not yet been implemented.

2.2.5 *Conceptual Graphs*

We continue with Conceptual Graphs (CG, J. Sowa (1976)), a "two-dimensional form of logic", that connects semantic networks discussed so far to the broader discipline of knowledge representation and logic. An excellent introduction is offered in John F Sowa (1992).

CG is a knowledge representation language designed as a synthesis of semantic networks; "logic-based techniques of unification, lambda calculus, and Peirce's existential graphs; linguistic research based on Tesniere's dependency graphs and various forms of case grammar and thematic relations; and data-flow diagrams and Petri nets, which provide a computational mechanism for relating conceptual graphs to external procedures and databases." The result is an expressive system of logic with a direct mapping between natural languages and e.g. expert

2 Logic is one of the main disciplines for meaning representation besides semantic networks and vector-space models. In this chapter, we assume familiarity with first order and intentional logic, and do not go into details, as this is not necessary for the main chapters.

3 The representation of high arity predicates by binary ones is one of the main characteristics `4lang`, see the elimination of "deep ditransitives" in Section 3.1.3.

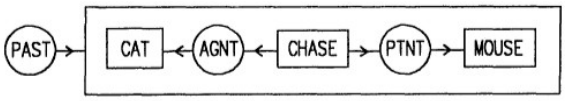

Figure 2: CS graph for *A cat chased a mouse* John F Sowa 1992, p.80

systems. By combining Peirce's contexts with the dependency graphs, CG provides a formalism that can represent Schank's scripts.

As exemplified in Figure 2, CG represents concepts by rectangular nodes and dependency relations ("conceptual" relations) by circular ones as a typed (a.k.a. sorted) version of logic. (As we have already seen, Schank's graphs show conceptual relations as various kinds of arrows instead of these labeled circles.)

2.2.6 *The naive physics manifesto*

Hayes (1979) proposes the construction of a formalization of a portion of common-sense knowledge about the everyday physical world (objects, shape, space, movement, substances, time, etc.) along with a theory of meaning. The main characteristics of the proposed theory are

- thoroughness, i.e. coverage,
- fidelity: the theory should be reasonably detailed,
- density: the ratio of facts to concepts needs to be fairly high (i.e. the units have to have lots of slots), and
- uniformity: a common formal framework (language, system, etc.) so that the inferential connections between the different parts (axioms, frames,...) can be clearly seen. It is methodologically important to allow the use of a variety of formalisms in sub-areas, but idiosyncratic formalisms should be systematically reducible to the basic formalism, and be regarded as 'semantic sugar'.

In this section, we introduce sections 3 to 6 of Hayes (1979). For modern advances in this direction, see Hobbs (2008), which we discuss in Section 2.2.7.

2.2.6.1 *The axiom-concept graph: clusters and density*

A naive physics formalization consists of many assertions and symbols (i.e. tokens: relation symbols, function and constant symbols) – or: frame headers, slot names, etc.; or: node and arc labels, etc. The meaning of the tokens is defined by the structure of the formalization, by the pattern of inferential connections between the assertions. The formalization is dense, if for each token, there are many axioms involving it, which pin down the meanings of the tokens. This view of

meaning differs profoundly from the view which holds that tokens in a formalization are words in a natural language.

The axiom-concept hypergraph (ACH) consists of nodes corresponding to tokens of the formalization; and arcs corresponding to axioms: an arc links the tokens that it uses. The formalization is dense if the ACH is highly connected. Hayes does not expect density to be uniform: there will be more dense clusters of concepts. Identifying these clusters is one of the most important and difficult tasks. E.g. what happens with liquids, is part of the liquids cluster, not part of some theory of 'what-happens-when': causality is not a cluster. Cluster identification is hard, since a large conceptual structure can be entered anywhere. If it seems hard to say anything very useful about the concepts, that can mean that one has entered the graph at a locally sparse place, rather than in a cluster. This thesis analyses a similar graphical representation of the `4lang` semantic network and the connected components of the graph in Section 3.3.

Clustering is hierarchical: e.g. the collection of concepts to do with three-dimensional shape and orientation ('above', 'below', 'tall', 'fat', 'wide', 'behind', 'touching', 'resting on', 'angle of slope', 'edge' (of a surface), 'surface' (of a volume), 'side', 'vertical', 'top', 'bottom', which have many internal relationships) must appear significantly in conceptual frameworks that underlie visual perception and locomotion, describing assemblies, the theory of liquids, and that of physical actions and events.

Hierarchical organization is a point where we disagree with Hayes: while the lexical relations this thesis is about could theoretically be organized in a hierarchy, we think that such an organization is disadvantageous from a practical point of view, as different contrast will show up in different branches of the hierarchy, and even the relations that appear in more branches may enter in different vertical order.

2.2.6.2 *The a/c ratio and reductionist formalizations*

The ratio of axioms to concepts (the a/c ratio) will be large for a dense axiomatization. Any interesting axiomatization will have a/c greater than one; but there are interesting axiomatizations in which a/c will be very close to 1. E.g. in the Zermelo-Fraenkel set theory, $c = 2$ (the concepts are '$\in$' and 'set') and $a = 8$. This theory enables one to define many concepts (e.g. the integers; the rationals; the reals), and the desired properties of these concepts (e.g. the principle of induction for integers or the continuity of the real line) follow from the structure of these definitions, and the axioms as theorems of the axiomatization. The axiomatic approach to naive physics which Hayes proposes is different. Set theory is reductionist in the extreme: it is extraordinarily sparse. By adding definitions to a reduced theory, a/c tends asymptotically to unity. The resulting ACH has one very small cluster at the center, surrounded by a cloud of nodes each linked radially. This reduc-

tionist graph is quite a different 'shape' from the connected, clustered graph of a dense axiomatic theory. Hayes believes that there is no such small, reductionist theory for common sense reasoning.

Many approaches in the artificial intelligence literature, make a reductionist assumption or 'semantic primitives', exemplified by the work of Wilks (1977) and Schank (1975). The number of primitives is about 90 in Wilks, and 14 in Schank. Schank and his students associated inference molecules with the 14 primitive action-tokens, which play the same sort of central organizing role that the set axioms do. The desired properties of e.g. buying or giving follow from their definitions, and the meaning given to the primitives by the core theory. Hayes criticizes Wilks for merely presenting a list of tokens with a brief description, i.e. the semantic primitives being English words. A reductionist, semantic-primitives based approach to meaning may be adequate for some subtasks, but in real human language understanding at some point we will have to represent detailed knowledge of the world.

2.2.6.3 *Meanings, model theory, and fidelity*

If the meanings of tokens are not specified by definitions, then how? A token means a concept to the extent that the formalization enables a sufficient number of inferences to be made whose conclusions contain the token. But Hayes assumes that a formalization has an adequate model theory as well, i.e. tokens have extension. Hayes highlights the widespread delusion of confusing a formal description of a model found in the textbooks with the actual model. If axiomatization has a very much simpler model than the intended one, then the tokens mean no more than they mean in the simple model. This is what Hayes means by 'fidelity'. E.g. an adequate formalization of a blocks world will be such that any model of it must have an essentially three-dimensional structure. Fidelity is how closely the simplest model resembles the intended one.

A related problem is that the meaning of a token depends upon the entire formalization, a *change* to any part of the formalization can change every other part. People with different formalizations in their heads may understand the same token in different ways. Find a substance and a set of circumstances such that I would call it 'water' and you would not! It is even possible when our beliefs about water (i.e. all the assertions which actually contain the token 'water') are identical. The difference may lie in some related concept (such as viscosity, or drinkability) which we understand differently. It may not even be possible to say exactly which tokens we differ on. One of the good reasons for choosing naive physics to tackle first is that there seems to be a greater measure of interpersonal agreement here.

If you change the meaning of 'water', the change in the meanings of other tokens is less, the further away the token is from 'water'. As a working hypothesis, you may identify this distance with shortest-

path distance in the ACH hypergraph. Thanks to this distance-dilution effect, it seems a reasonable strategy to, first, work on clusters more or less independently. You can introduce concepts, which occur in some other cluster, fairly freely, assuming that their meaning is reasonably tightly specified there. E.g. in considering liquids, I needed to talk about volumetric shape: our concept of a horizontal surface would hardly be complete if we had never seen a large, still body of water — but we assume of a fairly autonomous theory of shape. The 'definitions' view of meaning is theoretically wrong, but a good method. Finally, Hayes talks about the body and sensory input. As any consistent first-order axiomatization has a model with only symbols, 'motor tokens' — symbols which describe bodily movements — should directly be related to the body.

2.2.6.4 *Thoroughness and closure*

One way to have a high a/c ratio, it might seem, would be to keep c small: find some small, self-contained groups of concepts which could be formalized in total isolation to a reasonable degree of fidelity. But in a typical situation, one quickly needs to introduce tokens, and in order to pin down their meanings, yet more concepts. The proliferation of tokens seems to be getting out of hand. If one thinks of exploring the ACH, one needs a sense of direction, to stay within the current cluster. During the formalization process, the proliferation must slow down eventually. The 'thoroughness' requirement is to go on until this slows down, when our collection of concepts has closed upon itself, so that all the things one wants to say in the formalization can be said using the tokens which have already been introduced. This means we have spanned the entire graph, and need only to add new arcs, filling out the graph until its density is sufficient to capture the meanings of its tokens. Hayes's program is to get a formalization which is closed and has high fidelity (so, high density): then it must also be thorough.

To achieve greater fidelity, one will need greater thoroughness. E.g. to really capture the notion of 'above', you probably have to go into analogies to do with interpersonal status: (Judge's seats are raised; Heaven is high, Hell is low; to express submission, lower yourself, etc.) Imagine a world in which the 'status' analogy was reversed. That is a possible model of naive physics, but not of common sense. A formalization cannot be deep without being broad, and must be deep to be dense: so a dense formalization must be deep and broad. The cluster hierarchy mentioned before depends upon the fidelity, the level of detail. The programme of tackling naive physics in isolation is based on the belief that there is a level of detail at which naive physics forms a close cluster in a rich but tractable level of detail.

Composite Entities	perfect, empty, relative, secondary, similar, odd
Scales	step, degree, level, intensify, high, major, considerable
Events	constraint, secure, generate, fix, power, development
Space	grade, inside, lot, top, list, direction, turn, enlarge, long
Time	year, day, summer, recent, old, early, present, then, often
Cognition	imagination, horror, rely, remind, matter, estimate, idea
Communication	journal, poetry, announcement, gesture, charter
Persons	leisure, childhood, glance, cousin, jump
Microsocial	virtue, separate, friendly, married, company, name
Bio	breed, oak, shell, lion, eagle, shark, snail, fur, flock
Geo	storm, moon, pole, world, peak, site, sea, island
Material World	smoke, shell, stick, carbon, blue, burn, dry, tough
Artifacts	bell, button, van, shelf, machine, film, floor, glass, chair
Food	cheese, potato, milk, bread, cake, meat, beer, bake, spoil
Macrosocial	architecture, airport, headquarters, prosecution
Economic	import, money, policy, poverty, profit, venture, owe

Table 1: Concepts in Hobbs (2008)

2.2.7 *Deep Lexical Semantics*

Now we turn to Deep Lexical Semantics (Hobbs 2008), motivating it from a more recent perspective. HellaSwag (Zellers et al. 2019) tests pre-trained deep language models like BERT (Section 4.3) with questions like which of the alternatives below finishes the short text *A woman is outside with a bucket and a dog. The dog is running around trying to avoid a bath. She...* the most appropriately.

1. rinses the bucket off with soap and blow dry the dog's head.
2. uses a hose to keep it from getting soapy.
3. gets the dog wet, then it runs away again.
4. gets into a bath tub with the dog.

The good answer is 3. Models struggle with this task. The authors note that while the wrong endings are on-topic, with words that relate to the context, humans consistently judge their meanings to be either incorrect or implausible. These problems suggest that for understanding, we need something beyond the meaning of the words, and their probability in different sentence contexts. We saw in Section 2.2.6 that Hayes (1979) suggested the construction of a formalization of a portion of common-sense knowledge about the everyday physical world along with a theory of meaning. Deep Lexical Semantics (Hobbs 2008) is a further step in this direction.

Hobbs (2008) took a basic core of the about 5000 most frequent synsets in WordNet; categorized these into sixteen broad categories,

e.g. time, space, scalar notions, composite entities, and event structure; and sketched out the structure of some of the underlying abstract core theories of commonsense knowledge (see Table 1). The latter includes the basic predicates in terms of which the most common word senses need to be defined or characterized; axioms that link the word senses to the core theories; and a kind of "advanced lexical decomposition", where the "primitives" into which words are "decomposed" are elements in coherently worked-out theories. Hobbs (2008) focuses on the 450 synsets that are concerned with *events* and their structure.

Hobbs has very similar principles to Hayes (1979): We must have underlying theories and axioms that link these to words. Concepts and axioms include domain-dependent knowledge, of course, but 70-80% of the words in most texts, even technical texts, are words in *ordinary* English. Hobbs chooses the core theory of *scales*, which will provide axioms involving predicates such as 'scale', '<', 'subscale', 'top', 'bottom', and 'at'. These are abstract notions that apply to partial orderings as diverse as heights, money, and degrees of happiness.

Some lexical and world knowledge can be acquired automatically, e.g. the correlation between "married" and "divorced". The corresponding predicate-argument structures may also be acquired, along with which way the implication goes and with what temporal constraints. But consider a more complex relation to illustrate his axiomatization method, that of a range. In Hobbs's view, it is feasible to manually axiomatize the meanings of several thousand words, what can achieve the desired complexity and reliability of the core theories and the linking axioms.

Section 3 describes the following core theories that are crucial in characterizing *event* words:

- Eventualities and their Structure: states and events,
- Set Theory (modeled in a standard fashion),
- Composite Entities, including the predicate 'partOf' and the figure-ground relation 'at',
- Scales: partial orderings, monotone functions, the construction of composite scales, the characterization of qualitatively high and low regions of a scale (related to distributions and functionality), and constraints on vague scales,
- Change of State
- Cause. Recall that Hayes (1979) explicitly warned against trying to formalize causality, saying that what happens e.g. with liquids, is part of the liquids cluster, not part of some theory of 'what-happens-when'.

  In Hayes view, causality is characterized by two properties: If every eventuality in a causal complex happens, the effect happens;

and everything in the causal complex is *relevant* to the effect in a way that can be made precise. Hobbs's approach to causality includes force-dynamic notions (Section 2.3.4) like *enable, prevent, help, obstruct, attempts, success, failure, ability*, and *difficulty*.

- Events. Changes of state and causality compose into more complex events: conditional, iterative, cyclic, and periodic events. This part of the theory is linked with several well-developed ontologies for event structure.
- a well-developed theory of *time*,
- a rather sparse theory of *space*, and
- a large number of theories explicating a commonsense theory of *cognition*,
- the predicates 'possess' and 'remain' would be explicated in a commonsense theory of *economics*.

2.2.8 KL-ONE: *super-concepts and local restrictions*

We have already seen a computational formalizations of the Aristotelian *genus* versus *differentia specifica*: in the form of supersets and properties in Quillian (1969)'s semantic memory. The next section will discuss semantic markers and distinguishers in Katz and Fodor (1963)'s approach to meaning decomposition as well. A third example is KL-ONE (Brachman and Levesque 1985), where *concepts* are described by their subsuming concepts (their super-concepts), their local internal structure expressed in *roles* (which describe relationships like properties or parts) and structural descriptions, which express the interrelations among the roles.

> A Concept must have more than one super-concept (if there are no local restrictions), differ from its super-concept in at least one restriction, *or* be primitive. A Concept with no local restrictions is defined as the conjunction of its super-concepts. Super-concept serves as a proximate genus, whereas the local internal structure expresses *essential differences*, as in classical classificatory definition (Sellars 1917). The network structure formed by the subsumption relationships between Concepts [is] a *taxonomy*. (Brachman and Levesque (1985), emphasis added)

KL-ONE instigated first-class status for "Roles" a.k.a. slots. We conclude this description of KL-ONE with *contexts*. Individual concepts, i.e. concepts that uniquely describe individuals, are associated to some context. Assertions about co-reference and existence are also always relative to some context so as not to affect the taxonomy of generic

knowledge. Context provides the mechanism for reasoning about hypotheticals, beliefs, and desires.

2.2.9 *Cyc*

Most of the problems discussed in Section 2.2.4 have not been solved to this day, and though expert systems in specific domains brought a second summer in AI, the second winter also arrived due to brittleness outside these narrow domains. The drop in reputation and funding was a sign of the need to represent commonsense knowledge, the wisdom of a kindergarten child in a knowledge base (KB). Cyc (Lenat and Guha 1990) is an early example of effort in this direction. In retrospect, their success lies between the two extremes they formulated as at least providing some insight into issues involved in ontology population with "an indication as to whether the symbolic paradigm is flawed" and the more optimistic one that "no one in the early twenty-first century even considers buying a machine without common sense". For `4lang`, Cyc is relevant especially for the status of primitives, see Section 3.2. While there is a related open resource, and even the framework is not much older than WordNet (Section 2.4.3), its impact on `4lang` is rather theoretical, so we discuss it here, among the early works that laid down the principles of semantic networks rather than in Section 2.4, where modern lexical resources are introduced.

Lenat and Guha (1990) organize their paper along the three tasks in building a KB: the (logical) language (CycL), the procedures for manipulating knowledge, and populating the KB. The authors frame understanding as including "beliefs, knowledge of others' [...] limited awareness of what we know, various ways of representing things, [and] knowledge of which approximations (micro-theories) are reasonable in various contexts". In our description of Cyc we concentrate on its aspects with the greatest impact on `4lang`, the language and the database, rather then inference.

The two systems are already similar in their methodologies: the core of the Cyc ontology was built manually, and it was extended in the 80s by knowledge entered in primarily automatic fashion.

> We developed our representation language incrementally as we progressed with [the task of knowledge encoding]. Each time we encountered something that needed saying but was awkward or impossible to represent, we augmented the language to handle it. Every year or two we paused and smoothed out the inevitable recent "patchwork." (Lenat and Guha 1990)

The language is summarized as frame-based and embedded in a more expressive predicate calculus framework along with features like representing defaults or reification (allowing one to talk about propositions

in the KB). As for the inference machine, they abandon the AI tradition of a single, very general mechanism (e.g. resolution) for problem solving and prefer special data structures and algorithms for problems of varying complexity as done in traditional computer science.

The main difference between the Cyc KB and `4lang` is that we distinguish the core vocabulary from the broader one, while this distinction is not made in the Cyc KB where, though many of the one or two million assertions are general rules, some are specific facts dealing with particular entities and events (e.g. famous people and battles.)

A great heritage of `4lang` from Cyc is the use of non-monotonic reasoning: most assertions are *default* beliefs and the addition of new facts can cause them to be retracted. Cyc is also similar to present day question answering systems in that inference is based upon a (quickly identified) small subset of relevant sentences.

Though Cyc is *strongly typed* (as opposed to the type-free `4lang`), it offers us many insights. Lenat and Guha frequently use "set-theoretic notions to talk about collections, but these collections are more akin to what W. Quine (1969) termed *natural kinds*, e.g. *dog* or *lemon*, that are usually assumed not to be completely definable as intersections of more primitive classes. Collections are organized in a generalization-specialization hierarchy" (Brachman and Levesque 1985).

Cyc handles *time and actions* analogously to space: time and events are substances. "One could take a glob of peanut butter and separate out all the peanut chunks, and these alone do not form a glob of peanut butter. [...] The substancehood principle applies only to pieces larger than the *granule* of that substance." 'Walking' is a type of temporal substance by the same token.

As there are "orthogonal ways of breaking down a physical object, there [are two] orthogonal ways of breaking down an action:" actors and subEvents. There are separate categories of slots that are used in order to relate actors to actions and subEvents to events. To put it so simply that may seem brutal from a strongly typed point-of-view, but excellent for `4lang` purposes: actor slots are roles like *performer, victim, and instrument* and sub-event slots are 'before', 'during' and 'after' the action. The later are the predecessors of the `4lang` concepts representing event structure with the same names (except for 'during' which is unmarked in `4lang`), and can also be compared to the PRE- and POST- procedures (conditional execution) and WHEN- (side effects) in KL-ONE (see the previous section).

2.3 COGNITIVE SEMANTICS

Recall the overview of this section in Section 2.1.2.

2.3.1 *Semantic markers and distinguishers*

We start this section with the standard model of the featural decomposition of lexical meaning due to Katz and Fodor (1963). The paper describes its aim as the organization of *facts contributed by diverse fields* including philosophy, linguistics, philology, and psychology. The first part of the paper describes *the domain, the descriptive and explanatory goals, the mechanisms, and the empirical and methodological constraints* upon a semantic theory. They want to find a balance of *strict formalism* (developed some years later in Montague Grammar) and *great explanatory power* (like traditional lexicography). The input to their semantic model is a sentence analyzed by a recursive compositional grammar, in modern terms, a parse tree. These authors require a semantic theory be capable of recognizing (and resolving) *ambiguity, paraphrase, and anomaly* (e.g. *The paint is silent*) but other aspects like the computation of truth values is deferred.

The difference between syntax and semantics is that the latter may rely on *context*, mainly linguistic one (the dialog), and to a restricted degree, extra-linguistic one (world knowledge). Their notion of *world knowledge* subsumes facts like 'buildings do not jump', which is needed for comprehending the sentences *Joe jumped higher than the Empire State Building* and *Joe jumped higher than you* differently. The theory should "interpret discourses just so far as the interpretation is determined by grammatical and semantic relations which obtain within and among the sentences of the discourse."

The components of the proposed semantic theory include the *dictionary* (the same module we will call the lexicon) and one that could be called a *word-sense disambiguation* method in present-day terms. The most important part of this theory is the structure of dictionary entries. Besides part-of-speech (POS) specification and, optionally, explicit cross-references to synonyms, dictionary entries consist of *sense characterizations* like that in Figure 3. The key notion is that of the *semantic markers* (in parenthesis, e.g. $(Human)$) that represent relations between meanings of the same polysemous word and between different dictionary entries. *Distinguishers* (in brackets) assigned to a lexical item are intended to reflect what is idiosyncratic about its meaning. This distinction is analogous (Kornai 2019, Chapter 5) to the Aristotelian notion of *genus* (a *mirror* is a 'plain surface') and a *differentia specifica* (. . . that 'reflects'). The unenclosed elements are *grammatical "markers"* (features). Semantic markers play a role in disambiguation, selectional restrictions, and, in a limiting case of selectional restrictions, the detection of semantic anomaly. The formalism also allows restrictions for the arguments of the items, e.g. $\langle(Female)\rangle$ in the representation of one of the senses of *honest* designates that the corresponding meaning of *honest* applies only to arguments with the $(Female)$ marker. In the concluding section, the authors mention that there may exist a

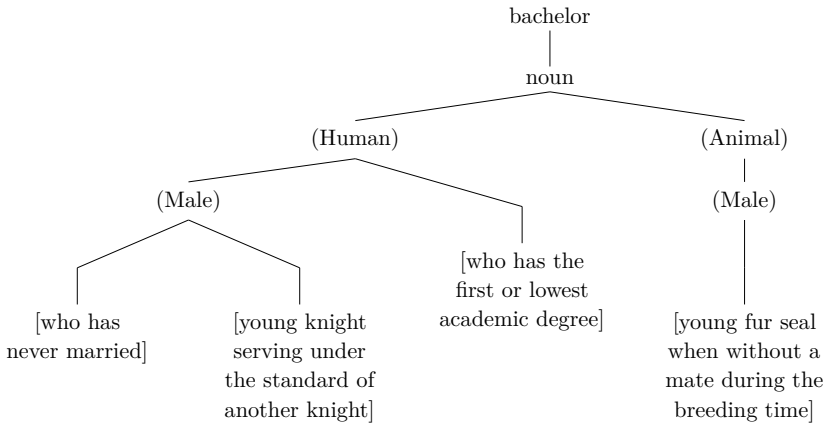

Figure 3: The sense-characterization of *bachelor* by Katz and Fodor (1963)

universal inventory of semantic markers from which the markers of each particular language are drawn, a goal `4lang` (the semantic representation framework that we introduce in the next chapter) shares with this theory.

2.3.2 *Case Grammar*

One of the main chapters of this thesis (Chapter 5) introduces the semantic roles used in `4lang`, the concept network of the research group the author belongs to. Our system has been heavily influenced by Case Grammar (Fillmore 1968), which this section introduces.

Our introduction is based on Palmer, Gildea, and Xue (2010, Chapter 1), who investigate *semantic roles* (semantic relations and predicate-argument structure) and the controversies surrounding them. They start with the example that from a sentence like *John threw a ball to Mary in the park*, an NLP system should identify a throwing event, John as the Agent or Causer of the event, Mary as the Recipient, the ball as the item being thrown, and the *location* of the throwing event. The linguistic theory of mapping from the syntactic analysis of the sentence to the underlying predicate argument structures is known as *Linking*. On the syntactic side, we have *alternations* like *John broke the window/The window broke* with the same semantic role (or conceptual relation) in both sentences. (In the example, it would typically be labeled as the Patient.)

Case Grammar originated with Fillmore's paper on "deep" cases, i.e. *semantically typed verb arguments* (Fillmore 1968). The theory involves types of nouns with different types of cases, e.g. the Agentive and *Dative* roles are most likely to be of type *animate*. Argument *frames* specify the number, type and obligatory/optional nature of roles associated with a verb. *Linguists developed tests* for determining whether two noun phrases have the same case. For instance, members of a conjunction have the same case. Representing *alternative role assignments*

(e.g. *Mother is cooking the potatoes/The potatoes are cooking/Mother is cooking*) by the same deep cases can result in a more compact lexicon. Even *like* and *please* can be considered semantically equivalent, distinguished only by their preferred mappings. Within the semantic domain, generalizations can be exploited in the form of commonalities e.g. between the Agentive cases and the Objective cases of actions such as hitting, breaking, and cutting.

The inventory of roles differ between flavors of the theory, only the Agent and the Patient being relatively straightforward. The *Agent* is the initiator of the action, the doer, typically acting deliberately or on purpose. The question *What did X do?* can be applied, with *X* being the Agent. The *Patient*, on the other hand, is being acted upon. It is likely to change state as a result of the Agent's actions. The questions *What happened to Y?* or *What did X do to Y?* would apply.

2.3.3 *Natural Semantic Metalanguage*

Lenat and Guha (1990) formulate one of the greatest problems remaining in modern semantic networks as follows.

> Programs often use names for concepts such as predicates, variables, etc., that are meaningful to humans examining the code; however, only a shadow of that rich meaning is accessible to the program itself. For example, there might be some rules that conclude assertions of the form `laysEggsInWater(x)`, and other rules triggered off of that predicate, but that is only a fragment of what a human can read into `laysEggsInWater` (Lenat and Guha 1990)

A solution to the problem of arbitrary node-labels has been offered outside of the computational realm, by the Natural Semantic Metalanguage (NSM) approach (Wierzbicka 1972) that we introduce following the first two chapters in a more recent collection (Goddard and Wierzbicka 1994), "the first attempt ever to empirically test a hypothetical *set of semantic and lexical universals* across a number of genetically and typologically diverse languages" with "parallel and strictly comparable answers to the [questions of] a shared set of concepts, forming the common conceptual foundation of all cultures".

The principles of the work are specified in their Section 1.1 and enumerated below. Goddard includes a discussion of the opinion of the main semanticist of the century on these principles.

1. Semiotic Principle. "A sign cannot be reduced to or analyzed into any combination of things which are not themselves signs". Goddard lists some examples of what meaning can*not* be decomposed to: reference or denotation, truth conditions, neurophysiological data, and usage. This principle is opposite to the goal of this

thesis that searches for connections between symbolic representations and distributional ones.

2. Decomposition into discrete terms without (circularity and) residue. Exhaustive analysis distinguishes NSM from *componential analysis* which attempts to capture only *systematic oppositions* or "Katz and Jackendoff, who both believe that for many words an unanalysable residue of meaning remains" (the Distinguishers, recall Section 2.3.1). Commitment to discrete terms distinguishes NSM from *scalar notations*, the topic of Chapter 7.

3. Semantic Primitives Principle. There exists a finite set of undecomposable meanings and semantic primitives have an elementary syntax whereby they combine to form 'simple propositions'. While this is a key point of the collection, `4lang` does not require the elements of the core/defining vocabulary to be primitive.

4. NSM approach. "The proper metalanguage of semantic representation is [...] a minimal subset of ordinary natural language." Goddard lists examples for positions taken in the literature regarding the problem of the (meta-)semantics of the representational elements.

    - Proposals which represent primitives by *obscure technical terms* like symbols borrowed from logic ($\exists, \forall$) or those like Schank's PACT, CACT and TACT that need explanation in ordinary English (e.g. 'physical act', 'communication act' and 'transfer act', respectively).
    - Predicates in generative semantics, like CAUSE, NOT, BECOME and ALIVE whose intended meanings were not (exactly) those of the English words, but were more 'abstract'
    - Katz (1987) uses semi-technical labels to identify the 'conceptual components' of e.g. the English word CHASE: 'Activity', 'Physical', 'Movement', 'Fast', 'Direction', 'Toward location of', 'Purpose', 'Catching'. It has to be made clear what we gain by formalization as opposed to natural syntax.

5. "The NSMs derived from various languages will [...] have the same expressive power."

6. The linguistic exponents of semantically primitive meanings in different languages can be placed into one-to-one correspondence (modulo differences like allolexy and POS membership), thus they share a common set of combinatorial properties.

7. Strong Lexicalization Hypothesis. Every semantically primitive meaning can be expressed through a word, morpheme or fixed phrase in every language. Exponents may be homonyms with different POSs or bound morphemes. Goddard follows Chomsky

> (1965) in distinguishing *formal universals* concerning the principles by which sense-components are combined to yield the meanings of lexemes from *substantive universals* concerning the identity of semantic components. The collection tests the thesis of the most extreme form of substantive universalism that "there is a fixed set of semantic components, which are lexicalized in all languages".

In the NSM approach, words (morphemes, etc.) can be *identical in meaning despite different POS*, ranges of use, or patterns of polysemy. Differences in the range of use does not invalidate the claim of semantic equivalence, as far as it is caused just by lexical blocking or social and cultural factors. The project has introduced *canonical contexts* to specify the sense of each polysemous words that should be used in explications, and the contexts in which the proposed meaning is expected to be found.

As admitted in the last chapter, the greatest problem with the NSM approach is polysemy as basic, everyday words are particularly likely to be polysemous because of Zipf's law. They require polysemy always to be justified on language-internal grounds, and to prove that a word is polysemous, one has to demonstrate that the putative senses call for distinct reductive paraphrase explications or syntactic frames (and distribution). We note that patterns of polysemy show similarities among languages (Youn et al. 2016).

Similarly to `4lang`, NSM has to reconcile the existence of language-specific morphosyntactic categories with the claim that the semantic metalanguage is isomorphic across languages, e.g. in the natural semantic metalanguage based on Latin, VOLO would never occur without an explicit subject. We will discuss this problem in Chapter 5.

Goddard and Wierzbicka (1994, Section 2.2) investigates The Proposed Primitive Inventory in the groups shown in Table 2. Albeit we are not interested in whether an element is primitive, it is useful to discuss how the core definitions in `4lang` handle the areas where these groups have proved indispensable in NSM.

2.3.4 *Force dynamics in language and cognition*

Talmy (1988) draws the attention to what he calls force dynamics: linguistic, psychological, and social phenomena related to physical ones, like the exertion of force, resistance to such exertion and the overcoming of such resistance, blockage of a force and the removal of such blockage, etc. Talmy offers a framework that also includes *letting*, *hindering*, *helping*. The theory builds upon the parallelisms between how we refer to physical and psychosocial matters.

In English, force dynamics is present in different grammatical categories: closed-class words (conjunctions, prepositions, modals), open-class lexical items, semantics of course (physical force psychological

Substantives	I, you, someone, something, people
Mental predicates	think, say, know, feel, want
Determiners/quantifiers	this, the same, other, one, two, many, all
Actions/events	do, happen
Meta-predicates	no, if, can, like, because, very
Time/place	when, where, after, before, under, above
Partonomy/taxonomy	have parts, kind of
Evaluators/descriptors	good, bad, big, small

Table 2: Primitives of Natural Semantic Metalanguage in groups.

and social interactions, psychosocial "pressures"), and discourse (patterns of argumentation, discourse expectations and their reversal). The theory brings these together into systematic relationships.

Talmy attributes his method to "cognitive semantics" or "cognitive linguistics", which analyzes the cognitive process and its surface linguistic realizations together. Force dynamics is among the fundamental notional categories that languages use to structure and organize meaning, while they exclude other notional categories from playing this role. For cognitive semantics, it is important, how the linguistic structuring relates to perceptual modalities and reasoning, space, time, and visual perception, or, in this case, physics and psychology. The paper goes from conceptually basic physics dynamics to psychological and social interactions, the grammatical category of modals, discourse factors (argumentation), and other cognitive and conceptual domains.

The simplest force dynamics model consists of the following:

- two forces, and an Agonist and an Antagonist. The salient issue is whether the Agonist is able to manifest its force.
- The Agonist is toward action or toward inaction. The Antagonist opposes the Agonist.
- The relative strengths of the Agonist and the Antagonist is a third parameter.
- The result is either action or inaction.

More complex force-dynamic patterns change through time: a stronger Antagonist can come in or go out, or the balance of forces can shift.

An additional kind of pattern is in which the Antagonist remains away. Corresponding to each of the steady-state patterns introduced so far, there is a secondary steady-state pattern with the Antagonist steadily disengaged. E.g. where the Antagonist is stronger, we have the patterns for the Antagonist *letting* the Agonist to move or rest.

There are alternatives of Foregrounding different subsets of the factors, e.g. making the Agonist, the Antagonist, or the result the grammatical subject or the object.

Examples with a weaker Antagonist: with the Agonist as the subject: *despite, although*, with the Antagonist as subject: *hinder, help, leave alone*.

*Psycho*dynamics generalizes notions of physical pushing and blocking to wanting and refraining; psychological 'pressure', and 'pushing'. The self may be divided to an Agonist and an Antagonist, where the Agonist represents the desires, and the Agonist's role is suppression. In language, this is extended to physical entities without sentience such as wind, a dam, or a rolling log. A psychological component is normally included and understood as the factor that renders the stronger participant. The body has an intrinsic tendency toward rest, requiring animation by the psyche.

Two additional factors are the *phase* along a temporal sequence, and 'factivity': the occurrence or non-occurrence of portions of the sequence and the speaker's knowledge about this. With the Antagonist as subject: *try* involves focus at the initial phase without knowledge of its outcome, while *succeed* and *fail* focus on a known occurring or non-occurring outcome.

The force dynamics in discourse (argumentation and expectations) is based on the metaphor of an *argument space*: each point can oppose or reinforce another point, and each encounter can move the argument state closer to or further from one of the opposing conclusions.

The last part of the paper compares conceptual models of physics implicit in language to the real physical theory. One great difference is the asymmetry between the privileged Agonist and the Antagonist so natural in language-based conceptualizing, which has no counterpart in physical theory. The real theory is based on objects' impetus in motion, while the naive theory assumes a tendency to come to rest. In modern physics, stationariness is not a distinct state but is simply zero velocity. In language either the Agonist or the Antagonist has greater relative strength, while in physics, two interacting objects must be exerting equal force. The linguistic expression of causation has a tripartite structure: a static prior state, a discrete state-transition, and a static subsequent state. This is based on the notion of an 'event': a portion conceptually partitioned out of the continuum of occurrence, which is autonomous, without causal processes during its occurrence. Blocking and letting, resistance and overcoming, some of the most basic force-dynamic concepts, have no principled counterpart in physics, because these concepts depend on the ascription of entityhood to a conceptually delimited portion of space, and the entity's intrinsic tendency toward motion or rest.

2.3.5 *Conceptual Structures*

"I think that an overview of the ideas about the nature of argument structures and the mechanisms that lead from semantic argument structures to syntactic arguments", which will be investigated in Chapter 5, "must not miss Ray Jackendoff's proposal, which evolved in many articles and books since cca. the mid-1980s; if only because the notion of the *lexical conceptual structure* to be distinguished from the semantic structure is also relied on by those who otherwise propose different mapping mechanisms from that of Jackendoff (e.g. the Levin–Rappaport pair). Jackendoff (1990) is a relatively early (and, thankfully, fairly easy to understand) review of his views. You don't have to 'learn' this, but getting to know the basic ideas (it is enough to just go through the first 60 pages) can, in my opinion, get everyone to rethink new perspectives" (András Komlósy, personal communication, translated from Hungarian by thesis author).

"Building on ideas about semantics first expounded by Gruber (1965), Jackendoff (1972, 1983) elaborated significantly on the notion of cases by treating them as arguments to a set of *primitive conceptual predicates* such as GO, BE, STAY, LET, and CAUSE." (Palmer, Gildea, and Xue 2010)

GO can be used to describe changes of location, possession, or state, in any situation where both a "before state" and a different "after state" can be defined. It basically takes three arguments, the object undergoing the change and the before and after locations, possessors, or states. (`4lang`, the semantic representation framework we introduce in the next chapter, does not use and explicit GO predicate, but it shares CAUSE, BEFORE, and AFTER with CS.) Later versions introduced subtypes of primitive predicates that add more information, e.g. the manner of a motion. Jackendoff 's intent was not to provide detailed representations of all of meaning but, to focus on the *mapping* between syntax and semantics. The remainder of this section discusses the theory based on Jackendoff (1990).

2.3.5.1 *Ontological categories or conceptual parts-of-speech*

Instead of a division of formal entities into logical types like constants, variables, predicates, and quantifiers, the theory of Conceptual Structures (CS) sorts constituents to a few major ontological categories (or conceptual parts-of-speech) like Thing, Event, State, Action, Place, Path, Property, and Amount.

Each major syntactic constituent maps into a conceptual constituent: NP correspond to Thing-constituents, the PP to a Path-constituent, and the entire sentence to an Event. The converse of this correlation does not hold, e.g. many conceptual constituents of a sentence's meaning are completely contained within lexical items. The mapping between conceptual and syntactic categories is many-to-many but it is

subject to markedness conditions. Each conceptual constituent has an argument structure feature, which allows for recursion of conceptual structure and hence an infinite class of possible concepts.

2.3.5.2 *Localism*

A second cross-categorial property of conceptual structures goes back to the *localistic* theory. The formalism for encoding concepts of spatial location and motion can be abstracted/generalized to many other semantic fields. Many verbs and prepositions appear in more semantic fields and in intuitively related paradigms.

Many implicative properties of verbs (such as *factive*, *implicative*, and *semifactive*) follow from generalized forms of inference rules developed to account for verbs of spatial motion and location. Each semantic field has its own particular inference patterns, e.g. in the spatial field, one fundamental principle stipulates that an object cannot be in two disjoint places at once. It follows that an object that travels from one place to another is not still in its original position. In the field of information transfer, this inference does not hold. A similar conceptual structure may apply to different parts-of-speech, as exemplified by the parallelism between the iteration of actions and the plural of things, or the bounded/unbounded distinction among verbs (event/process, telic/atelic) and the count/mass distinction among nouns.

2.3.5.3 *Preference Rule Systems*

CS involves something similar to prototype theory or fuzzy set theory: Verbs have more "fuzzy truth conditions": climb = move up & grasp, see = gaze & realize. An event which satisfies both conditions at once, is more *stereotypical*. An example from an other part of speech is nouns that denote form and function as two conditions (e.g. *book*). When one lacks information about the satisfaction of the conditions, they are invariably assumed to be satisfied as default values.

2.3.5.4 *Argument Structure and Thematic Roles*

THE STATUS OF THEMATIC ROLES CS has a notion of thematic roles which has greatly influenced `4lang`. In Jackendoff (1990)'s approach, thematic roles are structural configurations in CS (See his Section 2.2.).

```
DO(John,CAUSE(HAVE(Bill,book)))
```

E.g. the traditional Source/Goal, "the object from/to which motion proceeds", can be structurally defined as the argument of the Path-function FROM/TO. Agent is the first argument of the Event-function CAUSE, and Experiencer is an argument of some function having to do with mental states.

The list of a verb's arguments can be constructed simply by extracting the indices from the verb's lexical conceptual structure. The hierarchy of thematic roles is "cca. provided" by the relative depth of the embedding of the indices in the conceptual structure. Each kind of argument position plays a distinct role in rules of inference.

Not only NPs receive thematic roles. For instance, *green* is a Goal in *The light changed from red to green.*, and *shut up* is a Goal in *Bill talked Harry into shutting up.*, not the thematic role for a subordinate clause, as suggested in Lexical Functional Grammar. Clauses can occur in various thematic roles, just as Things can.

There's no "default" thematic role in the sense that Objective is "default" or "neutral" in Fillmore (1968): in CS, an NP must correspond to a specific argument position in conceptual structure and therefore must have a specific thematic role. Even *Theme* or *Patient*, which have been taken to be such a default role, have a specific structural definition.

ARGUMENT FUSION AND SELECTIONAL RESTRICTIONS In CS, and similarly in `4lang`, a verb's lexical representation can include information about a participant which is not even syntactically expressed. In order for a sentence to be understood, this fine CS must exist. Selectional restrictions are explicit pieces of information that the verb supplies about its arguments. Formally, they correspond to the conceptual structure that occurs within an indexed conceptual constituent.

CS is a unification-based system: if two conceptual structures contain incompatible information, (if the offending features are sisters in a taxonomy of mutually exclusive possibilities, such as Thing/Property/Place/Event/etc. or solid/liquid/gas) their fusion is anomalous. `4lang` does not implement such hard constraints.

E.g. the transitive verbs *drink* and *butter* both mean "cause something to go someplace". They differ semantically in what they stipulate about the Theme and the Path. The direct object of *butter* is the Goal, and the Theme is completely specified by the verb, while the direct object of *drink* is the Theme, and the Path is (almost) completely specified by the verb. It is part of the meaning of *order* that the recipient (or Goal) of an order is under obligation to perform the action described by the complement clause, and that of *promise* that the issuer (or Source) of a promise undertakes an obligation to perform the action described by the complement.

CS has many ways of expressing conceptual structure within arguments of the verb (which is part of the verb's meaning): the positions of the indices (which is analogous to `4lang`'s deep cases, i.e. the way the verb links its arguments to syntactic structure), selectional restriction, and and implicit arguments.

MULTIPLE THEMATIC ROLES FOR A SINGLE NP Chapter 3 of Jackendoff (1990) investigates the q-Criterion, i.e. that each subcatego-

rized NP (plus the subject) corresponds to exactly one argument position in conceptual structure, and that each open argument position in conceptual structure is expressed by exactly one NP. In Jackendoff's view, the q-Criterion must be weakened, e.g. because of transaction verbs such as *buy, sell, exchange*, and *trade*, where there are two giving actions (that of the merchandise and the money), and the seller and the buyer have two semantic roles apiece; or *chase*, where both the Agent and the Patient move. We will see that deep cases in `4lang` are closer to the surface: *buy* has an agentive subject, while its source is unspecified for animacy, even if it gives money voluntarily. In contemporary computational systems, we can assume a sentence analyzed for syntactic dependency, and the task of deep cases is to mediate between the dependency annotation and the semantic representation.

UNIFYING LEXICAL ENTRIES Chapter 4 in Jackendoff (1990) investigates argument structure alternations, where the alternatives can be captured by the same lexical entry. `4lang` goes the same path. Optional modifiers (of place, time, and manner) are not encoded anywhere in the lexical entry. The problem of causatives *(The box slid/Bill slid the box down the stairs)* is solved following the Unaccusative Hypothesis.

A more special example is *climb* with three syntactic frames: null complement, direct object, or PP. CS wants to account for the difference that only the direct object entails that the subject reaches the top. `4lang` disregards such differences, not in order to codify such a coarse level of mental representation, but as an engineering shortcut. More productive lexical processes, e.g. passive participles from verbs can be expressed in terms of manipulations on the argument indices. In `4lang`, passives are already handled by the dependency parse. Jackendoff also discusses verbs with some spatial feature in their meaning (*point, surround, cover, support*) which would go beyond the limits of the present thesis.

SOME FURTHER CONCEPTUAL FUNCTIONS Section 5.2 in Jackendoff (1990) investigates verbs of manner of motion like *curl, writhe*, or *dance*. These are less interesting for our present purposes, as the working method of `4lang` is to define manually only some defining vocabulary, which can be used to define all other words automatically.

While this topic is beyond the scope of the present thesis, we quote some ideas by Jackendoff on *conceptual clause modification*. Jackendoff offers a partial taxonomy of functions that convert a State or Event into a restrictive modifier of another State or Event (syntactically: subordinating conjunctions that turn sentences into restrictive modifiers).

- Cause (why?) has logically two types: reason, represented with FROM, a variant of the usual FROM; and purpose, goal, or rationale

(the intention may be the speaker's or attributed to the Agent), represented with FOR, a variant of TO or TOWARD.

- In accompaniment *(Bill came with Harry)* there is a mutual dependence between Bill's coming and Harry's, and Bill is "foregrounded". This asymmetrical relation is "more than conjunction but less than causation".
- Exchange, reward or punishment are voluntary acts of social cognition, based on assessment in legal and economic systems, which is worth a separate satus in cognitive semantics.

More of these subordinators are similar to spatial functions both in their morphology and the inferences associated with them. Cross-linguistic study is important here, of course: if the same apparently idiosyncratic fact appears in language after language, something is being missed. Conversely, if an apparently principled English fact is violated in other languages, the principle must be questioned.

FEATURAL ELABORATIONS OF SPATIAL FUNCTIONS Jackendoff aims at a featural decomposition of verb meaning. E.g. he introduces a feature opposition in spatial location, say Location versus Contact (or ±contact) which is present in the prepositional system, where *on* and *against* contrast with *in, next to, alongside, above*, and in the verb lexicon, where *stroke*, *scratch*, *rub*, and *brush*, unlike some other verbs, specify motion while in continuous contact with the object. `4lang`, in contrast, tries to capture words with other words instead of features.

THE ACTION TIER AND THE ANALYSIS OF CAUSATION Section 7.1 in Jackendoff (1990) decomposes thematic roles to two dimensions: The Action Tier distinguishes the Actor and Patient, while the thematic tier (Theme, Source, and Goal) deals with motion and location. Thus *What happened to* Pat*?* or *What did* Agt *do to* Pat*?* is orthogonal to *What moves where?*.

2.3.6 *English Verb Classes and Alternations*

As another source of semantic knowledge, Levin (1993) points out that the expression and interpretation of arguments is to a large extent determined by the verb's meaning. The introduction of the book exemplifies this with *break, cut, hit,* and *touch.* Each verb shows a distinct pattern with respect to three alternations, the middle alternation (*This bread cuts easily.*), the conative construction (*cut at*), and the body-part alternation (*Margaret cut Bill on the arm.*). There are other verbs that show the same pattern of behavior: Break Verbs: *break, crack, rip, shatter, snap*, Cut Verbs: *cut, hack, saw, scratch, slash*, Touch Verbs: *pat, stroke, tickle, touch*, and Hit Verbs: *bash, hit, kick, pound, tap, whack.*

Levin's analysis is based on relevant meaning components. The body-part possessor ascension alternation needs 'contact', while the conative alternation needs both 'motion' and 'contact'. *Touch* is a pure verb of contact, *hit* is a verb of contact by motion, *cut* is a verb of causing a change of state by moving something into contact, and *break* is a pure verb of change of state. This explains which verb participates in which alternation.

These phenomena are manifested across languages by verbs of the same semantic types. To the extent that languages are similar, the same meaning components – and hence the same classes of verbs – figure in the statement of regularities concerning the expression of arguments. The classes have a range of properties in common, including the possible expression and interpretation of their arguments, and the existence of certain morphologically related forms.

The meaning component analysis is related to "semantic bootstrapping" models of child language acquisition built on the assumption that a word's syntactic properties are predictable from its meaning. Meaning components identified via the study of semantic/syntactic correlation show considerable overlap with those posited in language acquisition.

Levin investigates intricate and extensive patterns of syntactic behavior: the subcategorization frame of a verb, diathesis alternations, morphological properties and extended meanings.

Part I of the book introduces diathesis alternations that are relevant to lexical knowledge, subdivided into groups on the basis of the syntactic frames involved: transitivity alternations, alternate expressions of arguments (mostly within the verb phrase), alternations that permit "oblique" subjects, and a variety of other types. Part II presents a large number of semantically coherent classes of verbs[4]. Levin tries to strike a balance between breadth and depth of coverage. He ignores verbs taking sentential complements except when they show interesting behavior with NP or PP complements; verbs derived by productive morphological processes, such as zero-derivation, prefixation (*un-*, *de-*, *dis-*, *re-*, etc.) or suffixation (*-ify*, *-ize*, *-en*, etc.); and inherent lexical aspect of verbs (aktionsart). It is left as an open research question whether a complete *hierarchical* organization of English verb *classes* is possible or even desirable.

4 Put; Remove; Send and Carry; Exert Force: Push/Pull; Change of Possession; Contribute; Learn; Hold and Keep; Concealment; Throw; Contact by Impact; Hit; Poke; Contact: Touch; Cut; Combine and Attach; Separate and Disassemble; Color; Image Creation; Illustrate; Creation and Transformation; Engender; Calve; Verbs with Predicative Complements; Perception; Psych-Verbs (Psychological State); Desire; Judgment; Assessment; Search; Social Interaction; Communication; Sounds Made by Animals; Ingest; Involve the Body; Groom and Bodily Care; Kill; Emission; Destroy; Change of State; Lodge; Existence; Appearance, Disappearance, and Occurrence; Body-Internal Motion; Assume a Position; Motion; Avoid; Linger and Rush; Measure; Aspectual Verbs; Weekend; Weather

2.3.7 *The generative lexicon*

Both traditional and most computational lexicons (the latter will be discussed in Section 2.4) tend to have very fine-grained sense distinctions, and the relations between different senses are mostly not represented. Pustejovsky (1995) call these resources *sense enumeration lexicons (SELs)*, and proposes the *generative lexicon (GL)* as an alternative, where lexemes have richer structure, and the virtually infinite semantic types a lexeme may have arise in context, by co-composition with the similarly flexible representations of other words, similarly to how infinitely many sentences are generated from a finite lexicon by recursive generative grammars in syntax. While the core of the GL is organized among semantic types, and is thus less interesting in the context of `4lang`, the theory has many features worth studying from our more association-based point of view as well.

GL builds on *a classification of word polysemy* to homonymy and polysemy proper, or, in Weinreich (1964)'s terms, contrastive and complementary ambiguity. *Contrastive ambiguity (i.e. homonymy)* is the coincidence of unrelated meanings, while *complementary ambiguity (polysemy)* refers to logically related word senses, manifestations of the same basic meaning in different contexts, possibly different parts of speech. Whether homonymic senses are historically related or accidents of orthographic and phonological blending, is largely irrelevant for the purposes of lexicon construction and the synchronic study of meaning. The two types of ambiguity also differ in whether the disambiguation of co-occurring words help each other: homonymy works so that once the context or domain for one item has been identified, the ambiguity of the other items is also constrained (contextual priming). This does not hold for sense narrowing in polysemy, where one sense may be entailed by the other sense. Pustejovsky mentions classes of (complementary) polysemy where the senses correspond to different semantic types like Count/Mass (*lamb*), Container/Containee (*bottle*), Gap/Frame (*door, window*), Product/Producer (*newspaper, Honda*), Plant/Food (*fig, apple*), Process/Result (*examination, merger*), Place/People (*city, New York*), and Change-state/Create (*bake*).

Pustejovsky (1995, Section 3) goes further to define *logical polysemy* as a complementary ambiguity where there is no change in lexical category, and the multiple senses of the word have overlapping, dependent, or shared meanings.

Pustejovsky lists three arguments showing *the inadequacies of SELs for semantic description* and that to maintain compositionality, we must enrich the representations of the lexical items:

- *The Creative Use of Words*, that words assume new senses in novel contexts,

- *The Permeability of Word Senses*, that Word senses are not atomic definitions but overlap and make reference to other senses of the word, and
- *The Expression of Multiple Syntactic Forms*, that a single word sense can have multiple syntactic realizations.

GL involves four *levels of lexical representation:* argument, event, qualia, and inheritance structure.

Argument structure specifies the number and type (semantic and syntactic) of arguments a predicate takes. This is by far the best understood of the four levels in generative linguistics (e.g. Chomsky's Theta-Criterion, Lexical Functional Grammar (Bresnan 1978, 2001)), and argument structure is also the strongest determinant or constraint on the acquisition of verb meaning by children. Pustejovsky distinguishes four types of arguments (illustrated for verbs), *true* obligatory arguments subject to the theta criterion; *default* arguments that are necessary for the logical well-formedness of the sentence, but may be left unexpressed on the surface; *shadow* arguments, e.g. incorporated semantic content (the instrument of *kick* or *butter*); and *(true) adjuncts* that are associated with verb classes and not with the representation individual verbs, including temporal or spatial modifiers. The categorization of arguments induces a corresponding categorization of verb alternations as well: those that result in the expression of true arguments versus which involve the expression of optional ones.

Event structure is for the representation of information related to Aktionsarten and event type, in the sense of Vendler (1967): event type (state, process, and transition) and subeventual structure. Besides the relation between an event and its subevents, GL involves overlap and inclusion of subevents as well, and one of the subevents may be the *head* of the event. In `4lang`, as we already mentioned in Section 2.3.5, there are three potential subevents, the unmarked (present) one, the one represented under an `after` node, and represented under `before`.

Qualia structure is the set of properties or events associated with a lexical item which best explain what that word means, such as its constituent parts, purpose and function, mode of creation, etc. More formally, these aspects are

- CONSTitutive, the relation between an object and its constituent parts (e.g. "text in a novel is characteristically a narrative or story, while a dictionary is by definition a listing of words"), its material, and also what this object is part of.
- FORMAL: orientation, magnitude, shape, dimensionality, color, position;
- TELIC: purpose and function, how we use a thing, or the purpose that an agent has in performing an act. Direct TELIC, e.g. beer is

$$\begin{bmatrix} \textit{novel} & \\ \text{QUALIA} = & \begin{bmatrix} \text{CONST} & = \textit{narrative} \\ \text{FORMAL} & = \textit{book} \\ \text{TELIC} & = \textit{reading} \\ \text{AGENT} & = \textit{writing} \end{bmatrix} \end{bmatrix}$$

Figure 4: Qualia structure of *novel*.

made in order that it will be drunk, is distinguished from instruments, e.g. knives are made to cut with them; and finally,

- AGENTive specifies how things come into being, a mode of explanation that will distinguish natural kinds from artifacts (e.g. cookies, cakes, and bread are typically baked);

see Figure 4. The model is inspired in part by Moravcsik (1975)'s interpretation of Aristotle's modes of explanation. The qualia structure plays an important role in how we understand sentences, e.g. by knowing that the TELIC of *movie* is *watch*, we understand *John enjoyed the movie* in the way that he enjoyed *watching* it.

The last one of the four levels, inheritance identifies how a lexical structure is related to other structures in the type lattice.

The four levels are connected by generative devices providing for the compositional interpretation of words in context. Though, unlike `4lang`, GL is a strongly typed model, and these devices (e.g. type coercion and shifting, selective binding and co-composition) play the role of fitting items in novel type environments, the basic idea that predicate-argument binding can refer to subevents in the semantic representation is a feature shared with `4lang`.

While Pustejovsky criticizes the lexical semantic literature for overemphasizing the role of verbs, their classes and alternations, he also devotes a chapter to this topic, more concretely causation. The point is that members of alternations, e.g. the transitive and the intransitive variant of a verb, are generated from the same item in GL.

2.4 MODERN LEXICAL RESOURCES

> Usage seems to be inversely proportional to representational complexity. — *(Russell and Norvig 2002)*

The final section of this chapter mostly introduces modern lexical resources, which serve as the basis of any kind of supervised NLP research. Every experiment reported in the main part of the thesis relied on one of them.

2.4.1 *Computational lexicography for NLP*

While it is neither especially modern nor a lexical resource, we start with a reflection on the Introduction chapter of Boguraev and Briscoe (1989), because it is very closely related to the methods used in the `4lang` project, the semantic network that we will introduce in the next chapter. This book dates from the dawn of corpus linguistics, and the chapter discusses lexical resources, both their theoretical role and applications in traditional linguistics and NLP-based systems. This book analyses *The Oxford Advanced Learner's Dictionary of Contemporary English (LDOCE)*, which is also important for `4lang`, because both our hand-written definitions and automatically extracted representations heavily relied on it. Traditional lexicons contain tens or hundreds of thousands of lexical items, and computational lexicography and lexicology have developed disciplines with their own workshops and conferences. While NLP has established new lexical knowledge bases (KBs) for a wide variety of researchers and applications, reusing existing lexical resources offers further room for improvement. Machine readable dictionaries (MRDs) represent a considerable tradition where much work has already been done, however difficulties arise because these resources are produced for human use, and they may make inconvenient assumptions, and rely on the users' linguistic and common sense knowledge which machines do not have. The book has made a great influence on `4lang`, as both lines of research strive to make information in MRDs accessible for machine use, and evaluate and improve computational semantic systems and linguistic theories based on these resources. The decades since Boguraev and Briscoe (1989) have proven that lexicons derived from MRDs for machine use are different from conventional dictionaries in how they organize and represent information, but the same dictionary database (DB) can be used for both automated and human use. Some reoccurring themes of the book are the division between lexical semantics and pragmatic knowledge, the border between rules and the lexicon, and the acquisition of POS and subcategorization information with syntactic features.

2.4.1.1 *The nature of a dictionary entry*

In Boguraev and Briscoe (1989)'s view on the lexicon vs. rules division, a general-purpose dictionary DB should be as inclusive and theoretically uncommitted as possible. E.g. either one assumes a rule of *re-* prefixation or one needs to list elements like *reissue, reclaim* and *repay*.

The entries in most dictionaries distinguish 'homographs' of a word form when it serves as noun, verb or some other POS. Entries start with the form (headword, spelling, hyphenation, phonetic variants, allomorphs, stress) and information on the distributional behaviour (either with a simple word class tag, e.g. in The Collins English Dictionary,

or with elaborate subcategorization information, e.g. in LDOCE, or in The Collins COBUILD English Language Dictionary).

Regarding the content, dictionaries tend to provide definition(s), examples, cross references; grammar and stylistics of usage; synonyms, antonyms, related words; a picture, etymology; and derived words, compound terms, idiomatic or common phrases, expressions and collocations. LDOCE also provides semantic notions in the form of so called *subject* and *box* codes, which specify the semantic field (e.g. politics, religion, language) and selectional restrictions (e.g. the verb *sandwich* prefers an abstract or human subject). The language of dictionary definitions tends to be of a restricted form. In LDOCE, the vocabulary is restricted to approximately 2200 words used mainly in their most common sense, which theoretically would cut down circularities (but see the next paragraph). Unfortunately, derivational morphology is applied to these words in a rather liberal way. Representation is made difficult by the fact that there is a continuum between the minimal semantic knowledge implied by the use of a particular word (word sense) and the special (or expert) knowledge relevant to its use in the context of a specific domain.

2.4.1.2 *Reliability and utility of MRDs*

The preface to the published version of the Longman Defining Vocabulary (LDV) claims that 'a rigorous set of principles was established to ensure that only the most 'central' meanings of [a controlled vocabulary of] 2000 words, and only easily understood derivatives, were used'. 'Body' is part of the definitional vocabulary and has as its central (1) meaning "the whole of a person". However, Boguraev and Briscoe (1989) point out that *parliament* is defined as "a law-making body", utilizing the meaning of body (5) "a number of people who do something together". To make things worse, about 30 non-LDV words are used in definitions, e.g. *aircraft* is used 267 times.

Besides the already mentioned liberal use of derivatives ('container' is used for the definition of box2(1), even though only the verb *contain* is considered to be primitive), circularity (*container* ↔ *box*) also arises. Another related problem is the use of phrasal verbs made up from verbs and particles taken from the restricted vocabulary but, of course, with a non-compositional meaning.

In another chapter of Boguraev and Briscoe (1989), Vossen, Meijs, and Broeder (1989) derive a syntactic typology for the structures of the meaning descriptions of each of the major parts-of-speech (POS) in a dictionary. The typology combines hyponyms and adjectives, with subject field, speech register, and sociolect codes.

2.4.1.3 *Connectionism, word ambiguity, and knowledge*

The final chapter of the book, (Wilks et al. 1989) investigates the relation between connectionism and word ambiguity. The authors realize that connectionism shares properties with compositional semantics, and they do not expect to distinguish representations for particular word senses, but to be simply different aspects of a single non-symbolic representation, and to correspond (if to anything) to a selection of different weighted arcs. They advocate weighted symbolic representations. This view applies to issues of word sense for compositional semantics (discreteness of word senses vs. continuity and vagueness).

The position in the chapter is that the inseparability of knowledge and language goes far, and knowledge for certain purposes should be stored in text-like forms. The authors compare the semantic structure of dictionaries to the underlying organization of knowledge representations, and observe similarities: computational semantics converges with knowledge acquisition and computational lexicography. The chapter investigates whether it is right to assume the notion of a word 'sense' directly from traditional lexicography and MRDs. (The answer is yes.) Another question is whether a dictionary is a strong enough *knowledge base*. Not directly, but its content can be made explicit by additional information. Collecting the initial information (bootstrapping) is needed from the dictionary itself or some external resource.

2.4.2 *Frame semantics*

Jurafsky (2014) introduces *frames* as a rather general representation that expresses the background contexts or perspectives by which a word or a case role could be defined. The name came from the pre-transformationalist (1974) view of sentence structure as consisting of a frame and a substitution list. Frames were also called *scripts* or *schemata*.

In Kornai (2008, Section 5.3)'s reflection, the original intention was to use scripts as repositories of commonsense procedural knowledge: what to do in a restaurant, what happens during a marriage ceremony, etc.; represent the actors fulfilling roles, e.g. that of the waiter or the best man; and decompose the prototypical action in a series of more elementary sub-scripts such as 'presenting the menu' or 'giving the bride away'. Kornai relates scripts to "linguistically better motivated models", in particular discourse representation theory, whose scope is more modest, being concerned primarily with the introduction of new entities (the owner, the best man). Scripts have also influenced studies of rituals.

Turning to Jurafsky (2014)'s account of verbal case frames, Fillmore was also inspired by lists of slots and fillers used by early information extraction systems, but his version of this idea was more linguistic. The

motivating example was the Commercial Event frame *(buy, sell, cost, pay, charge)*. Frames could represent perspectives on events, e.g. *sell* vs *pay*. Alternative senses of the same word might come from their drawing on different frames. The perspective-taking aspect of frame semantics influenced framing in linguistics and politics.

2.4.3 *WordNet*

Probably the most popular lexical NLP resource is the (English Princeton) WordNet (Miller 1995). As we will discuss Hungarian as well in Section 7.4, the Hungarian WordNet (Miháltz et al. 2008)[5] has to be mentioned as well. WordNet follows the lexicographic tradition of treating POSs separately, and words are grouped by semantic equivalence to 117 000 *synsets* with a definition ("gloss") each, and, in most of the cases, sentences illustrating the use of the words in the set. WordNet disambiguates word forms to many senses (synsets) to account for fine distinctions in their usage. This opposes to the monosemic approach `4lang` follows, see our discussion in Kornai and Makrai (2013) as well. An aspect of WordNet which is more instructive for `4lang` is its inventory of binary relations.

Palmer, Gildea, and Xue (2010, Section 1)'s introduction to linguistic theories and semantic representations of roles "ends where it began, with Charles Fillmore". In this and the following two sections, we introduce a couple of verb-related resources. Palmer, Gildea, and Xue (2010, Section 2) describe these resources as having differing goals, and yet being surprisingly compatible. They differ primarily in the granularity of the semantic role labels. FrameNet labels the arguments of *approve* as Grantor and Action. PropBank uses very generic labels such as Arg0, Arg1, ... VerbNet, on the third hand, has several alternative syntactic frames and a set of semantic predicates. VerbNet marks the PropBank Arg0 as an Agent and the Arg1 as a Theme. The three resources can be seen as complementary.

2.4.4 *FrameNet*

Based on Fillmore's Frame Semantics (Section 2.4.2), FrameNet (Baker, Fillmore, and Lowe 1998) describes a particular situation or event along with its participants. Semantic roles are called *Frame Elements (FE)*, and they are defined for each semantic frame. The predicate is called *Lexical Unit (LU)*. All LUs in a semantic frame share the same set of FEs. FEs are fine-grained semantic role labels, e.g. the Apply-heat Frame includes a Cook, Food, and a Heating Instrument.

A frame can also have adjectives and nouns such as nominalizations. FEs are classified in terms of how central they are: core (conceptually

5 https://github.com/dlt-rilmta/huwn

necessary for the Frame, roughly similar to syntactically obligatory), peripheral (such as time and place; roughly similar to adjuncts) or extra-thematic (not specific to the frame and not standard adjuncts but situating the frame with respect to a broader context, e.g. *The cottage still looks very much the same* ***from the outside***. (Ruppenhofer et al. 2006)).

Lexical items are grouped together without consideration of similarity of syntactic behavior, resulting in rich, idiosyncratic descriptions. E.g. *buy* and *sell* both belong to the semantic frame 'Commerce_buy', which involves a Buyer and Seller exchanging Money and Goods. Buyer and Goods are core FEs for this frame while Seller and Money are Non-Core FEs. Other Non-Core FEs include Duration (the length of time the Goods are in the Buyer's possession), Manner, Means, Place, Rate, and Unit, the unit of measure for the Goods.

2.4.5 *VerbNet*

VerbNet (Kipper et al. 2008) is midway between PropBank and FrameNet in lexical specificity, but it is more similar to PropBank with its close ties to syntactic structure. VerbNet consists of hierarchically arranged verb classes, extended from the Levin classes (see Section 2.3.6): Levin has 240 classes, with 47 top level classes and 193 second and third level. Original Levin classes constitute the first few levels in the VerbNet hierarchy, with each class subsequently refined. VerbNet has added almost 1000 lemmas as well as 200 more classes. There is now a 4th level of classes and several additional classes at the other three levels.

VerbNet adds to each Levin class an abstract representation of the syntactic frames with explicit correspondences between syntactic positions and the semantic roles (e.g. *break*: Agent REL Patient, or Patient REL *into pieces*). An argument list in VerbNet consists of semantic roles (Agent, Patient, Theme, Experiencer, etc., 24 in total), and selectional restrictions on the arguments, expressed using binary predicates that describe the participants during *stages of the event*.

VerbNet has class-specific interpretations of the semantic roles; 3,965 verb lexemes with 471 classes; links to similar entries in WordNet, OntoNotes groupings, FrameNet, and PropBank; and coherent syntactic and semantic characterization of the classes, which facilitate the acquisition of new class members.

Each VerbNet class contains a set of *syntactic frames*. Constructions such as transitive, intransitive, prepositional phrases, resultatives, and a large set of diathesis alternations listed by Levin are represented by the corresponding semantic roles (such as Agent, Theme, and Location), the verb, other lexical items required for a construction or alternation, and semantic restrictions (such as animate, human, and organization). Syntactic Frames specify which prepositions are allowed, and the syn-

tactic nature of the constituent (NP, PP, finite and nonfinite sentential complements).

Semantic predicates denote the relations between participants and events in the form of a conjunction of semantic predicates, such as motion, contact or cause, and START(e), END(e) and DURING(e), to indicate when the semantic predicate is in force.

2.4.6 *PropBank*

PropBank consists of an annotated corpus (to be used as training data) and a lexicon. Semantic role labels are chosen to be quite generic and theory neutral, Arg0, Arg1, etc. The same semantic role is kept across syntactic variations. The lexicon lists, for each broad meaning of each annotated verb, its frameset, i.e. the possible arguments in the predicate and their labels (its "roleset"), all possible syntactic realizations, and a set of verb-specific guidelines for annotators. PropBank is similar in nature to FrameNet and VerbNet although it is more coarse-grained, and more focused on literal meaning – as opposed to metaphorical usages and support verb constructions – than FrameNet.

PropBank defines semantic roles on a verb-by-verb basis:

- Arg0 is generally a prototypical Agent (Dowty 1991) while
- Arg1 is a prototypical Patient or Theme.
- There are no consistent generalizations for the higher numbered arguments, e.g. Arg2 can be beneficiary, goal, source, extent or cause.
- There are several more general ArgM (Argument Modifier) roles that can apply to any verb, and which are similar to adjuncts, e.g. LOCation, EXTent, ADVerbial, CAUse, TeMPoral, MaNneR, and DIRection.

These generic labels make high inter-annotator agreement possible. A roleset corresponds to a distinct usage of a verb. It is associated with a set of syntactic frames, the frameset.

There is a verb-specific descriptor field for each role, such as *baker* for 'Arg0' in *bake*, for use during annotation and as documentation, without any theoretical standing. The neutral, generic labels facilitate mapping between PropBank and other more fine-grained resources such VerbNet and FrameNet, as well as Lexical-Conceptual Structure or Prague Tectogrammatics.

Most rolesets have two to four numbered roles, but as much as six can appear, in particular for certain verbs of motion. PropBank lacks selectional restrictions, verb semantics, and inter-verb relationships.

Verb-Specific labels have their limitations. Inter-verb labels make inferences and generalizations based on role labels possible, because some

encoded meaning is associated with each tag, which helps in training automatic semantic role labeling systems. Researchers using PropBank as training data for the most part ignore the "verb-specific" nature of the labels, and instead build a single model for each numbered argument. This is feasible, because Arg0/Arg1 constitute 85% of the arguments, ArgMs are also labeled quite consistently. Arguments Arg2-Arg5 are highly overloaded, and performance drops significantly on them.

2.4.7 *ConceptNet*

The most relevant comparison for `4lang`, the basically word-level meaning representation framework we will introduce in the next chapter, is ConceptNet (Liu and Singh 2004). ConceptNet is a knowledge graph, i.e. it connects words and phrases with labeled edges. It is designed to represent the general knowledge involved in understanding language in the form of relations between words such as *'A net is used for catching fish'; '**Leaves** is a form of the word **leaf**'; 'The word **cold** in English is **studený** in Czech';* or *'O alimento é usado para comer'*, i.e. 'Food is used for eating'. In version 5.5 (Speer, Chin, and Havasi 2017), this piece of knowledge has been collected from many sources that include expert-created resources, crowd-sourcing, and games with a purpose.

The authors combine ConceptNet with word embeddings (Section 4.2) to get understanding that they would not acquire from distributional semantics alone, nor from narrower resources such as WordNet or DBPedia. The word embedding has been trained using a generalization of the *retrofitting* method (Faruqui et al. (2015), see Section 4.2.10). They demonstrated results on (i) intrinsic evaluations of word relatedness, which was a popular way of evaluating word embeddings before the introduction of contextualized word representations (Section 4.3), and on (ii) applications of word vectors, including solving SAT-style analogies.

In the remainder of this section, we describe the ConceptNet representation based on Speer and Havasi (2012, Section 3). Assertions in ConceptNet can be seen as edges that connect the concepts (words or phrases) corresponding to its nodes. Assertions can be justified by other assertions, knowledge sources, or processes. Predicates (i.e. edge labels) can be interlingual relations, such as *IsA* or *UsedFor* (see Table 3); or automatically-extracted relations that are specific to a language, such as *is known for* or *is on*. Processes that read knowledge from free text, will produce relations that are not aligned with multilingual relations. These unaligned relations specify the language and a normalized form, e.g. *A bassist performs in a jazz trio* translates to a `/c/en/perform_in` relation.

Negation in ConceptNet is a bit tricky. Conjunctions of assertions come with a positive or negative score, where a negative weight means we should conclude that the assertion is not true. The negation of a

Relation	Sentence pattern
IsA	NP is a kind of NP.
UsedFor	NP is used for VP.
HasA	NP has NP.
CapableOf	NP can VP.
Desires	NP wants to VP.
CreatedBy	You make NP by VP.
PartOf	NP is part of NP.
Causes	The effect of VP is NP|VP.
HasFirstSubevent	The first thing you do when you VP is NP|VP.
AtLocation	Somewhere NP can be is NP.
HasProperty	NP is AP.
LocatedNear	You are likely to find NP near NP.
DefinedAs	NP is defined as NP.
SymbolOf	NP represents NP.
ReceivesAction	NP can be VP.
HasPrerequisite	NP|VP requires NP|VP.
MotivatedByGoal	You would VP because you want VP.
CausesDesire	NP would make you want to VP.
MadeOf	NP is made of NP.
HasSubevent	One of the things you do when you VP is NP|VP.
HasLastSubevent	The last thing you do when you VP is NP|VP.

Table 3: The interlingual relations in ConceptNet, with example sentence frames in English. Table from (Speer and Havasi 2012)

conjunction with a high negative score is not necessarily true either: it may in fact be nonsensical or irrelevant. To represent a true negative statement, such as *Pigs cannot fly*, ConceptNet 5 uses negated relations such as `/r/NotCapableOf`.

2.4.8 *Abstract Meaning Representation for Sembanking*

Now we turn to one of the most popular meaning representation frameworks, Abstract Meaning Representation (AMR, Banarescu et al. (2013)). The original paper illustrates the AMR method with a syntactic analogue. Syntactic *treebanks* have had tremendous impact on natural language processing. *Whole sentence parsing* unified separate tasks (e.g. base noun identification) and their evaluations. Now smaller tasks are naturally solved as a by-product of whole-sentence parsing, and in fact, they are solved better than when they used to be approached in isolation. By contrast, a decade ago semantic annotation used to be balkanized with separate annotations for named entities, co-reference, semantic relations, discourse connectives, temporal entities, etc. Each annotation had its own associated evaluation, and training data was split across many resources. The idea behind AMR has been to unify the semantic landscape.

The authors wrote down the meanings of thousands of English sentences in simple, whole-sentence semantic structures. AMR and the tools associated with it have the following principles:

- Rooted, directed, edge-labeled, leaf-labeled graphs, which are easy for people to read, and for programs to traverse. This traditional format is equivalent to feature structures, conjunctions of logical triples, directed graphs, and PENMAN inputs. The latter is used for human reading and writing. The *root* of an AMR represents the focus of the sentence or phrase.
- AMR trees abstract away from syntactic idiosyncrasies, attempting to assign the same AMR to sentences that have the same basic meaning, e.g. *he described her as a genius*, *his description of her: genius*, and *she was a genius, according to his description* are assigned the same tree.
- Extensive use of PropBank framesets (see Section 2.4.6). For example, AMR represents `bond investor` using the frame `invest-01`, even though no verbs appear in the phrase.
- Agnostic about how to analyze/generate.
- Heavily biased towards English, originally not an interlingua.

2.4.8.1 *AMR Content*

In neo-Davidsonian fashion, AMR introduces variables (or graph nodes) for entities, events, properties, and states. Leaves are labeled with concepts: `(b / boy)` refers to an instance (called `b`) of the concept 'boy'. Relations link entities: `(d / die-01 :location (p / park))` means there was a death d in the park p. When an entity plays multiple roles in a sentence, AMR employs re-entrancy in graph notation (nodes with multiple parents) or variable re-use in PENMAN notation.

Concepts are either English words (`boy`), PropBank framesets (`want-01`), or special keywords. The latter include special entity types (`date-entity`, `world-region`, etc.), quantities (`monetary-quantity`, `distance-quantity`, etc.), and logical conjunctions (`and`, etc.). There are approximately 100 relations:

- Frame arguments, following PropBank conventions. :arg0, :arg1, . . . , :arg5
- General semantic relations: :accompanier, :age, :beneficiary, :cause, :compared-to, :concession, :condition, :consist-of, :degree, :destination, :direction, :domain, :duration, :employed-by, :example, :extent, :frequency, :instrument, :li, :location, :manner, :medium, :mod, :mode, :name, :part, :path, :polarity, :poss, :purpose, :source, :subevent, :subset, :time, :topic, :value
- Relations for quantities. :scale, :quant, :unit,
- Relations for date-entities. :day, :month, :year, :weekday, :time,. . .
- Relations for lists. :op1, :op2, :op3, :op4, :op5, :op6, :op7, . . . , :op10
- The inverses of all relations, e.g. :arg0-of,
- Every relation has an associated reification, which is used when we want to modify the relation itself.

AMR's hundred relation types are in contrast with the sparse inventory of `4lang`, the semantic network we will introduce in the next chapter. In graphs, `4lang` uses 0-, 1-, and 2-arrows, see Section 3.1.3,

> but the difference between `4lang` and AMR is less severe than it may appear at first blush, since the overwhelming majority of AMR relations like `:employed-by` are simply treated as ordinary transitive predicates in `4lang` . . . Considerable technical differences remain, e.g. `4lang` does not countenance overt semantic passives like 'employed by'. (Kornai et al., manuscript)

The authors give examples of how AMR represents various linguistic phenomena. AMR handles some level of derivational morphology. Besides nominalizations that refer to a whole event or a role player in an

event, *-ed* adjectives frequently invoke verb framesets, e.g. *acquainted with* and *-able* adjectives often invoke the AMR concept *possible*, but not always.

Most *prepositions* simply signal semantic frame elements, but they are kept if they carry additional information. Cases when neither PropBank nor AMR has an appropriate relation, e.g. *The man was sued in the case* are solved like this:

```
(s / sue-01
:arg1 (m / man)
:prep-in (c / case))
```

NAMED ENTITIES. Any concept in AMR can be modified with a :name relation. There are standardized forms for about 80 named-entity types, e.g. person or country. Multiple forms of a concept are not normalized (*US* versus *United States*), nor are semantic relations inside a named entity analyzed. This offers a uniform treatment to titles, appositives, and other constructions.

REIFICATION The sentence *The marble was not in the jar yesterday* is represented as

```
(b / be-located-at-91
:arg1 (m / marble)
:arg2 (j / jar)
:polarity -)
:time (y / yesterday))
```

If AMR would not use the reification, we would run into trouble, e.g.

```
(m / marble
:location (j / jar
:polarity -)
:time (y / yesterday))
```

cannot be distinguished from the representation of *yesterday's marble in the non-jar*. Some reifications are standard PropBank framesets (e.g., `cause-01` for :cause, or `age-01` for :age).

2.4.8.2 *Limitations of AMR*

AMR does not represent inflectional morphology and universal quantification, it does not distinguish between real events and hypothetical, future, or imagined ones, e.g. in *the boy wants to go*, `want-01` and `go01` have the same status, and noun compounds do not have a systematic representation, e.g. *history teacher* and *history professor* translate to

```
(p / person :arg0-of (t / teach-01 :arg1 (h / history)))
```

and

```
(p / professor :mod (h / history))
```

respectively, because `profess-01` is not an appropriate verb. It would be reasonable in such cases to use a NomBank (Meyers et al. 2004) noun frame.

2.4.8.3 *Creating AMRs*

The AMR Editor allows rapid, incremental AMR construction. To assess inter-annotator agreement, as well as automatic AMR parsing, AMR developed the Smatch metric and an associated script that measure the overlap between two AMRs by viewing each AMR as a conjunction of triples. Smatch takes the variable mapping that yields the highest F-score.

2.4.9 *Enhanced English Universal Dependencies*

`4lang`, the meaning representation formalism we will introduce in the next chapter, is a semantic model, and the division of labor principle suggests that a semantic project should defer the task of syntactic analysis to existing tools. Interfacing with syntax remains an important problem. Kovács, Gémes, Kornai, et al. (2022) discuss how more recent `4lang` graphs are created from the Universal Dependencies (UD) representation created by Stanza (Qi et al. 2020). This section introduces recent developments in syntactic analysis which is relevant for semantics.

In creating so-called ***`enhanced++`*** English Universal Dependency graphs, Schuster and Manning (2016) are motivated by that many shallow natural language understanding tasks use dependency trees to extract relations between content words. They revisit and extend these dependency graph representations in light of the Universal Dependencies initiative, and provide an enhanced and an `enhanced++`English UD along with a converter from basic UD trees to these latter two kinds of graphs, which are part of Stanford CoreNLP and the Stanford Parser.

The authors point out that the usage of the Stanford Dependencies (SD) representation falls into two categories: syntactic and a shallow semantic representations. Syntactic tasks proper, such as source-side reordering for machine translation or sentence compression, require a syntactic tree: a sound syntactic representation is more important than the relations between individual words. These trees need to be strict surface syntax trees. For *shallow semantic tasks* on the other hand, such as biomedical text mining, open domain relation extraction, or unsupervised semantic parsing, the relations between content words are more important than the overall tree structure. These tasks use collapsed or so called CCprocessed SD representations, which may be general graphs instead of trees, and may contain additional and augmented relations. E.g. in *Fred started to laugh*, the relation between

the controlled verb *laugh* and its controller, *Fred* is made explicit in the CCprocessed SD representation.

The enhanced UD representation has the following features:

- There are additional relations and augmented relation names.
- Augmented modifiers: The collapsed SD graphs also include the preposition in the relation name. This helps to disambiguate the type of the modifier. All nominal modifiers (`nmod`) also include the preposition in their names. The same is true for more complex PPs which are either analyzed as adverbial clause modifiers (`advcl`) or as adjectival clause modifiers (`acl`). Conjunct relations are augmented, e.g. `conj:and`.
- Governors and dependents are propagated to clauses with conjoined phrases.
- Subjects of controlled verbs are linked.

2.4.9.1 *The `enhanced++` UD representation*

The `enhanced++` UD representation is more interesting for natural language understanding systems that try to extract relationships between entities, e.g those in open domain relation extraction, or relationships between objects in image descriptions.

PARTITIVE NOUN PHRASES are phrases such as *both of the girls*, in which *both of the* acts semantically as a quantificational determiner. In the basic UD representation, however, *both* is the head while *both girls* is headed by *girls*. In order to obtain a similar analysis for these phrases, `enhanced++` UD changes the structure of the basic dependency trees, which is not allowed according to the guidelines for enhanced dependency graphs. They treat the first part of the phrase as a quantificational determiner, promote the semantically salient NP to be the head of the partitive, and analyze the quantificational determiner as a flat multiword expression that is headed by its first word. The quantificational determiner is attached using the special relation `det:qmod`.

LIGHT NOUN CONSTRUCTIONS such as *a panel of experts* or *a bunch of people* are treated similarly.

MULTIWORD PREPOSITIONS such as *the house* ***in front of*** *the hill* traditionally contain a relation between *house* and *front*, and *front* and *hill*. Here the enhancement++ lies in representing the relation between *house* and *hill*.

CONJOINED PREPOSITIONS such as *I bike* ***to and from*** *work* also pose some challenges. Ideally there is an `nmod:to` as well as an

`nmod:from` relation: *bike to work* and *bike from work* are conjoined by *and*. CCprocessed Stanford Dependencies representation introduced copy nodes which `enhanced++` UD adapts. The intended meaning can be illustrated as 'I bike and bike$^{(o)}$ to and from work, respectively'.

CONJOINED PREPOSITIONAL PHRASES such as *She flew **to Bali or to Turkey*** should encode that the two `nmod:to` relations are conjoined by *or*. For these reasons, `enhanced++` UD also analyze such clauses with copy nodes.

`enhanced++` UD attaches both the referent of a *relative pronoun* directly to its governor, and the relative pronoun to its referent with a referent (`ref`) relation. E.g. the analysis of *The boy who lived* includes both `++boy ref who` and `++boy nsubj lived`.

`enhanced++` UD does not propagate object or nominal modifier relations in clauses with conjoined verb phrases such as *the store buys and sells cameras* because of many cases such as *she was reading or watching a movie*, where *movie* is not the object of *reading*. In contrast to AMR (Section 2.4.8), `enhanced++` UD does not distinguish between comitative and instrumental: AMR requires semantic role labeling, which is very hard.

`enhanced++` UD is limited regarding generalized quantifiers and controlled verbs, such as *Everybody wants to buy a house*

```
Everybody nsubj:xsubj buy,
```

where the UD graph encodes approximately *Everybody wants that everybody buys a house*. The graph for *Everybody sleeps or is awake* approximately encodes *Everybody sleeps or everybody is awake*. Another imitation regards whether a conjoined subject (*Sue and Mary are carrying a piano*) should be interpreted distributively or collectively, which depends on world knowledge and the context.

2.4.10 *The State of the Art in Semantic Representation*

In this filnal section of the chapter, we follow two recent papers in overviewing semantic representation schemes. Finally, in Section 2.4.10.4, we shortly discuss a framework with and emphasis on quantification.

Abend and Rappoport (2017) clarify the general goals of research on semantic representation (except for vector space models), and compare them with syntactic schemes.

The paper discusses the goals of semantic representations (SRT), the components, (predicate-argument relations, discourse relations and logical structure), the concrete SRT schemes and annotated resources, the criteria for evaluation, and the relation to syntax. They focus on the level above the words, i.e. the meaning relationships between lexical items, rather than the meaning of the lexical items themselves. The

main differences between SRTs are the formalism, the interface with syntax, the ability to abstract away from formal and syntactic variation, the level of training required for annotators, and the level of cross-linguistic generality.

In Abend and Rappoport's view, SRTs should be paired with a (computationally efficient) method for *extracting* information from them that can be directly evaluated by humans. *Applications* include inference, as in textual entailment or natural logic; supporting knowledge base querying; and defining semantics through a different modality, images, or embodied motor and perceptual schemas. (They defer emotions and sentiment.)

2.4.10.1 *Semantic Content*

As we have seen in Section 2.4.2, *events* (sometimes called frames, propositions or scenes) include the predicate (main relation, frame-evoking element), arguments (participants, core elements) and secondary relations (modifiers, non-core elements). There are ontologies and lexicons of *event types* (also a predicate lexicon), which categorize semantically similar events evoked by different lexical items. FrameNet, defines frames as schematized story fragments evoked by a set of conceptually similar predicates. The Richer Event Descriptions framework is another event resource.[6]. This notion of events should not be confused with events as defined in Information Extraction and event coreference, such as a political or a financial event.

SRTs differ in which *nominal and adjectival predicates* are covered. Recent versions of PropBank cover eventive nouns and multi-argument adjectives. FrameNet covers all these, and also covers relational nouns that do not evoke an event, such as "president". SRTs may represent arguments that appear outside sentence boundaries, or do not explicitly appear anywhere in the text.

Core and non-core arguments are distinguished semantically rather than distributionally. Core arguments are whose meaning is predicate-specific and are necessary components of the described event, while non-core arguments are predicate-general. FrameNet defines core arguments as conceptually necessary components of a frame, that make the frame unique and different from other frames; and peripheral arguments, which introduce additional, independent or distinct relations, e.g. time, place, manner, means and degree.

Semantic roles in FrameNet are shared across predicates that evoke the same frame type, e.g. "leave" and "depart"; PropBank roles are verb-specific, and the set was extended by subsequent projects such as AMR; and VerbNet and subsequent projects use a closed set of abstract semantic roles for all predicate arguments, such as AGENT, PATIENT and INSTRUMENT.

6 Citations can be found in the original Abend and Rappoport (2017).

Abend and Rappoport discuss *temporal relations* in details. This kind of analysis may mean timestamping according to time expressions found in the text, or by predicting their relative order in time. The main resources are TimeML, a specification language for temporal relations; and annotated corpora by the TempEval series of shared tasks. The theory goes back to scripts, schematic, temporally ordered sequences of events associated with a certain scenario, e.g. going to a restaurant (Section 2.4.2). Causal relations between events have applications (including planning and entailment) and annotation schemes, also integrated with TimeML-style temporal relations. The internal temporal structure of events has been less frequently tackled, but Moens and Steedman (1988) defined an ontology for the temporal components, e.g. a preparatory process (e.g., "climbing a mountain") and its culmination ("reaching its top"). Statistical work on this topic is unfortunately scarce but involves aspectual classes, and tense distinctions.

Spatial Relations have their cognitive theories and applications in geographical information systems or robotic navigation. The task of Spatial Role Labeling with its shared task SpaceEval subsumes the identification and classification of places, paths, directions, and motions, and their relative configurations.

In the papers running example, *Although Ann was leaving, she gave the present to John.*, the leaving and the giving events are sometimes related through 'CONCESSION', evoked by "although". *Discourse* analysis is useful but overlooked for summarization, machine translation and information extraction. Resources include the Penn Discourse Treebank, which classifies the relations between discourse units using high-level relation types like TEMPORAL, COMPARISON and CONTINGENCY; and finer-grained ones such as JUSTIFICATION and EXCEPTION. This treebank focuses on local discourse structure. The RST Discourse Treebank puts more focus on higher-order discourse structures and deeper hierarchical structures.

Attila Novák (personal communication) drew the attention in his prereview to Discourse Representation Theory (DRT, Kamp, Genabith, and Reyle (2011)). Parsing to Discourse Representation Structures, a formal meaning representation introduced by DRT, is a complex task, comprising other NLP tasks, such as semantic role labeling, word sense disambiguation, co-reference resolution, and named entity tagging. We also learn from the introduction at nlpprogress[7] that DRSs show explicit scope for certain operators, which allows for a more principled and linguistically motivated treatment of negation, modals and quantification, as has been advocated in formal semantics.

A narrower but better studied field is the segmentation of scientific papers into parts like background and discussion. Some schemes, e.g. the Groningen Meaning Bank (Basile et al. 2012) and UCCA (see Section 2.4.10.2) support cross-sentence semantic relations.

7 http://nlpprogress.com/english/semantic_parsing.html#drs-parsing

Logical structure, i.e. quantification, negation, coordination and their associated scopes are important in applications that require mapping text into an executable language, such as a querying language or robot instructions, and in recognizing entailment relations. We will shortly discuss an example representations framework in Section 2.4.10.4. Approaches to *inference and entailment* include Recognizing Textual Entailment, and Natural Logic with different annotation principles and resources.

2.4.10.2 *Semantic Schemes and Resources*

- As we saw in Section 2.4.8, AMR has predicate-argument relations, including semantic roles (adapted from PropBank) that apply to a wide variety of predicates (including verbal, nominal and adjectival predicates), modifiers, co-reference, named entities and some time expressions, but currently no relations above the sentence level. It is English-centric, which results in an occasional conflation of semantic phenomena realized similarly in English, and difficulties with invariance across translations. Abend and Rappoport illustrate this with the pair of sentences *I happened to meet Jack in the office* and *I asked to meet Jack in the office*, which have similar syntactic forms. When translating the sentences to German, the divergence between the semantics of the two sentences is clear: in the first one "happened" is translated to an adverb: *Ich habe Jack im Büro zufällig getroffen*, and in the second *asked* is translated to a verb: *Ich habe gebeten, Jack im Büro zu treffen.*

- Universal Conceptual Cognitive Annotation (UCCA, Abend and Rappoport (2013)) is a cross-linguistically applicable scheme for semantic annotation, building on typological theory, primarily on Basic Linguistic Theory. It includes argument structures of various types and relations across languages, but no semantic role information. UCCA distinguishes between primary and aspectual verbs, e.g. *happen to*, and it supports annotation by non-experts.

- Universal Decompositional Semantics (UDS) provides semantic role annotation, word senses, and aspectual classes (e.g., ±realis) collected through crowd-sourcing. UDS uses feature bundles e.g. +volition and +awareness instead of agent.

- The Prague Dependency Treebank (PDT) Tectogrammatical Layer (PDT-TL) represents argument structure (including semantic roles), tense, ellipsis, topic/focus, co-reference, word sense disambiguation, and local discourse information.

- There are schemes based on Categorial Combinatory Grammar.

- HPSG-based Schemes use feature bundles. Annotated corpora and manually crafted grammars exist for multiple languages along with broad-coverage Semantic Dependency Parsing shared tasks and corpora.
- OntoNotes has multiple inter-linked layers of annotation, borrowed from different schemes.

UNIVERSALITY. Besides remarkable cross-lingual resources like BabelNet, UBY (Gurevych et al. 2012), and Open Multilingual WordNet, semantic role labeling (SRL) schemes and AMR have also been studied for their cross-linguistic applicability. PropBank and FrameNet have been translated to multiple languages, and there are SRT schemes that set cross-linguistic applicability as main criteria, e.g. UCCA, and the LinGO Grammar Matrix, both of which draw on typological theory.

2.4.10.3 *Anchoring graph fragments to tokens*

Finally, we would like to follow Koller, Oepen, and Sun (2019) in distinguishing three flavors by the degree of anchoring. The strongest form of anchoring is bi-lexical dependency graphs, when *graph nodes injectively correspond to surface lexical units (tokens)*. In such graphs, each node is directly linked to a specific token (but there may be semantically empty tokens), and the nodes inherit the linear order of their corresponding tokens. Linguistic frameworks in this flavor include CCG word–word dependencies, Enju Predicate–Argument Structures, DELPH-IN MRS Bi-Lexical Dependencies (which we will shortly discuss in Section 2.4.10.4), and Prague Semantic Dependencies.

The middle flavor relaxes the correspondence relations between nodes and tokens, while still explicitly annotates the correspondence between nodes and parts of the sentence, but nodes may align with subtoken or multi-token sequences, e.g. (derivational) affixes or phrasal constructions. Nodes may correspond to overlapping spans, enabling lexical decomposition (e.g. that of causatives or comparatives). Representatives include Universal Conceptual Cognitive Annotation and two variants that reduce underspecified logical forms into directed graphs: Elementary Dependency Structures and Dependency Minimal Recursion Semantics (Section 2.4.10.4).

AMRs – on the other extreme – are *unanchored*, in that the correspondence is not explicitly annotated. AMR deliberately backgrounds notions of compositionality and derivation. The framework frequently invokes lexical decomposition and represents some implicitly expressed elements of meaning, abstracting furthest from the surface signal.

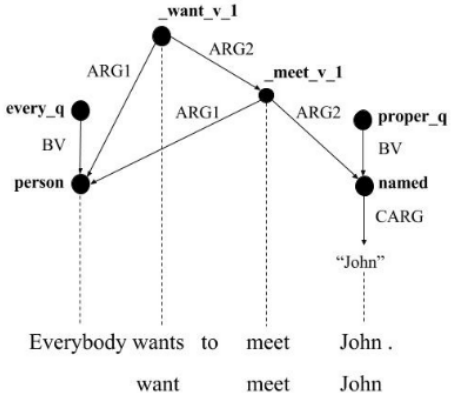

Figure 5: Semantic representation of the sentence "Everybody wants to meet John" from Buys and Blunsom (2017). The graph is based on the Elementary Dependency Structure (EDS) representation of Minimal Recursion Semantics (MRS). The alignments are given together with the corresponding tokens, and lemmas of surface predicates and constants.

2.4.10.4 *Quantification, Minimal Recursion Semantics, and its variants*

Most of the information content of the sentences is not in the structure (syntactically disambiguated, provided with quantifiers), but in the (content) words (Kornai 2019, Sec. 1.3). This is what `4lang`, the semantic formalism introduced in the next chapter, tries to represent. Nevertheless, we end this chapter with a short discussion of Minimal Recursion Semantics, which is more closely related to logic-based semantic theories.

The linguistic structures targeted in semantic parsing are predominantly shallow, restricted to relations between surface word tokens. An exception is provided by Buys and Blunsom (2017), who propose a neural encoder-decoder transition-based parser for Minimal Recursion Semantics (MRS, Copestake et al. (2005) and Copestake et al. (2016)). MRS also serves as the semantic representation of the English Resource Grammar (ERG, Flickinger (2000)). Buys and Blunsom define a common framework for semantic graphs for MRS-based graph representations (more precisely Dependency MRS and Elementary Dependency Structures, EDS) and AMR (Section 2.4.8).

MRS is a framework for computational semantics that can be used for both parsing or generation. As Figure 5 shows, instances and eventualities are represented with logical variables. Argument labels are drawn from a small, fixed set of roles. Arguments are either logical variables or handles. Handles are designated formalism-internal variables. Handle equality constraints support scope underspecification; multiple scope-resolved logical representations can be derived from one MRS.

MRS was designed to be integrated with feature-based grammars like HPSG (Section 2.4.10.2) or Lexical Functional Grammar. EDS (Oepen and Lønning 2006) is a conversion of MRS to variable-free dependency graphs which drops scope underspecification.

Mi generáltunk. Legalábbis azt hittük, hogy generálunk.
'We generated. At least we thought we generated.'

— Ferenc Kiefer on generative linguistics before Chomsky (1970).

3

THE 4LANG SEMANTIC NETWORK

As we pointed out in Section 1.1, some of our contributions are related to the 4lang theory and formalism for representing the semantics of natural language, which has been developed in the Human Language Technologies Research Group Budapest, and published along with partial implementation in many research papers (Kornai 2010a, 2012; Nemeskey et al. 2013; Kornai et al. 2015; Recski et al. 2016; Kovács, Gémes, Iklódi, et al. 2022) and two books (Kornai 2019, 2023).

The more important 4lang-related contributions of this thesis (Sections 3.3, 7.2 and 7.3) take derivatives of 4lang – the definition graph, or a word embedding created from the graph – as input. Besides, the author of this thesis had a great role in the manual creation of a set of core definitions for 4lang, but our claims related to this part of the work will focus on to the problem of thematic roles (Chapter 5). Now we give some background by introducing 4lang.

In the previous chapter, we introduced some fundamentals of symbolic meaning representation systems, notably Quillian (1969)'s seminal experiments with his semantic network (Section 2.2.1). While the 4lang theory involves other formalisms, for the purposes of the present thesis, 4lang can be primarily viewed as a semantic network, practically a graph, whose nodes are labeled with (names of) concepts (similarly

to AMR, see Section 2.4.8), and the edges with 0, 1, or 2, roughly corresponding to the most basic syntactic relations.

The name `4lang` refers to that the core dictionary, the object of inquiry in all of the `4lang` related work in this thesis, has bindings in four languages, representatives "of the major language families spoken in Europe; Germanic (English), Slavic (Polish), Romance (Latin), and Finno-Ugric (Hungarian)". More recently, Kornai (2023) added Japanese and Chinese. "The relative ease of creating these new bindings goes some way onward ameliorating concerns of eurocentricity."

"`4lang` is an algebraic (symbolic) system that puts the emphasis on lexical definitions at the word and sub-word level, and on valency (slot-filling) on the phrase and sentence level" (Recski et al. 2016). These two levels are the focus of this chapter and Chapter 5 respectively. "Historically, `4lang` falls in the AI/KR tradition, following on the work of Quillian (1969, Section 2.2.1), Schank (1975, Section 2.2.3), and more recently Banarescu (2013, Section 2.4.8). Linguistically, it is closest to Wierzbicka and Goddard (1972, 2002, Section 2.3.3) and to modern theories of case grammar and linking theory (see Butt (2006) for a summary)." (References to sections in the present thesis added.)

3.1 NODES AND EDGES

3.1.1 *Concepts: monosemy and language- and POS-independence*

The backbone of `4lang` consists of 1942 defined words and bound morphemes (see Section 3.2). However, the version of the Longman dictionary that was available to us (Bullon 2003) uses other elements (Section 2.4.1.2), so we further expanded the vocabulary with 197 simple words (e.g. *dimension, two, communicate, conform , mammal, item, artifact*), 188 proper names, the definition of which is essentially just a reference to the corresponding element of the encyclopedia (e.g. *Greenland, Greenwich, Guy Fawkes*) and 147 compounds (*bell-shaped, bitter-tasting, blue-black*). The latter are uninteresting from our present perspective.

The definitions in `4lang` were made by human labor, consulting classical dictionaries in the most cases, especially the Longman dictionary. This part of the work is unfortunately unreproducible. We quoted from the Natural Semantic Metalanguage project (Section 2.3.3) that words (morphemes, etc.) can have the same meaning representation there even if the part of speech, the scope of use, or the polysemy pattern is different. `4lang` definitions follow this line: (unary) predicates in `4lang` represent language- and POS-independent, monosemic concepts. We discuss language-independence and monosemy in this subsection, while POS-independence will be investigated in the next one.

Monosemy means that `4lang` tries to grasp the abstract meaning of the words, from which specific uses can be deduced. Kornai and

Makrai (2013) cite the definition of *potash* from Webster's Third (Gove 1961) to show how words considered there to be polysemous are defined in traditional lexicology. *Potash* has four meanings there. In the presentation of (Makrai 2013), we provided a similar example from the English WordNet (Miller (1995), see Section 2.4.3) with six meanings of the word *stomach*. According to the principles of `4lang`, most words are monosemic. Disambiguation is only done for pure homonyms, e.g. the word form *state* corresponds to separate entries in the senses related to 'country' and 'condition'. The disambiguation of homonyms have not been implemented. How the distinction between polysemy and homonymy can be made on the basis of data and word embeddings will be discussed in Chapter 8 in the frame of multilingual word sense induction, the computational task of clustering word occurrences to lexical items based on two corpora in different languages.

Rather than including as much information on different uses as possible in disambiguation, we prefer representing each surface morpheme with a single graph. In Ruhl (1989)'s view, the elements of the meaning of a word in a context that is not present in the monosemic lexical item should be deduced from the similarly abstract representations of context words. We think that computing the meaning of occurrences of words that are usually called metaphoric is the basic mechanism behind human linguistic capabilities, and artificial understanding should work with a similar goal, possibly with the use of non-lexical components to handle extra-linguistic knowledge and pragmatic implicatures. The interested reader should consult Recski (2018, especially Section 4.4.3) and Kovács, Gémes, Iklódi, et al. (2022, especially Sections 5 and 6).

The `4lang` dictionary strives to be *language-independent*. When defining the words, we tried to take into account a couple of languages, and the word forms of the terms were indicated in Hungarian, English, Latin and Polish. Since the creation of the definition formulas, colleagues have expanded the dictionary to more languages, Ács, Pajkossy, and Kornai (2013) to 40 languages, and Kornai (2023) to Japanese and Chinese.

Language-independence may be contrasted with the Saussurean definition of a linguistic sign which is an ordered pair consisting of a cluster of (spoken or written) forms in a specific language and an extra-linguistic category in the mind. Whether human categorization is dependent of the mother tongue and other languages learned by the speaker early on is a classical topic in psycho-linguistics. Common experience shows that people can express the same content in any language, and the greatest problem one faces in finding translational equivalents is that an ambiguous word in some language may (not surprisingly) translate to some other language in multiple ways, depending on context.

3.1.2 *Syntactic and semantic type*

`4lang` contains a single concept where two words differ only in their parts of speech, e.g. action nouns are the same concept as the verbal stem, since `4lang` describes the conceptual meaning. This approach obviously deviates from Montague grammar (Montague 1970), where syntactic types correspond to semantic types. `4lang` is a conceptual network, so its representations try to factor out pure morpho-syntactic differences on the word level.[1] This avoidance of types is in contrast to lexicographic practice, both traditional or symbolic computational, that splits usages of words by parts-of-speech. Furthermore, unlike in Conceptual Structures (Jackendoff (1972, 1990), Section 2.3.5), our concepts are free of semantic type as well.

The `4lang` approach to the lexicon can be illustrated in relation to the phenomenon that a great part of the English core vocabulary consists of words that appear as nouns and verbs as well, with semantically equivalent meanings: a $divorce_N$ is exactly a situation when some people $divorce_V$. The corresponding pairs in Hungarian are derivational ones: remaining with the same example, the noun *vál-ás* is derived from the verb *vál(ik)* by a compositional suffix.

Formal semantics is organized along the principle of compositionality: the representation of a phrase or a sentence is computed from the representations of the immediate constituents and the way of their composition. Montague Grammar formalizes the compositional requirement by associating rewrite rules over syntactic forms to semantic rules. Terminals of the semantic sub-grammar are semantic types, most notably entities and truth-valuable states of affairs.

Compositionality also applies to `4lang` graphs. Formulas in the handwritten core vocabulary, which we discuss in Section 3.2, are parsed to graphs in a rule-to-rule fashion, and the representations of phrases and sentences are composed of those of the words. The main operation in both is to draw a link from a node in the graph corresponding to the macro-structure of the linguistic unit to the so-called head-node of the constituent. The head-node corresponds to the genus (Sections 2.2.1, 2.2.8 and 3.1.2). The theory allows the link to point to a sub-graph, motivated e.g. by *accusativus cum infinitivo* sentences like *I see the father coming* where the object of seeing can argued to be the coming of the father as well as the father himself, but this idea is unrelated to the present thesis, and it is not implemented.

The lack of semantic types can be seen as an instance of radical lexicalism: `4lang` concentrates on the meaning of words and phrases at the expense of type consistency in the graph. Our definitions can of course turn out to be less exact that those applying POS distinc-

1 The interested reader may learn about the syntactic part of the `4lang` theory, motivated by functional programming and formalized in Eilenberg Machines, in Section 6.3.2 of Kornai (2008) and in Kornai (2019).

tions. Another problem is when the head-node depends on the POS: the head of *cook* has to be 'person' if the noun is meant, and 'make' if the verb. Nevertheless, `4lang` representations still turn out to capture enough lexical content to be useful in application, especially in word and sentence level similarly and entailment, see Section 3.7.

3.1.3 *Edges*

In the `4lang` meaning representation framework, the meaning of words and greater linguistic units is formalized in pointed directed graphs with nodes labeled by concepts and edges colored in three colors: 0, 1, and 2. Pointedness means that one node, the *head*, is distinguished for compositional purposes, as already discussed in Section 3.1.2.

In Section 2.2.4, we introduced Woods (1975)'s argument that a too large inventory of edge types (colors) makes reasoning with graphs computationally unfeasible. This problem is avoided in `4lang` by splitting relations to various levels. At the deepest level, there are only three types of edges (0, 1, and 2). When there is an edge $c_1 \xrightarrow{i} c_2$ with label $i \in \{0, 1, 2\}$ from concept c_1 to concept c_2, we will also say that c_2 is on the ith *partition*[2] of c_1. Binary predicates and lexical relations,[3] whose appearance in static word embeddings is the topic of Chapter 7, are represented with nodes (typically with 1 and 2-edges leading out of them to their first and second argument). These relations represent kinds of information including the type of general knowledge ConceptNet (see Section 2.4.7) represents.

Ditransitives (ternary and higher arity verbs) are eliminated by decomposition to at most binary ones (Kornai 2012) with methods pioneered in generative semantics. Following Jackendoff (1972), who defined *kill* as 'cause to die', with a `4lang` formula,

```
=AGT CAUSE [=PAT[die]],
```

we define *put* as 'cause to (be) at', (`=AGT CAUSE [=PAT AT =TO]`), and the two classes of Schank (1972) as *give*: 'cause to have', (`=AGT CAUSE [=DAT HAS =PAT]`) and *tell*: 'cause to know', (`=AGT CAUSE [=DAT KNOW =PAT]`). As we will see later, =AGT, =PAT, and =DAT are what we

2 Those who are familiar with gold-age meaning representation, especially Hendrix (1975), should note that in `4lang`, partition is meant much more simply than for Hendrix, who introduced a machinery with the same name to provide an adequate quantification mechanism for semantic network concepts. In `4lang`, more concepts on a partition of a concept (out-neighbors with a fixed edge label) are interpreted as a conjunctive bundle of properties.

3 `4lang` can represent both the static and constant lexical meaning of words as well as the contextual and dynamic properties they acquire in context. Lexical relations belong to the former ('cows make milk'), while binary relations can represent context-dependent properties as well ('John has a cow'). As the present thesis investigates lexical meaning, binary relations above lexical ones are out of the scope, and we use the two terms basically in an interchangeable way.

call deep cases, placeholders for the representation of the agent, the patient, and the recipient ("dative") of the verb respectively. There are eight deep cases is total, some of which represent arguments of relational nouns or function morphemes (Chapter 5).

Turning to the edge-colors, 0 denotes every relation in which a concept modifies some other as a whole: we draw an abstraction over the traditional genus/hypernym/IS-A (e.g. `dog` $\xrightarrow{0}$ `animal`, see Sections 2.2.8, 2.2.9 and 2.3.1), (generic) unary predication (`dog` $\xrightarrow{0}$ `bark`), and attribution (`dog` $\xrightarrow{0}$ `faithful`). The interested reader may learn more about IS-A, genus, and hypernym in Section 4.5 of Kornai (2019). 0 is used for verbs as well as nouns. Unlike Levelt, Roelofs, and Meyer (1999), where `escort IS-TO accompany`, in `4lang` we simply state that `escort` $\xrightarrow{0}$ `accompany`.

1 and 2 represent two arguments of a function that play asymmetric roles, e.g. the agent and patient role of a verb (e.g. `cow` $\xleftarrow{1}$ `make` $\xrightarrow{2}$ `milk`), or the figure and the ground in tempo-spatial relations (`star` $\xleftarrow{1}$ `at` $\xrightarrow{2}$ `sky`). Nodes (concepts) with an 1 or 2-labeled out-edge will be called *binary*, while the rest will be called *unary* because these concepts correspond to unary predicates of truth-conditional logic

TWO REMARKS RELATED TO THE FORMALISM It can be argued that in terms of predication, the direction of the 0 versus 1 and 2 edges is somewhat inconsistent: in `dog` $\xrightarrow{0}$ `animal`, the link goes from the argument to the predicate, while in `cow` $\xleftarrow{1}$ `make` $\xrightarrow{2}$ `milk`, the edges lead from the function to the arguments. In the view of the thesis author, this discrepancy may be an accident in the development of the system, but need not corrupt empirical results in applications. Nevertheless, András Kornai writes in personal communication that what the argument and what the predicate is in the case of 0, and also in the case of intransitives in general, is debatable/changeable, e.g. in the first two articles of Montague, there is boy(sleep) and sleep(boy), respectively. Both can be argued for. The interested reader may refer to Recski (2016b) as well.

Another remark has been made by Tibor Szécsényi (personal communication), related to the representation of "ergative (and other strange)" expressions (e.g. *Peter likes Mari* vs. *Peter is pleased by Mari*)".

> The thematic role–argument correspondence is not always clear! Wouldn't it have been enough to assume a single unary predicate-argument relation: `cow` $\xrightarrow{0}$ `(make` $\xleftarrow{0}$ `milk)`? This would have been more in line with the idea of light verb construction in syntactic theory, with Currying in logic, and with the type $\langle e, \langle e, t \rangle \rangle$ of transitive verbs. Or, once we get to type logic, what would the `4lang` representation corresponding to the type-raised, generalized quantifier transitive verb type $\langle \langle e, \langle e, t \rangle \rangle, \langle e, t \rangle \rangle$ look like? Would that make sense? (Tibor Szécsényi, translated by the thesis author.)

The ergative problem was one of the motivations for the introduction of the thematic roles like `=AGT`, a shallower level of binary relations. There are theoretical motivations for using hypergraphs like `cow` $\xrightarrow{o}$ `(make` $\xleftarrow{o}$ `milk)`, where edges can point to edges. We already mentioned an example in Section 3.1.2. However, disjunction, negation, and all forms of quantification are considered secondary phenomena in `4lang` (Kornai 2010b) which would make the model computationally more complex without much benefit in terms of accuracy in text understanding. The emphasis of `4lang` is on the lexical/conceptual content rather than an elaborated type theory involving raising and generalized quantifiers. The interested reader may consult Section 4.5 of Kornai (2023) as well.

3.2 THE RECURSIVE PROCESS OF WORD DEFINITION

Symbolic representations define concepts by other concepts (Section 2.2.1). Some methods take this circularity as a basic property of language, while others break it by using primitives, words that play the same role in semantics as primitive notions do in mathematics. The first approach includes disciplines ranging from structuralist semantics to semantic networks (Chapter 2) and information retrieval (Section 3.3). In Section 2.2.6 we reviewed Hayes (1979)'s analysis of the axiom-concept graph, and his considerations on which direction the definition process should follow. The primitive-based approach is exemplified in this thesis by the Natural Semantic Metalanguage (Section 2.3.3), and the Longman Defining Vocabulary (Section 2.4.1). The `4lang` approach is closer to the latter, but it is important that we do not specify the defining vocabulary on theoretical grounds, but we derive it from the definition graph (Section 2.2.6) with an iterative process (see Kornai et al. (2015, Section 2.1), Ács, Nemeskey, and Recski (2017, Section 2.2), and Kornai and Makrai (2013), the latter is in Hungarian).

The meaning of a sentence is composed of the meaning of its words, but the word inventory is still too great to give a `4lang` account of each item manually. Now we describe our method for vocabulary reduction from the, say, 80–160 thousand (disambiguated) words in a traditional dictionary to a defining vocabulary for which we can create `4lang` representations manually, constituting the main contribution in this chapter.

It must be noted that members of the defining vocabulary are not primitives of definition. This is in accordance with some other approaches: the structuralist notion of word sense; that "the full meaning of any concept is the whole network as entered from the concept node" (Collins and Loftus 1975); and what Lenat and Guha (1990) say about the lack of primitive actions in Cyc: "actions are not merely macros introduced for notational convenience, for use instead of more complex sequences of primitive actions. [Our] approach is motivated by two rea-

sons: we wish to be able to reason at different levels of abstraction and a priori assigning of a set of actions as primitives goes against this".

Our methods for defining the whole vocabulay in terms of a more restricted set (as well as previous work in this field) are discussed in Section 2.1 of Kornai et al. (2015). There are two basic approaches: bottom-up methods use a defining vocabulary specified on some theoretical basis, but our group has done top-down computations as well to discover the defining vocabulary of both traditional dictionaries and our manually written definitions themselves.

> The first modern efforts in [the direction of a basic vocabulary] are Thorndike (1921)'s Word Book, based entirely on frequency counts (combining TF and DF measures), and Ogden (1944)'s Basic English, based primarily on considerations of definability. The Swadesh (1950) list puts special emphasis on cross-linguistic definability, as its primary goal is to support glottochronological studies.
>
> [...]
>
> The idea that there is a small set of conceptual primitives for building semantic representations has a long history both in linguistics and AI as well as in language teaching. The more theory-oriented systems, such as Conceptual Dependency (Schank 1972) and NSM (Wierzbicka 1985) assume only a few dozen primitives, but have a disquieting tendency to add new elements as time goes by (Andrews 2015). In contrast, the systems intended for teaching and communication, such as Basic English (Ogden 1944) start with at least a thousand primitives, and assume that these need to be further supplemented by technical terms from various domains. [...] A trivial lower bound [on the number of primitives] is given by the current size of the NSM inventory, 65 (Andrews 2015), but as long as we don't have the complete lexicon of at least one language defined in NSM terms the reductivity of the system remains in doubt.
>
> For English, a Germanic language, the first provably reductive system is the Longman Defining Vocabulary (LDV), some 2,200 items, which provide a sufficient basis for defining all entries in LDOCE (using English syntax in the definitions). (Ács, Pajkossy, and Kornai 2013)

The core vocabulary of the `4lang` meaning representations framework is a set of about three thousand concepts with English, Hungarian, Latin and Polish exponents[4] and formal definitions that can be compiled to `4lang` graphs with the pymachine software package. The

4 Ács, Pajkossy, and Kornai (2013) describe how bindings in other languages can be created automatically.

original vocabulary (words with ID up to 2692) was specified in the Hungarian Unified Ontology (MEO) Project based on theoretical considerations similar to those mentioned in the previous citation. This process is also described in the paper:

> We[5] built a seed list composed of the Longman Defining Vocabulary (2,200 entries), the most frequent 2,000 words according to the Google unigram count (Brants and Franz 2006) and the British National Corpus, as well as the most frequent 2,000 words from Polish (Halácsy et al. 2004) and Hungarian (Kornai et al. 2006). [For Latin,] we added the classic Diederich (1939) list and Whitney (1885)'s *Roots*. (Ács, Pajkossy, and Kornai 2013)

Turning to the top-down method, in the same Kornai et al. (2015), we formalized the defining vocabulary in graph-theoretic terms, based on the definition graph, whose nodes correspond to (disambiguated) words, and a directed edge $u \to v$ represents if v is used in the definition of u. The mathematical formulation of the defining vocabulary is a feedback vertex set (FVS) that contains all nodes without out-edges (these are definitional primitives) and one node from each directed cycle. We found that in definition graphs there are much smaller FVSs than there may be if the graph was random: "For example, in the English Wiktionary, 369,281 definitions can be reduced to a core set of 2,504 defining words, and in Collins English Dictionary we can find a defining set of 6,490 words." Gold-age versions of the Longman Dictionary were created with a pre-specified defining vocabulary (LDV), what still shows its advantages in the newer, non-LDV-based version we have access to, as the defining vocabulary of the not strictly LDV-based version still consists only of 1,061 words.[6] The interested reader may read more details on the possible gains of a smarter parsing of implicit cross references in dictionaries, handling compositional derivations of Latinate stems, disambiguation, and multiword expressions in the paper. The key point is that a cca. 3000-word vocabulary that we defined with `4lang` formulas at the middle of the past decade (Section 3.6) covers the defining vocabulary of traditional dictionaries. Further refinements of the `4lang` defining vocabulary can be found in Appendix 4.8 of Kornai (2019) and in Kornai (2023).

5 *The inventory* had been essentially compiled before the thesis author joined the group

6 This set roughly corresponds to the words that are marked with `u` (for uroboros (Ács, Pajkossy, and Kornai 2013; Kornai et al. 2015; Kornai 2019, 2023)) in the 6th column of the `4lang` dictionary file.

3.3 THE IMPORTANCE OF EACH CONCEPT IN DEFINITION

Symbolic representations define concepts and relations by other concepts and relations, possibly with the help of formal devices like our deep cases. In the previous sections, we introduced the `4lang` formalism and our approach to the iterative process of defining words by each other. In this section, which applies the ideas of the Hungarian paper by Makrai (2013) to the definitions accompanying Makrai (2014b), we quantitatively describe how important each node of the semantic network is for the definition of the whole vocabulary. Intuitively, the importance measure tells us which unary and binary[7] predicates (e.g. *exist* or the comparative *-er*) and thematic roles (e.g. the patient) have to be defined (if possible) or used (in the case of primitives) with the greatest caution.

3.3.1 *Introduction*

In order to quantify how important each concept is in sentence comprehension, we transform the definitions that represent the meaning of each word into a directed graph, with the concepts as nodes. PageRank is a method in computer science, especially traditional web search, originally introduced to measure the relevance of websites. We apply PageRank to the definition graph to obtain values assigned to each vertex, which can be interpreted as the importance of the corresponding concept in understanding other words and phrases.

The remainder of the section is organized as follows. In Section 3.3.2 we present the definition graph, and in Section 3.3.3 we present the PageRank method used to calculate the weight of each concept. Finally, we report the quantitative results in Section 3.3.4.

We work at the word level, yet it is important to talk about arguments of words (typically those of verbs and relational nouns). As it follows from the principle of compositionality, we require that the representation of the meaning of a structure consisting of a function and its arguments is composed from the representations of the meaning of the function and that of the arguments. To make this possible, the definitions of functions should indicate where the representation of each argument has to be inserted. We do this by referring to the deep cases of the arguments (Chapter 5).

This section repeats the experiments in a Hungarian paper of ours (Makrai 2013)[8]. Repeating the experiments was made necessary by a change in the deep cases system. In Makrai (2013), the names of the

7 Recall from our Section 3.1.3 or from Kornai (2012) that `4lang` accounts for the meaning of ditransitive (and higher arity) verbs using deep predicates of at most two variables.

8 Thanks for helpful comments by Ágota Fóris, Dávid Nemeskey, Gábor Prószéky, Tibor Vámos, and Tamás Váradi.

deep cases abbreviated Hungarian surface cases (NOM denoted the subject, ACC denoted the object, and DAT denoted the dative argument) or classes thereof (OBL denoted the oblique). Makrai (2014b) introduced a more theoretically grounded system (i.e. more easily comparable to Fillmore's idea and invariance with respect to alternations). This linker inventory is introduced in this thesis in Chapter 5. For the purposes of this section of the thesis, we repeated the experiments by Makrai (2013) with the 2014 definitions.

3.3.2 *The definition graph*

In this section, we first show how we transformed the `4lang` dictionary into a directed graph and a corresponding matrix, which enabled us to characterize the semantic importance of the concepts. Thereafter we describe the graph.

The vertices of the definition graph are concepts from the dictionary. Recall that *concepts* have to be understood in the broad sense, rather technical than the cognitive one of Section 2.2.6: they include unary and binary predicates and deep case symbols. Whenever the word 'metal' is used in the definition of 'steel', there will be a directed edge

$$\text{'steel'} \rightarrow \text{'metal'}$$

in the graph. This graph has 3,185 (2013 version: 2,897) vertices and 11,023 edges (2013 version: 7,816). These numbers show that the graph is sparse, i.e. there are relatively few edges between pairs of vertices, recall Section 2.2.6. (Nevertheless, the development of the definitions between Makrai (2013) and Makrai (2014b) resulted in a slight increase of edge density to $e/n^2 = 1.0866 \cdot 10^{-3}$ in 2014 from $9.3130 \cdot 10^{-4}$ in 2013 of (n and e are the number of nodes and edges respectively), which means that the concepts got better anchored.)

The mathematical concept of strongly connected components will play an important role later. Two vertices are called *strongly connected* if a path (a sequence of edges) connects them in both directions. This relation is an equivalence relation, it classifies the vertices into classes, which are called *strongly connected components*. The strongly connected components of the `4lang` graph are interesting by themselves as they give an intuition about the graph, so we briefly present them.

Table 4 shows some the strongly connected components. The largest component consists of quite mixed words (*yellow, four, sleep, under, lack, month...*). The next largest strongly connected components consist of cycles such as months, days of the week, or seasons. The definition of e.g. a month consists of the pieces of information that the definiendum is a month and which the previous and the next months are.

In some of the mid-sized strongly connected components, the lexicographer can single out a central concept. E.g. the components of

january, february, . . . , december	12
monday, tuesday, . . . , friday	7
bed, chair, cupboard, furniture, table	5
cereal, flour, grain, wheat	4
draw/2707, pen, pencil, write	4
king, monarch, queen, royal	4
autumn, spring/2318, summer, winter	4
buttocks, seat, sit	3
camera, lens, photograph	3
calm, disturb, upset	3
answer, question, reply	3
bake, bread, cake	3
female, male, sex	3
justice, right/1191, wrong	3
actor, stage/2220, theatre	3
many, much, quantity	3
husband, marriage, wife	3
poem, poet, poetry	3
cutlery, fork, spoon	3

Table 4: The mid-size strongly connected components of the `4lang` definition graph, version 2014. The largest component consists of 623 words, there are many 2-cycles, and 2430 primitives.

furniture and *cereal* can be intuitively analyzed as consisting of the mentioned abstract concept and examples (*bed, chair, cupboard*, and *table* or *flour, grain*, and *wheat*) thereof. Lexical definitions generally do not contain examples. *Furniture* and *cereal* are exceptional in that they need examples. Cycles appear as the definition of the abstract concept contains the examples, and the definitions of the examples contain the abstract notion as a genus.

In most of the small (2–4) components, we can only see that the words mutually depend on each other conceptually, especially in the case of 2-cycles.[9] Finally, most concepts have no out-edges, i.e. they are primitives of definition, so they form singleton strongly connected components.

3.3.3 *Weighting the concepts*

Circularity is an old problem of lexicography: if we say that a 'child' is one who has a 'parent', and 'parent' is one who has a 'child', we have not said much. As we have discussed in Section 3.2, some modern dictionaries avoid this problem by limiting the vocabulary of the definitions to the so called *defining vocabulary*, the meaning of which is assumed to be know. In the experiment of this section, we choose the opposite direction by characterizing the importance of defining words based on the dictionary, i.e. the structure where the words define each other.

The mathematical method used for this can be thought of as a *random walk* in the definition graph. We start the walk in a randomly chosen concept. The problem of the probability distribution from which this start concept is drawn, will be dealt with in the next paragraph on the so called damping. During the steps of the walk, we randomly take one of the concepts defining the current concept with a uniform distribution (taking multiplicity into account). For each concept, consider the probability that we will end up there after a long time. This is called the limit distribution of the random walk. This just expresses how important a given concept is in defining all the concepts, taking into account, recursively, how important is the concepts that we want to define.

The limit distribution is unique (that is, independent of the initial distribution) if and only if the graph consists of a single strongly connected component. We have seen that this is not the case in the `4lang` definition graph. PageRank is adapted for weighting the vertices of graphs

9 cause–reason, exist–real, hill–mountain, book–page, electricity–wire, programme–television, acid–sour, bottle–glass, now–this, attention–interesting, level–scale, balance/1607–weigh, dirt–dust, door–entrance, bell–ring/2735, brush–paint, thick/2134–thin/1038, problem–solve, hang–swing, dig–spade, elephant–trunk/1910, guest–host/2605, horse–ride, rat–rodent, news–newspaper, president–republic, school–student, soap–wash

consisting of more than one strongly connected components. Intuitively, there are two possibilities at each step of the abstract walk that defines PageRank: With a high probability d, we still go to one of the vertices directly accessible (uniform distribution with multiplicity). With the remaining low probability $1 - d$, we can go to any of the nodes. Note that the name is a bit misleading: a smaller damping factor leads to more uniform distribution. More formally, this means that if we go to node j with probability $P(i, j)$ in the original walk given we are currently in node i, the same transition probability in the damped walk will be

$$P_d(i, j) = \frac{1 - d}{n} + d \cdot P(i, j)$$

d is called the *damping factor* (most often $d = 0.85$) and n is the number of nodes. If the graph is strongly connected, and d goes to 1, the limit distribution approximates that of the original walk.

3.3.4 *Results*

Of course, the PageRank value depends on the damping factor d. We computed the PageRank of each node in the definition graph with three values of the damping factor: besides the standard $d = 0.85$, we took 0.9 and 0.95. In Figure 6 we explore the PageRank distributions. The vertical axis corresponds to the PageRank values obtained with different damping factors. The horizontal axis is aligned: the vocabulary is sorted by PageRank obtained with $d = 0.85$. Both axes are logarithmic. (Encyclopedic references in `4lang` like `@Koran` are omitted for reasons to be discussed later in this section.)

This plot shows, that all the tree PageRank distributions are approximated by a power law,[10] i.e. a few items receive quite a great weight and very many get very little. A higher damping factor increases the contrast: the highest ranks increase and the lowest ones decrease. This is as expected, if we recall that the name is misleading: smaller damping factors smooth the distribution better.

The ordinal rank of the 8 most important elements does not depend on the damping factor, however we see some instability with respect to the damping factor especially in the 100–1000 range. (The PageRank of many (121 out of 185) encyclopedic references like `@Koran` are very close to $4 \cdot 10^{-4}$ regardless of the damping factor. These 121 are used only once, e.g. `@Koran` is used only to define the word *Koran*, while some appear in more definitions, e.g. `@Arabia` is used in the definition of both *Arabia* and *Arabian*. These elements deviate from the power-law distribution. We omitted them from the figure to make the power law clearer.)

10 In this thesis, power-law is written with a hyphen, whenever it is an adjective, and without one, when it is a noun.

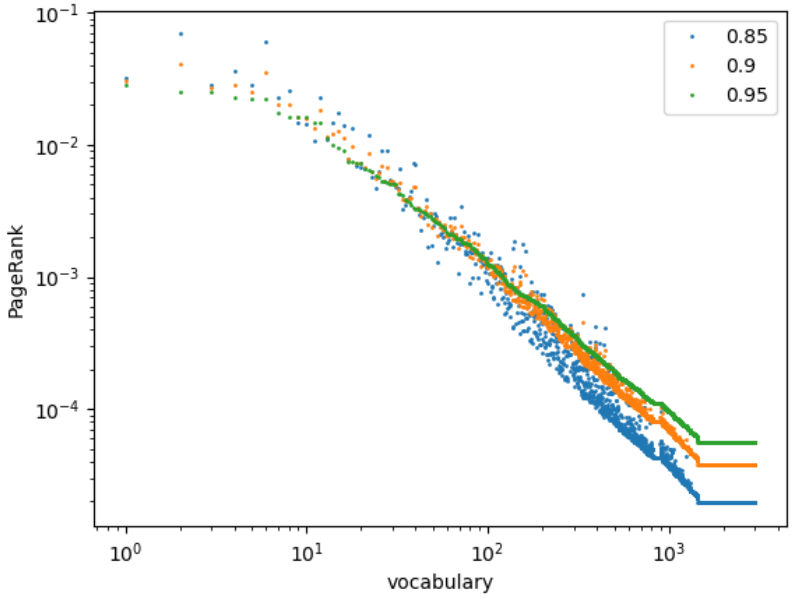

Figure 6: The three PageRank distributions. The vertical axis corresponds to the PageRank values, and colors correspond to different damping factors. The horizontal axis shows to the vocabulary sorted by the PageRank value obtained with $d = 0.85$. Both axes are logarithmic. Encyclopedic references were removed (see the main text). The plot shows, that all the tree PageRank distributions can be approximated with a power law, and a higher damping factor increases the contrast among concepts: the highest ranks increase and the lowest ones decrease.

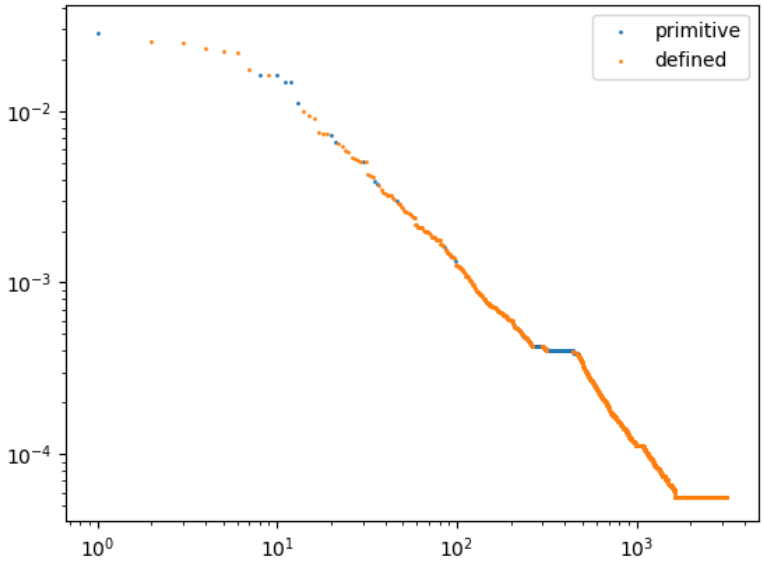

Figure 7: The PageRanks of primitive and defined concepts ($d = 0.85$), encyclopedic references included.

The tail consists of more than a thousand words with constant PageRank. This includes words like *presence, orange-colored, persuade, complicated*, and *-able*, which are defined, but not used in any definition.

The reader can investigate individual concepts in Figure 7 and Table 5, where we also indicate whether each concept is defined or left as a primitive. Clearly, the continuous values of PageRanks provide a more subtle description of the defining vocabulary then the binary distinction between primitive and defined concepts. Among the highest-ranking nodes of the definition graph, we can see binary relations (HAS, AT), deep cases (=PAT, =AGT), and unary concepts proper, especially members of 2-cycles (*exist, reason*). The high rank of some binary relations and deep cases are in line with the intuitive idea that understanding lexical relations and the arguments of a verb structure play a significant role in natural language understanding. These results are instructive in the sense that in order for a symbolic artificial intelligence system to be able to draw the right conclusions, it must first handle well the items at the top of the rankings.

3.3.5 *Conclusion*

We proposed a quantitative method to measure the importance of each word in the recursive process (Section 3.2) of word definition. The method is based on PageRank with the definition graph as an input. We applied the method to the graphical representation of the `4lang`'s manually-written word definitions. Most nodes in this graph represent

	index	$d = 0.85$	$d = 0.9$	$d = 0.95$	primitive
0	has	0.030417	0.031716	0.032239	True
1	=pat	0.028574	0.030711	0.032011	True
2	exist	0.025244	0.040721	0.069136	False
3	at	0.025138	0.027126	0.028517	False
4	reason	0.023064	0.028577	0.035985	False
5	cause	0.022473	0.025304	0.028615	False
6	real	0.022016	0.035551	0.059786	False
7	place/1026	0.017554	0.020090	0.022533	False
8	er	0.016251	0.020077	0.025601	True
9	in	0.016165	0.016037	0.014831	False
10	=agt	0.016092	0.015693	0.014193	True
11	lack	0.014789	0.013393	0.010658	True
12	=poss	0.014709	0.018331	0.023043	True
13	=rel	0.011136	0.011443	0.011095	True
14	quantity	0.009905	0.012027	0.014770	False
15	degree	0.009466	0.012834	0.017398	False
16	point	0.009130	0.011350	0.013958	False
17	man/659	0.007482	0.007818	0.007804	False
18	many	0.007392	0.009732	0.013207	False
19	after	0.007338	0.007264	0.006860	False
20	want	0.007262	0.007306	0.006783	True
21	part-of	0.006579	0.006701	0.006628	True
22	big	0.006449	0.008655	0.011969	False
23	object	0.006213	0.006275	0.005758	False
24	instrument	0.005876	0.005566	0.004672	False
25	contain	0.005749	0.006167	0.006241	False
26	large	0.005355	0.006930	0.009123	False
27	follow	0.005237	0.005347	0.005248	False
28	much	0.005152	0.006743	0.008978	False
29	move	0.005103	0.005299	0.005413	False
30	for/2782	0.005057	0.005303	0.005156	True
31	person	0.005057	0.004948	0.004791	False
32	do	0.004259	0.004536	0.004724	False
33	sex	0.004237	0.005192	0.006541	False
34	good	0.004148	0.003830	0.003253	False
35	before	0.003874	0.003773	0.003477	True
36	other	0.003758	0.003837	0.003824	True
37	live	0.003685	0.003953	0.004008	False
38	change	0.003494	0.003949	0.004467	False

Table 5: The PageRanks of primitive and defined concepts ($d = 0.85$).

concepts (unary predicates) and lexical relations (binary predicates, see the footnote in Section 3.1.3 for the difference). The place-holders of potential syntactic arguments are also represented by nodes (labeled with the deep cases – the thematic-role-style classes – of the argument, see), and so are encyclopedic references. This section adopts the Hungarian paper Makrai (2013) to the version of `4lang` which uses our deep cases. The most important nodes turn out to be lexical relations and deep cases, what motivates their further investigation in Chapters 5 and 7 respectively.

3.4 ANALYTIC PROPERTIES

What kind of information is included in `4lang` representations? Language philosophy and lexicography distinguish word meaning from other kinds of knowledge, while cognitive science and NLP put the emphasis on grounding linguistic knowledge in other capabilities of entities with natural or artificial intelligence such as vision and memory.

Kant (1781) introduced the distinction between analytic propositions, which are true by virtue of their meaning *(All bodies occupy space.)*, and synthetic propositions, that are true of their references in the real world *(All creatures with hearts have kidneys.)*. Within synthetic propositions, *a priori* and *a posteriori* propositions can be distinguished based on whether their justification relies upon experience. Logical positivists revisited the definition of analytic proposition as a proposition that is made true (or false) solely by the conventions of language.

W. v. Quine (1951) argued that the analytic–synthetic distinction is untenable despite "one [being] tempted to suppose in general that the truth of a statement is somehow analyzable into a linguistic component and a factual component." Wikipedia summarizes Quine's argument so that the notion of an analytic proposition requires a notion of synonymy (e.g. the proposition 'Bachelors are unmarried' is analytic because *bachelor* is synonymous with something like *older unmarried man*), but establishing synonymy inevitably leads to matters of fact via semantic equivalence.

Grice and Strawson (1956) offer a pair of thought experiments to restore the distinction. The protagonist of the first experiment says that *My neighbor's three-year-old child understands Russell's Theory of Types.* The other one says *My neighbor's three-year-old child is an adult.* The intended distinction is that it is *logically* impossible for a child of three to be an adult, and it is *naturally* impossible for a child of three to understand Russell's Theory of Types. "In both cases we would tend to begin by supposing that the other speaker was using words in a figurative or unusual or restricted way; but in the face of [their] repeated claim to be speaking literally, it would be appropriate in the first case to say that we did not *believe* [them], while in the second case [we would] say that we did not *understand* [them]."

For a deeper understanding of the `4lang` principle that the lexicographer should record analytic properties and disregard synthetic ones, the reader may refer to Section 5.7 of Kornai (2019), which heavily builds on the philosophical work in Putnam (1976), who "restored the honor of the analytical/synthetic distinction".

3.5 THE NAIVE MODEL AND AN ONTOLOGY

Our system is similar to truth-conditional semantics in that it can interact with models. There are more models: an internal one modeling linguistic meaning, and external models in charge of specific domain specific knowledge and reasoning (Nemeskey et al. 2013). In the preceding chapter, we reviewed many works emphasizing the role of naive theories in natural language semantics (Sections 2.2.6, 2.2.9 and 2.3.4). The internal model is different from that of modern sciences of the corresponding domains. E.g. the `4lang` definition of *heart* includes, besides the scientific truth that 'heart is an organ' and 'heart moves blood' the naive fact that 'love is in heart'. We define *death* as the end of life, though theology may state that life continues after death. As a third example, 'speed' is related to 'move' in `4lang`, but the exact nature of this relation which is explained in physics is not part of the naive world model, neither can it be expressed in `4lang`.

Gruber et al. (1993) defines an ontology as a formal, explicit specification of a shared conceptualization. In such Knowledge Representational terms, the core definitions, the main protagonists of this chapter, constitute the *top-level ontology* of the `4lang` meaning representation framework, keeping in mind that at this top level, we concentrate on linguistic meaning, and domain-specific knowledge can represented in external models.

3.6 FORMULAS

The main contribution of this chapter is the representation of a cca. 3000-word core vocabulary that, according to computations discussed in Section 3.2, is sufficient to define all the words in a dictionary. These core representations are written in `4lang` formulas that are compiled to `4lang` graphs by the pymachine software package.

`4lang` representations are graphs whose nodes are labeled by (mainly alphabetical) strings, the exponents of the concept that the node represents; edges have one of the colors 0, 1, and 2; and one node is distinguished as head-node. Such graphs can be specified by listing the nodes and the edges, but we maintain a formula representation as well which is more reminiscent of natural language definitions found in a dictionary. In this section, we describe the syntax of these formulas, i.e. the *minisyntax*, along with the graphs they are compiled to in `pymachine`, i.e. *minisemantics*. The minisyntax and the minisemantics together will

be called the *minigrammar*. (The terminology *metasyntax*, *metasemantics*, and *metagrammar* may be more familiar as they are the syntax and the semantics of some metalanguage, the object languages being natural languages, but we think that *meta* would suggest something impressive while minigrammar is a modest mechanism for creating `4lang` graphs in lexicographer-friendly fashion.)

The minigrammar was first published in Kornai et al. (2015) with the shortcoming that we did not make the head-node explicit, which made the formalism somewhat unclear. Figure 8 reproduces the grammar published there with some simplification in the system of non-terminals and indicating the head-node in each graph. The left column specifies how the graph representing the definiendum is built. There is always a *definiendu**m*** node denoted with m (labeled by the definiendum). The right column shows how a graph $g(X)$ representing the non-terminal X in the left side of the corresponding rule can be build from m and the graphs $g(Y)$ representing the yields of the non-terminals Y in the right side of the rule by drawing the edges from the head-node of some $g(Y_1)$ or m to that of some $g(Y_2)$ or m. The head-node of the resulting graph is emphasized by **boldface**.

Non-terminals of the minisyntax are D for a definition, E for an expression (subjunctive clause), E_u for a "unary expression" (subjunctive clause with unary head), U for (the label of) a unary node, B for (the label of) a binary node, and A for an argument of a binary node. The terminal `,` separates subjunctive clauses: a definition consist one or more clauses. Note that normal-font round parenthesises in this figure are used in regular expressions describing sentential forms, e.g. $(\texttt{,}E)^*$ is the Kleene-closure of `,`E, while the typewriter-font parens `(` and `)` are terminals of the minisyntax for 0-edges, e.g. `long(time)` compiles to `time` $\xrightarrow{0}$ `long`. Square brackets parenthesize arguments of nodes, mostly those of unary nodes (`air[move]`) and possibly those of binary ones (The definition of *put* is `=AGT CAUSE[=PAT AT/2744 =TO], =AGT MOVE =PAT, =PAT[object]`, in which the second argument of `CAUSE` is the patient (`=AGT`) being at the goal (`=TO`).)

Most unary predicates are lower-case strings that may include _ – see below for special cases. Ambiguous word-forms are disambiguated by appending the terminal `/` plus a numerical id to the end, e.g. `light/739` is the opposite of dark(ness) while `light/1381` is the opposite of heavy. In the theory, homonyms have to be disambiguated before the graphs are computed. This part was not implemented, the applications chose the fist `4lang` definition, when there were more.

From the point of view of the minigrammar, deep cases, the placeholders of arguments in representations of functions, are also unary labels despite their linking purpose. Deep cases are type-set as e.g. `=AGT` or `=TO`. Some unary nodes are encyclopedic references, these are prefixed with the terminal `@`, e.g. `@United_States`. Binary node labels are uppercase strings also allowing `_`. In this thesis, binaries are type-set

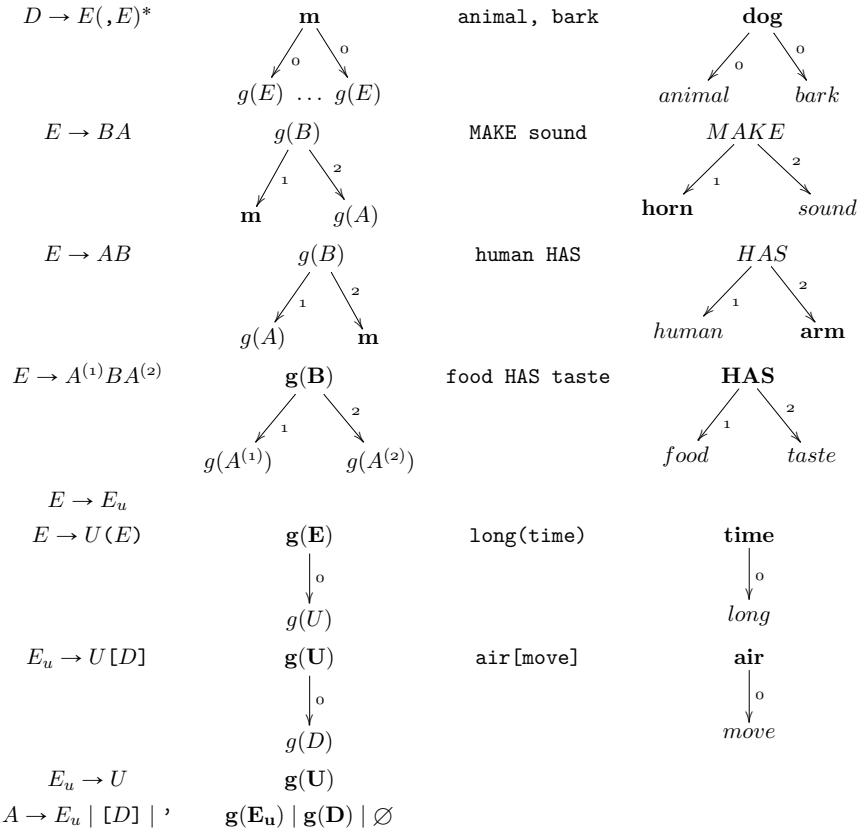

Figure 8: The original minigrammar

with SMALL-CAPS-AND-HYPHENS for aesthetic purposes), e.g. `HAS` and `PART_OF` are written as HAS and PART-OF.

In Section 2.2.9, we discussed to what extent concepts can be defined as conjunctions of other concepts. The first row in Figure 8 corresponds to the top level: the definition of a concept is a conjunction of properties. For the theoretic background, see section 3.3. of Kornai (2019) who defines 'dog' as 'four-legged, animal, hairy, barks, bites, faithful, and inferior'[11]. The next three lines represent binary predication. By default, the definition parser in `pymachine` draws a 0-edge from empty arguments of binary nodes to the definiendum m. This can be avoided by inserting the dummy argument `'` ($g(\text{'}) = \varnothing$). E.g. the definition of *place* is `point, ' AT/2744`: a place is a point, at which something else (something generic) is. Nodes with the same label get unified (see Section 2.3.5) unless there is the key-word `other` on their 0-th partition. E.g. the intended minisemantics of the definition of *think*, `=PAT IN/2757 mind, =AGT HAS mind`, is that the patient is in the mind that the agent has, because the to `mind` nodes are unified. Nevertheless, in the definition of *cross*, `shape, line AT/2744 other(line), symbol, <HAS upright(post/2740)>, <HAS horizontal>, <christian>`, the key-word `other` shows that the two occurrences of `line` should not be unified.

3.7 APPLICATIONS, INHERITANCE, AND NEGATION

No summer's high
No warm July
No harvest moon to light one tender August night
No autumn breeze
No falling leaves
Not even time for birds to fly to southern skies

— Stevie Wonder

We conclude this chapter with an overview of `4lang`'s applications along with related remarks on the representation of (word-level) negation and inference. Early work (Nemeskey, Recski, and Zséder 2012; Nemeskey et al. 2013) demoed `4lang` in a dialog system that answered questions about the time table and sold tickets. `4lang` was successfully applied to measure the similarity of English (Recski and Ács 2015; Recski et al. 2016; Recski 2016a, 2018) and Hungarian (Recski, Borbély, and Bolevácz 2016) words and English sentences (Kovács et al. 2020; Kovács, Gémes, Iklódi, et al. 2022). A key idea of these works is to represent words by their hand-written[12] or automatically extracted defini-

11 We give pre-theoretic meanings in 'single quotes', while `typewriter font` is kept for `4lang` formulas.

12 See the footnote in Section 5.5 on which applications used the hand-written definitions.

tions, and then replace defining words with their definitions (this kind of step is called *expansion*), and repeat this for some iterations. The principle is to reduce every word to the defining vocabulary. See the PhD thesis of Recski (2016b) as well.

Recski et al. (2016) discuss some features for inference in their Section 3.2. From the configuration `train` $\xrightarrow{0}$ `vehicle` $\xleftarrow{0}$ `car` they infer that *train* and *car* are somewhat similar (both are vehicles), and from `park` $\xleftarrow{1}$ `IN` $\xrightarrow{2}$ `town` and `street` $\xleftarrow{1}$ `IN` $\xrightarrow{2}$ `town` that so are *park* and *street* (both are in towns). A key point in inference is inheritance, which we introduced in Section 2.2.1. If we have `HAS wing` for all birds, `HAS wing` will also be true of all concepts for which $\xrightarrow{0}$ `bird` holds.

Inheritance is closely related to negation. Negation is expressed in the 2016 version of the hand-written `4lang` formulas by connecting a `lack` node to the 0th partition of the property which is lacking, e.g. by stating `diamond` $\xleftarrow{1}$ `HAS` $\xrightarrow{2}$ `color` $\xrightarrow{0}$ `lack` in the definition of *diamond*, we escape the (counterfactual) inference of concluding `diamond` $\xleftarrow{1}$ `HAS` $\xrightarrow{2}$ `color` from the conjunction of `diamond` $\xrightarrow{0}$ `mineral` $\xrightarrow{0}$ `substance` and `substance` $\xleftarrow{1}$ `HAS` $\xrightarrow{2}$ `color`. A broader discussion of negation in `4lang` can be found in Chapter 4 of Kornai (2023).

In the similarity experiments, Recski, Ács, Borbély, and Bolevácz utilized dependency parsing, and combined the manual definitions with those extracted from explanatory dictionaries. They refined the system with construction-specific rules. Both the agents (resp. patient) in the manual definitions and the subjects (resp. object) in the dependency analysis have been linked with a 1 (resp. 2) arrow. Combined with word embeddings and WordNet, they achieved the state of the art on SimLex 999 (Hill, Reichart, and Korhonen 2015) near to the correlation between a human annotator and the average of the other annotators.

Novák and Novák (2018a) transformed word embeddings into symbolic token-based semic representations, Their experiments involved `4lang`, as this framework „seemed to consists of a relatively coherent minimal set of semantic elements" (Attila Novák, in his pre-opponent's report on this thesis, translation by the thesis author).

Radim: "Was a story like nobody believed that it actually works, and you can do this sort of algebra with the vectors directly?"
Tomáš: "Oh, algebra, yeah, yeah, yeah."

— From a podcast with Tomáš Mikolov by Radim Řehůřek (21:20)

4

DISTRIBUTION AND VECTORS

Most contributions of this thesis are based on vector space language models (VSMs). This chapter provides a relatively complete history of these models going back to two interrelated families of word representations. The traditional method (Section 4.1) takes the co-occurrence matrix as a starting point, while more recent representations are learned as weights in shallow (Section 4.2) or deep (Section 4.3) neural networks.

The primary source of information about the meaning of a word is how often it is used in different contexts, an idea called the *distributional hypothesis* by linguists going back to Z. Harris (1951), and often quoted in the form that "You shall know a word by the company it keeps" (Firth 1957). The Saussurean definition of syntactic category (part of speech) is strikingly similar, the only difference in NLP practice appears to be how the context is defined (Sahlgren (2006), see Section 4.1.4): syntax is based on a short directed window (e.g. adjectives closely precede nouns) while semantic relations can be extracted from longer but symmetric windows (*dog* and *faithful* co-occur in sentences in any order).

One simple formalization of word distribution in a corpus is the *co-occurrence* matrix whose rows correspond to words in the vocabulary, columns to contexts, and cells contain the occurrence count of the word corresponding to the row appearing in the context corresponding to the column. What is meant by context depends on the application. In Latent Semantic Analysis (LSA, Deerwester, Dumais, and Harshman (1990), Section 4.1.3), columns of the original (unreduced) matrix correspond to documents. In matrix-based vector space language models (Turney and Pantel 2010) on the other hand, columns originally correspond to words, and counts express how often the words corresponding to the row and the column collocate in a window of some fixed length (say 5). Both in LSA and co-occurrence based VSMs, the number of contexts is at least in the thousands and gets reduced to some hundred dimensions for computation efficiency.

Neural language models (both shallow static ones, Section 4.2, and contextualized deep ones, Section 4.3), on the other hand, are neural nets, trained on gigaword corpora by iterating over words in their contexts and updating some weights of the model at each word. The resulting models represent similar words and sentences with similar vectors, and already static word embeddings reflect relational similarities between words like $\mathbf{king} - \mathbf{queen} \approx \mathbf{man} - \mathbf{woman}$ (Mikolov, Yih, and Zweig 2013).

4.1 MATRIX FACTORIZATION FOR WORD MODELING

4.1.1 *Semantic differential*

Vector space models of word meaning originate with psychological research by Osgood, May, and Miron (1975). In Osgood, May, and Miron's experiments, participants were asked to scale words like *freedom* on oppositional scales like *sturdy-fragile*, be the choice simple or abstract/metaphorical. Measurements were done in several languages with great typological care, and projected from the high-dimensional space of these oppositions to a three-dimensional space by principal component analysis (PCA). The emerging inter-lingual scales called EVALUATION,

POTENCY, and ACTIVITY turned out to explain much of the variation in the data. The method is called *semantic differential*. For details, see the last part of Section 2.7 in Kornai (2019).

4.1.2 *TF-IDF and PMI*

The next step in the history of VSMs was to gain the vectors from text corpora or, in the context of information retrieval, where the method got elaborated (Salton, Wong, and Yang 1975), from text documents. Classical methods start with a *frequency matrix*, more recent ones adjust association weights in artificial neural networks, but the mathematics these systems learn turn out to be variants of each other. Turney and Pantel (2010) discuss the history of VSMs arranged by what the rows and columns of the matrices correspond to, distinguishing *term–document*, *word–context* and *pair–pattern* matrices. Each cell contains the frequency of the term (or word, ...) corresponding to the row in the document (or context, ...) corresponding to the column.

Frequencies are adjusted to balance the effect of more frequent but less informative terms, or the variation in the length of the documents. The standard *weighting* technique comes from information retrieval, where the task is to return from a pool of documents the ones that are the most relevant for (similar to) a given query. (The query is also treated as a document) This weighting is tf-idf (term frequency–inverse document frequency) scoring, but there are other methods as well.

In NLP, the information-theoretic association scores *pointwise mutual information (PMI*, Church and Hanks (1990))

$$PMI(x,y) = logP(x,y)/P(x)P(y)$$

and *positive pointwise mutual information* (PPMI, Niwa and Nitta (1994))

$$\max\{0, PMI(x,y)\}$$

became standard, and Levy and Goldberg (2014c) showed (as we will see in Section 4.2.5) that the more recent word2vec is mathematically equivalent to a variant of PMI, *shifted PMI*.

Besides weighting, matrices also have to be *smoothed* to reduce the amount of random noise and to fill in some of the zero elements in a sparse matrix. Semantic differential (Section 4.1.1) applies PCA, which computes word representations from the raw term–document matrix. PCA requires inverting the data matrix. This became feasible for thousand-row matrices in the past decades, resulting in the method called Latent Semantic Analysis, what we turn to now.

4.1.3 *Latent semantic analysis*

The main pre-neural method, which has remained an important reference point in the word embedding era (Tsvetkov, Faruqui, and Dyer 2016; Antoniak and Mimno 2018), is Latent semantic analysis (LSA, Dumais et al. (1988) and Furnas et al. (1988)). Landauer, Foltz, and Laham (1998) introduce LSA in two ways.

On the practical side, it is a method for obtaining approximate estimates of the contextual substitutability of words in text, and similarities among words and text segments. On the cognitive side, it is a model of the computational processes and representations underlying the acquisition and utilization of knowledge. While we think that it rather depends on the scientific taste of the researcher whether they motivate their work with such acquisitional claims, the practical importance of LSA in pre-embedding NLP is beyond debate. For a recent overview of LSA methods in psychology, especially author modeling, automated grading, and change over time, see Iliev, Dehghani, and Sagi (2014, Section 1.4).

Closer to the mathematical content is the way to think of LSA as representing the meaning of a word as the average of the meaning of all the passages in which it appears, and dually, the meaning of a passage as an average of the meaning of all the words it contains. The choice of dimensionality can be of great importance. LSA can be motivated in a way that the resulting dimensions may be analogous to the semantic features often postulated as the basis of word meaning, but establishing specific relations to mentalisticly interpretable features poses daunting technical and conceptual problems. It may worth noting that LSA arrived at the same dimensionality (300), as word embeddings did (Section 4.2). The effective usage of LSA is a process of very sophisticated tuning and can be viewed as a kind of art. The main factors are preprocessing (stopwords, stemming), frequency matrix transformations, the choice of dimensionality, and, the choice of similarity measure. For an early study on the impact of weight functions choice, see Nakov, Popova, and Mateev (2001).

The authors point out that transformation of co-occurrence counts to log frequency divided by entropy and followed by dimensionality reduction is reminiscent of information retrieval methods, and the psycholinguistic reality of the dimensionality reduction step is often implicit and sometimes explicit in many neural net and spreading-activation architectures. The similar equivalence between word embeddings and pointwise mutual information will be discussed in Section 4.2.5.

4.1.3.1 *Singular Value Decomposition*

Data preprocessing transformations in LSA need to be described in more detail. LSA subjects the data in the raw word-by-context matrix to a $\log(x+1)$ transformation, and then each cell entry is divided by the

row entropy value. The result is an estimate of the word's importance in the passage, the degree to which knowing that a word occurs in a passage provides information about the passage.

Singular value decomposition (SVD) is the general method for linear decomposition of a matrix into independent principal components of which factor analysis is the special case for square matrices. For the reader who is not familiar with or interested in multivariate statistics, we cite Landauer, Foltz, and Laham (1998)'s elevator-pitch description of factor analysis as finding a parsimonious representation of all the intercorrelations between a set of variables in terms of a new set of abstract variables, each of which is unrelated to any other but which can be combined to regenerate the original data. SVD does the same thing for an arbitrarily shaped rectangular matrix, including the case when columns stand for words, and rows for contexts. (See the formulas in Section 4.2.8.3.) In the process, cells in the matrix originally contain the frequency. The raw cell entries f are first transformed to $\ln(1+f)/e$ where e is the entropy of the word over all contexts. This matrix is then submitted to SVD and the — for example — 300 most important dimensions are retained (those with the highest singular values, i.e. the ones that capture the greatest variance in the original matrix). The resulting vectors of 300 real values represent each word and each context. Similarity has been usually measured by the cosine of the angle between the vectors.

Related to LSA is a generative method called Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003), where each document is supposed to be composed of a mixture of topics. While the dimensions of LSA may be regarded as abstract and meaningless, the dimensions in LDA correspond better to latent topics that emerge from the corpus.

4.1.4 *Relation to structuralist linguistics*

Now we summarize Sahlgren (2006), who investigates the relation between the word-space model and structuralist linguistics.

4.1.4.1 *Rethinking the distributional hypothesis: syntagma and paradigm*

The distributional hypothesis, as motivated by the work of Zellig Harris, states that differences of meaning correlate with differences of distribution, but he neither specifies what kind of distributional information we should look for, nor what kind of meaning differences it mediates.

Syntagmatic relations concern positioning, as already the Greek word *syntag-matikos* 'arranged, put in order' shows.[1] They relate entities that co-occur in the text. They are linear, and apply to linguistic entities

1 The PDF version of Sahlgren (2006) I have access to transcribes the first vowel of the Greek word as *u*, but – as Attila Novák pointed out – the original must be an *υ*, and the transcription, accordingly a *y*.

Test	Which relation? (Is essential?)	Context
Thesaurus	both (−)	large
Association	syntagmatic (+)	small
Synonym	paradigmatic (+)	narrow
Antonym	paradigmatic (−)	wide
POS	paradigmatic (+)	narrow

Table 6: Test, relations they rely on, the degree to which the relations are essential to the test (− and +), and the context that yields the best results in the strict evaluation settings (Sahlgren 2006, Table 15.6). The thesaurus task is to list words with related meanings to the query.

that occur in sequential combinations. They are combinatorial relations, which means that words that enter into such relations can be combined with each other. A *syntagma* or syntagm is such an ordered combination of linguistic entities: written words are syntagms of letters, sentences are syntagms of words.

Paradigmatic relations, on the other hand, concern substitution. The Greek word *paradeigmatikos* means serving as a model. Saussure himself never used the word *paradigmatique*. It was Hjelmslev who coined the term as a substitute for Saussure's *associative meanings*. Paradigmatic relations are between entities that do not co-occur in the text. They hold between linguistic entities that occur in the same context but not at the same time. A paradigm is a set of such substitutable entities, usually depicted as orthogonal axes in a grid.

Although Harris was arguably more directly influenced by the works of Bloomfield than of Saussure, the latter's structuralist legacy is foundational for both Bloomfield's and Harris' theories. In Sahlgren's view, *the Saussurian refinement* of the distributional hypothesis clarifies the semantic requirements of the word-space model and the distributional methodology. A word-space model accumulated from co-occurrence information contains syntagmatic relations between words, while one from information about shared neighbors contains paradigmatic relations.

4.1.4.2 *The semantic continuum*

Sahlgren's point is that syntagmatic and paradigmatic relations between words should be discoverable by using co-occurrence information and information about shared neighbors in the word-space, respectively. *A qualitative comparison between different uses of context* e.g. in LSA (Section 4.1.3) or other models should be able to divulge the difference by empirical investigation. Sahlgren is interested in what these different

uses of context entail, what their differences are, and how they can be used to build word spaces.[2]

Sahlgren's thesis is split to background chapters, "setting the scene" chapters, and foreground chapters, a structure we followed in the present thesis to some extent. The latter contain *experiments* demonstrating the differences between syntagmatic and paradigmatic uses of context: small context regions yield more syntagmatic word spaces, while wide context windows yield more paradigmatic spaces, as can be seen in Table 6. Only a small percentage of the nearest neighbors occur in both syntagmatic and paradigmatic word spaces.

Sahlgren investigates three *parameters* of the characterization of paradigmatic contexts:

- the size of the context region,
- the position of the words within the context region, and
- the direction in which the context region is extended. The only experiment he was aware of exploiting the directional information in a words-by-words co-occurrence matrix was Schütze (1993).

In his experiments, Sahlgren compares different *weighting schemes* of the slots for the paradigmatic uses. The two extremes are constant weighting over the window, and aggressive distance weighting according to the formula 2^{1-l}, where l is the distance to the focus word. Possibilities in between include linear distance weighting and $1/l$.

In the concluding chapter, Sahlgren answers his research questions. Is it at all possible to extract semantic knowledge by merely looking at usage data? Clearly, yes. Does the word-space model constitute a complete model of the full spectrum of meaning, or does it only convey specific aspects of meaning? It is complete as far as it reflects a structuralist dichotomy of syntagma and paradigm. If we believe that meaning is essentially referential, then no.

4.1.4.3 *"Future" work*

The future work section lists problems related to which much has been achieved since 2006, but they still remain major problems. One is that word spaces may have (i) a common internal structure that can be

2 While a bit irrelevant for the purposes of the present thesis, it is interesting what Sahlgren thinks about the use of a *document* as a context. Word-space algorithms that prefer a syntagmatic use of context, such as LSA, hail from the information retrieval community, where a *document* is a natural context of a word. But "document" in the sense of a topical unit is an artificial notion that hardly exists elsewhere; before the advent of library science, the idea that the content of a text could be expressed with a few index terms must have seemed strange. In the "real" world, content is something we reason about, associate to, and compare. In the world beyond information-retrieval, text is a continuous flow where topics intertwine and overlap and the notion of a "document" is at best an arbitrary choice. In a whole document nearly every term can co-occur with every other.

utilized to differentiate between different types of relations within the word space; and (ii) a discoverable "latent" dimensionality. While compositionality is not without controversy in the philosophy of language, word-space models may be extended to handle phrase, clause, sentence, paragraph, "document" and text level meaning too. The word-space model may have the flexibility and ability to continuously evolve when subjected to a continuous data flow.

Finally, Sahlgren remarks that the word-space model is not a psychologically realistic model of human semantic processing. It is arguable that humans also use extra-linguistic context when learning, understanding, and using language. The inability to reach beyond the limits of textuality is the most disqualifying feature of the word-space model with regard to the referential aspect of meaning.

4.1.5 *A compression-based method*

Cilibrasi and Vitányi (2004) present a similarity measure between words and phrases based on information distance and Kolmogorov complexity, using Google page counts. In the Turney and Pantel (2010) classification, this is a term–document model. This similarity measure is the special case of a compression-based universal similarity metric among objects given as finite binary strings. These strings include genomes, music pieces in MIDI format, computer programs, pictures in simple bitmap formats, or time sequences such as heart rhythm. The universal metric is feature-free in the sense that it does not look for particular features, but analyzes all features simultaneously and determines the similarity between every pair of objects according to *the most dominant shared feature*. The word similarity measure is based on "the Google semantics of a word or phrase", i.e. the set of web pages returned by the query concerned.

They normalize the introduced distance to make it relatively stable with respect to the index size (Normalized Google Distance, NGD). The NGD of *horse* and *rider* is 0.443. The distance is usually between 0 (identical) and 1 (unrelated), but not always (see below). If the distance is calculated from the index of only one-half of the pages, this distance only deviates to 0.460.

A drawback of the Google semantics is that terms with different meaning may have the same semantics, especially *opposites* often have a similar semantics. The paper offers more literature (of course, from before 2005) on how representative Google hits are for language. The theoretical underpinning is based on the theory of Kolmogorov complexity, in terms of coding and compression. Let G denote the prefix-code word length defined from the relative frequency of the hits. The NGD formula

$$NGD(x,y) = \frac{G(x,y) - \min(G(x), G(y))}{\max(G(x), G(y))}$$

$$= \frac{\max(\log f(x), \log f(y)) - \log f(x,y)}{\log N - \min(\log f(x), \log f(y))}$$

is similar to many earlier formulas in this area, but not equivalent to any of them.

It has to be noted that the returned *Google counts are approximate*. The situation used to get worse if one used the boolean OR operator between search terms, but the measure is based on the AND operator, which is less problematic. When the paper was written, Google already estimated the number of hits based on samples, and the number of indexed pages already changed rapidly. To compensate for the latter effect, the authors have inserted a normalizing mechanism. Web searches for rare two-word phrases correlated well with frequency in traditional corpora, as well as with human judgments.

4.1.5.1 *Kolmogorov complexity, information distance, compression-based similarity*

Information can be compressed to different extents. The Kolmogorov complexity $K(x)$ is the length, in bits, of the ultimate compressed version from which x can be recovered by a general decompression program. An earlier paper considered the following information distance $E(x,y)$: given two strings x and y, what is the length of the shortest binary program in the reference universal computing system such that the program computes output y from input x, and also output x from input y. Up to a negligible logarithmic additive term, $E(x,y) = K(x,y) - \min K(x), K(y)$, where $K(x,y)$ is the binary length of the shortest program that produces the pair x, y and a way to tell them apart. This distance $E(x,y)$ is actually a metric.

E is *universal* for the family of computable distances, i.e. E minorizes every admissible distance up to an additive constant, where admissible means non-negative, symmetric, and computable. More intuitively, this means that the information distance determines the distance between two strings minorizing the dominant feature in which they are similar. This measure has to be normalized, because if small strings differ by an information distance which is large compared to their sizes, then the strings are very different. The normalized information distance (NID) has values between 0 and 1, and it is universal: minorizes, up to a vanishing additive term, every other possible normalized computable distance. The NID is uncomputable since the Kolmogorov complexity is uncomputable, but we can use real data compression programs to approximate the Kolmogorov complexities $K(x)$, $K(y)$, $K(x,y)$.

4.1.5.2 *Google distribution, Normalized Google Distance, and their universality*

We cannot use the probability of the events directly to determine a prefix code, or, rather the underlying information content implied by the probability because events overlap and hence the summed probability exceeds 1. But absolute probabilities allow us to define the associated prefix code-word lengths (information contents) for both the singletons and the doubletons.

The Google Similarity Distance has the following properties:

- The range of the NGD is basically in between 0 and ∞. More precisely, it is slightly negative if the Google counts are untrustworthy and state $f(x, y) > \max\{f(x), f(y)\}$.
- If $f(x) = f(y) = f(x, y) > 0$, then $NGD(x, y) = 0$.
- If frequency $f(x) = 0$, then for every search term y we have $NGD(x, y) = \infty/\infty$, which we take to be 1 by definition.
- NGD is always non-negative and $NGD(x, x) = 0$ for every x.
- NGD is symmetric ($NGD(x, y) = NGD(y, x)$).
- The NGD does not satisfy the triangle inequality, i.e. NGD is not a metric.

The paper includes clustering and classification experiments (against WordNet, see Section 2.4.3) to validate the universality, robustness, and accuracy of the proposal.

4.1.6 *Mathematical processing*

Now we summarize Turney and Pantel (2010, Section 4)'s discussion of the mathematical processing for distributed word models. This will be especially important in Chapter 6.

First the frequency matrix is built by scanning sequentially through the corpus, and recording events and their frequencies in a hash table, a database, or a search engine index. The frequency matrix has to be represented in a sparse way (i.e. most items are 0).

4.1.6.1 *Weighting the Elements*

The weights of the elements in the matrix have to be adjusted, because common words will have high frequencies, yet they are less informative than rare words. Information retrieval uses the tf-idf (term frequency $\times$ inverse document frequency) family of weighting functions, where an element gets a high weight when the corresponding term is frequent in the corresponding document (i.e. tf is high), but rare in other documents in the corpus (i.e. df is low). Document length has to be normalized.

Affixation, especially derivational affixation is problematic both from linguistic and computation point of view. The linguistic problem is to delineate the inventory of compositional affixes. The compositional problem is that though different forms of the same lexeme are correlated, yet we may not want to lemmatize them, because they may have slightly different meanings. An idea that did not become standard is to reduce the weights of derivatives when they co-occur in a document.

A key step in pre-neural machine learning was feature selection. One of the most popular word association scores remains Pointwise Mutual Information, which we will discuss in detail in Section 6.2.

4.1.6.2 *Smoothing the Matrix*

The goal of smoothing the matrix is to reduce the amount of random noise and to fill in some of the zero elements that are due to data sparsity. The other direction, sparsification is a hot topic today (Sanh et al. 2019), but it goes beyond the limits of this thesis. The mathematical method of truncated (or thin) Singular Value Decomposition (SVD) is standardly applied to either document similarity (Latent Semantic Indexing), or word similarity (Latent Semantic Analysis, Section 4.1.3).

SVD decomposes X into the product of three matrices $U\Sigma V^\intercal$, where U and V are in column orthonormal form (i.e. the columns are orthogonal and have unit length, $U^\intercal U = V^\intercal V = I$), and Σ is a diagonal matrix of singular values. If X is of rank r, then Σ is also of rank r. Let Σ_k, where $k < r$, be the diagonal matrix formed from the top k singular values, and let U_k and V_k be the matrices produced by selecting the corresponding columns from U and V. The matrix $U_k\Sigma_k V_k^\intercal$ is the matrix of rank k that best approximates the original matrix X, in the sense that it minimizes the approximation errors. That is, $\widehat{X} = U_k\Sigma_k V_k^\intercal$, which is called the *truncated SVD*, minimizes $|\widehat{X} - X|_F$ over all matrices $\widehat{X}$ of rank k, where $|\ldots|_F$ denotes the Frobenius norm.

The authors list four aspects of what SVD is looking for: latent meaning, noise reduction, indirect or high-order co-occurrence (when two words appear in similar contexts), or sparsity reduction. Truncated SVD implicitly assumes that the vectors have a Gaussian distribution – Minimizing the Frobenius norm $|\widehat{X} - X|_F$ will minimize the noise, if the noise has a Gaussian distribution – but this assumption is not satisfied by word frequencies.

4.1.6.3 *Comparing the Vectors*

There are many different ways to measure the similarity of two vectors, but the most popular one is clearly cosine similarity, while the most intuitive one remains the Euclidean distance. In classical information retrieval, it has been commonly said that, properly normalized, the difference in retrieval performance using different measures is insignificant. Distances include the Manhattan distance, or, from information theory,

Hellinger, Bhattacharya, and Kullback-Leibler. Dice $2xy/(|x|^2 + |y|^2)$ and Jaccard have set-theoretic motivation.

Lee (1999) gives the principle that measures that focused more on overlapping coordinates and less on the importance of negative features (i.e. coordinates where one word has a non-zero value and the other has a zero value) appear to perform better. In her experiments, the Jaccard, Jensen-Shannon, and L1 measures seemed to perform best.

Other researchers studied the linguistic and statistical properties of the similar words returned by various similarity measures and found that the measures can be grouped into three classes: high-frequency sensitive measures, low-frequency sensitive measures, similar-frequency sensitive methods. Given a word w_0, if we use a high-frequency sensitive measure to score other words w_i according to their similarity with w_0, higher frequency words will tend to get higher scores than lower frequency words. If we use a low-frequency sensitive measure, there will be a bias towards lower frequency words. Similar-frequency sensitive methods prefer a word w_i that has approximately the same frequency as w_0.

4.1.6.4 *Efficient comparisons*

One section in Turney and Pantel (2010) discusses methods like distributed sparse matrix multiplication and Random Indexing. Randomized algorithms are based on the idea that high-dimensional vectors can be randomly projected into a low-dimensional subspace with relatively little impact on the final similarity scores. Random Indexing (RI) is an approximation technique that computes the pairwise similarity between all rows (or vectors) of a matrix. There are index vector elements of which are mostly zeros with a small number of randomly assigned $+1$'s and -1's. The cosine measure between two rows r_1 and r_2 is then approximated by computing the cosine between two fingerprint vectors, fingerprint(r_1) and fingerprint(r_2), where fingerprint(r) is computed by summing the index vectors of each non-unique coordinate of r.

Locality sensitive hashing (LSH, Broder (1997)) is similar technique. LSH functions include the Min-wise independent function, which preserves the Jaccard similarity between vectors, and functions that preserve the cosine similarity between vectors.

4.2 NEURAL WORD EMBEDDINGS

4.2.1 *Symbolic structures in connectionism*

As a thesis submitted to a theoretical linguistics programme, this work may start its account of neural word models with Rumelhart and McClelland (1986), a paper from the same year as Hinton, McClelland, and Rumelhart (1986), which proposed a distributed representation of words. Rumelhart and McClelland's paper belongs to the infamous

past tense debate between connectionists and discrete-minded scholars. However, we prefer taking our ideological heritage from Smolensky (1990, Section I), which we summarize now.

4.2.1.1 *Discrete and continuous computations*

Connectionist models rely on parallel numerical computation rather than the serial symbolic computation of traditional artificial intelligence (AI) models. Smolensky argues that connectionist models will offer an opportunity to escape the brittleness of symbolic AI systems, and develop more human-like intelligent systems, but only if we can find ways of naturally instantiating the sources of power of symbolic computation within fully connectionist systems. The connectionist approach, on the one hand, is an excellent opportunity for formally capturing the subtlety, robustness, and flexibility of human cognition, and for elucidating the neural underpinnings of intelligence. The symbolic approach, on the other, has provided tremendous insights into the nature of the problems that must be solved in intelligent systems, and of techniques for solving these problems.

The paper is part of an effort to extend the connectionist framework to naturally incorporate symbolic computation, without losing the virtues of connectionist computation; i.e. integrate the discrete mathematics of symbolic computation and the continuous mathematics of connectionist computation. Language can be represented by objects like a phrase-structure tree, or even as a simple sequence of words. The representation problem is characterized as finding a mapping from the set of structured objects to a vector space.

Smolensky takes an analogy from mathematics: representing abstract groups as collections of linear operators on a vector space. Discrete group theory and the continuous vector space theory interact, and this relation extends to applications like quantum physics. In physics, elementary particles involve a discrete set of particle species which exhibit many symmetries, that are described by group theory. Yet underlying elementary particle state spaces are continuous.

In human language processing, the discrete symbolic structures that describe linguistic objects are actually "imbedded" in a continuous connectionist system that operates on them with flexible, robust processes that can only be *approximated* by discrete ones. Smolensky refers to structures as symbolic ones, because the principal cases of his interest are objects like strings and trees, however, his analysis is of structured objects in general; it applies equally well to objects like images and speech trains. (His view is not that mental operations are always serial symbol manipulations, but that the information processed often has useful symbolic descriptions.)

Smolensky seeks a fully distributed representation in which each output neuron participates in the representation of many different outputs. In the tensor product representation he proposes, both the variables

and the values can be arbitrarily nonlocal, enabling (but not requiring) representations in which every unit is part of the representation of every linguistic constituent in the structure. The representation can be used recursively, and connectionist representations of operations on symbolic structures and recursive data types, can be naturally analyzed.

4.2.1.2 *Why inject symbolic structure in a neural network?*

The motivation for pursuing the representation of symbolic structures in connectionist systems lies in the connectionist modeling of higher cognitive processes such as language. Here the central question is: What are computationally adequate connectionist representations of strings, trees, and sentences? The essence of the connectionist approach, people might say, is to expunge symbolic structures from models of the mind. But a reasonable starting point is to take linguistic analysis of the structure of linguistic objects seriously, and to find a way of representing this structure in a connectionist system: it is important to find adequate connectionist representations of these trees or strings. Smolensky' hope is that new connectionist representations of linguistic structures will rest on prior understanding of connectionist representations of existing symbolic descriptions of linguistic structure. The importance of representing linguistic structures exceeds NLP: these representations are the basis for connectionist models of conscious, serial, rule-guided behavior: all higher thought processes.

One argument against designing a connectionist representation of symbolic structures goes like this: Just as a child somehow learns to internally represent sentences with no explicit instruction on how to do so, so a connectionist system with the right learning rule will somehow learn the appropriate internal representations; The problem of linguistic representation is not to be solved by a connectionist theorist but rather a connectionist network. Smolensky's response is the following:

- In the short term, at least, our learning rules and network simulators do not seem powerful enough for unstructured learning,
- we will still need to explain how the representation is done,
- we should build bridges as soon as possible between accounts of language; the problem is just too difficult to start all over again from scratch,
- to experiment now with connectionist learning of rather complex skills (e.g. parsing, anaphoric resolution, and semantic interpretation, all in complex sentences), we need connectionist representation of the input and output. We want to study the learning of the operations without waiting for the discovery of the linguistic representations.

- Language is more than just a domain for building models: it is a foundation on which the entire traditional theory of computation rests. It is crucial for how the basic concepts of symbolic computation and formal language theory relate to connectionist computation.

4.2.2 *Neural language modeling*

At least before the neural revolution in NLP, the term *language modeling* was restricted to the task of "predicting the next word", which is equivalent to computing the probability (naturalness) of a word sequence. Probabilities are estimated using (relative) frequencies. As there are infinitely many possible sentences but the model is trained on a finite sample, the main point is in generalization. A simple and effective approach to language modeling is the family of n-gram models (Brown et al. 1992) that make the Markov assumption, i.e. the simplifying assumption that the probability of a word in a context depends only on preceding words of some fixed number (four in most applications of the time). Thus the probability of the Hungarian word string *minden madár társat választ* ('every bird is choosing a mate')[3] is computed as

$$P(\texttt{\^{} minden madár társat választ \$}) =$$

$$P(\texttt{ minden } | \texttt{ \^{} }) \cdot P(\texttt{ madár } | \texttt{ minden }) \cdot P(\texttt{ társat } | \texttt{ madár }) \cdot$$
$$\cdot P(\texttt{ választ } | \texttt{ társat }) \cdot P(\texttt{ \$ } | \texttt{ választ })$$

$P(\texttt{ madár } | \texttt{ minden })$ denotes the probability of the word *madár* given that the preceding word was *minden*. `^` and `$` denote the beginning and the end of the string, respectively. While n-gram models are easy to understand and useful in application, they have the disadvantage of not capturing morphological and semantic relations between words. This is the problem that the neural language model (Bengio et al. 2003) solved.

Bengio et al. (2003) implement the n-gram language model relying on shared-parameter multi-layer neural networks. Their network has millions of parameters, and it is trained on tens of millions of examples. Training such large-scale model is expensive but feasible, scales to large contexts, and yields good comparative results.

The idea of fighting the so called *curse of dimensionality* with distributed representations is summarized by the authors as associating with each word in the vocabulary a distributed word feature vector (a real-valued vector in R^m); expressing the joint probability function of word sequences in terms of the feature vectors of these words in the sequence; and learning simultaneously the word feature vectors and the parameters of the probability function. The objective can be the log-likelihood of the training data or a regularized criterion, e.g. by adding

3 This sentence is from the song that gave the title of the conference where Makrai (2014b) was published.

a weight decay penalty i.e. like in ridge regression, the squared norm of the parameters as a penalty.

The paper cites a rich collection of related work for the idea of using neural networks to model high-dimensional discrete distributions and, from the early days of connectionism, the idea of learning a distributed representation for symbolic data. In their view, neural networks for language modeling are not new either in the field of character-level LM based neural text compression with or without hidden units and one or more input words. What is more well known, generalization from training sequences is and has been been obtained in the form of similarities between words: clusterings of the words with words associated deterministically or probabilistically with classes. Vector-space representation for words has been well exploited in the context of an n-gram based statistical language model, using LSI to dynamically identify the topic of discourse. Finally, vector-space representation for symbols in the context of neural networks, and especially a parameter sharing layer, was pioneered in text-to-speech mapping.

Bengio et al. (2003) is the kind of paper whose future work section forecast the most important steps of the next 10-15 years, especially hierarchical softmax (Section 7.4.4.1, Morin and Bengio (2005)), the recurrent language model (Mikolov 2010), negative sampling (Mikolov, Sutskever, et al. 2013), "interpreting (and possibly using) the word feature representation" (Mikolov, Yih, and Zweig 2013), and sub-word encoding (Bojanowski et al. 2017). A section sketches an energy-based extension.

We used the Hierarchical Log-Bilinear extension (HLBL, Mnih and G. E. Hinton (2009)) of the neural word model for this thesis (Sections 7.2 and 7.3). The model is called log-bilinear because it models the co-occurrence of two words as proportional to $\exp(u^\top \cdot v)$ where u and v are the corresponding word vectors.

Probabilistic modeling proper means that the sum of the co-occurrence probabilities, $Z = \sum_{u \in V} p(u \mid v)$, a.k.a. the partition function (V is the vocabulary) should be equal to 1. Z is very costly to compute. Hierarchical modeling, most notably hierarchical softmax, solves this problem by organizing the vocabulary at the leaves of a binary tree, and reducing the choice of a word to a series of binary choices among the path leading to the corresponding leaf. The choice at each node is accounted for by a two-valued probabilistic variable, which makes the partition function trivial.

4.2.3 *Parameter sharing and noise-contrastive estimation*

One of the key components of the NLP advances in the last decade is parameter sharing (Bengio, Courville, and Vincent 2013) in the form of unsupervised pre-training introduced by Collobert et al. (2011), who train a single model for tasks including part-of-speech tagging, chunk-

ing, named entity recognition, and semantic role labeling. The system learns internal representations based on vast amounts of mostly unlabeled training data. This representation is then used as a basis for building a freely available tagging system with good performance. The architecture is similar to Bengio et al. (2003)'s language model discussed in the previous section, but it uses noise contrastive estimation to spare the computation of the normalization term needed for probabilistic modeling. A couple of years later, noise-contrastive estimation, or simply *negative sampling*, became an ingredient of the very influential skip-gram model we will see in the next section. Besides its great importance in the development of VSMs, Collobert et al.'s work has also relevance for this thesis because we used their vectors (SENNA) in work presented in Sections 7.2 to 7.4.

This work is also one of the most remarkable linguistic applications of one of the major neural architectures, *convolutional neural networks*, which was originally invented for computer vision. The window approach described so far performs well for most NLP tasks Collobert et al. choose, but it fails with semantic role labeling (SRL), where the predicate may fall outside the window. This task requires the consideration of the whole sentence. Among the main neural networks architectures, one of the natural choices to tackle this problem in a convolutional networks.

A convolutional network is a sequence of alternating convolutional and pooling layers. A convolutional layer is a generalization of a window approach: given a sequence represented by columns in a matrix, a matrix-vector operation is applied to each window of successive windows in the sequence, where the weight matrix is constant across all windows. Convolutional layers extract local features around each window, and they are often stacked to extract higher level features.

The size of the output of the convolutional layer depends on the number of words. Local feature vectors extracted by the convolutional layers have to be combined to obtain a global feature vector, with a fixed size, in order to apply subsequent layers. Traditional convolutional networks often apply a (possibly weighted) average or a max operation over "time". Average does not make much sense in the SRL case, as in general most words in the sentence do not have any influence on the semantic role of a given other word. So the authors used a max approach. The network finally produces one score per possible tag for the given task, as in the window approach.

4.2.4 *word2vec*

Deeper in its effect on the broad NLP community than in its architecture, the first wave of the neural revolution was pre-trained *word embeddings*, word models learned by shallow neural networks in an unsupervised way, which have become very popular since Mikolov, Sutskever,

et al. (2013), who implemented[4] a log-bilinear model to learn continuous representations of words on very large corpora efficiently. These more accurate variants of earlier VSMs, map "similar" word to similar vectors in a space of some hundred dimensions. Word similarity covers syntax and semantics, and vector similarity is mostly measured by cosine similarity. Embeddings also reflect analogical quadruples (Mikolov, Yih, and Zweig (2013), Section 4.2.7) like

$$\textbf{woman} - \textbf{man} \approx \textbf{queen} - \textbf{king}$$

Mikolov, Le, and Sutskever (2013) discovered that VSMs of different languages have such similarities that a *linear* mapping can map the representations of words in a source language to the representation of their *translations*, see Sections 7.4, 7.5 and 8.4 for details.

Most of the main contributions of this thesis are related to the word2vec line of research. Sections 7.2 and 7.3 investigate two lexical relations with the vector offset method of Mikolov, Yih, and Zweig (2013), Section 7.4.1 offers a Hungarian equivalent of the analogical test set, Section 7.4.2 to compares word embeddings based on the linear method for dictionary induction. Section 7.5 utilises the confidence score obtained in linear translations to develop the triangulation method of dictionary induction, and Chapter 8 puts linear translation in the context of cross-lingual word sense induction by computing an upper bound on the precision of multi-sense word embeddings as detectors of word ambiguity.

4.2.5 *Word embeddings as matrix factorization*

The series of papers Levy and Goldberg (2014c), Goldberg and Levy (2014), Levy and Goldberg (2014b), Levy, Goldberg, and Dagan (2015), and Levy et al. (2015) unfolded the series Mikolov, Chen, et al. (2013), Mikolov, Sutskever, et al. (2013), Mikolov, Yih, and Zweig (2013), Mikolov, Le, and Sutskever (2013), and Le and Mikolov (2014) as Zhuangzi unfolded Laozi. As we have already cited, Levy and Goldberg (2014c) showed that skip-gram with negative-sampling (SGNS) is implicitly factorizing a word-context matrix,

$$w \cdot c = \mathrm{PMI}(w, c) - \log k$$

whose cells are the pointwise mutual information (PMI) of the respective word and context pairs, shifted by a global constant. Similarly, an embedding model based on noise-contrastive estimation (Mnih and G. E. Hinton 2008) was shown to be implicitly factorizing a similar matrix, where each cell is the (shifted) log conditional probability of

4 https://github.com/tmikolov/word2vec/

a word given its context. SGNS is much less sensitive to extreme and infinite values than the pure SVD of a PPMI matrix, due to a sigmoid function surrounding $w \cdot c$, and the weighting function: rare (w, c) pairs affect the objective much less.

Levy and Goldberg (2014c) improved results on standard test sets of the time, two word similarity tasks and one of two analogy tasks, with a sparse Shifted PPMI word-context matrix representation of the words. (We introduced PPMI in Section 4.1.2.) They also showed that dense low-dimensional vectors from exact factorization with SVD provides at least as good as SGNS's solutions for word similarity tasks. On analogy questions, SGNS remains superior to SVD. They conjectured that this stems from the weighted nature of SGNS's factorization.

4.2.6 *Global optimization*

The interest in why SGNS can capture such fine-grained semantic and syntactic regularities using vector arithmetic inspired another implementation, *GloVe* (Pennington, Socher, and Manning 2014), which, besides its mathematical elegance, apparently became the most frequently applied word embedding, probably more frequently than the original set by Mikolov et al. Our experiments in Chapters 7 and 8 are no exception. The abbreviation stands for global vectors or, more precisely, globally optimized vectors. The authors claim that models, such as SGNS, that train on separate local context windows instead of on global co-occurrence counts, poorly utilize the statistics of the corpus. The global approach is made possible by training only on the non-zero elements in the word-word co-occurrence matrix.

The basis of GloVe is the logbilinear model

$$w_i^\intercal \widehat{w}_k + b_i + \widehat{b}_k = \log(X_{ik}),$$

where X is the co-occurrence matrix, w and $\widehat{w}$ are the focus and context vectors for each word, and b and $\widehat{b}$ are bias vectors. The two kinds of vectors w and $\widehat{w}$ are needed because words rarely appear in their own context, but we do not want $w^\intercal w$, the squared norm of w, to be small.

The objective above is approximated with weighted least-squares regression, where the weighting is motivated by that rare co-occurrences are noisy and carry less information than the more frequent ones. They introduce the weighting function $f(X_{ij})$, where

$$f(x) = \begin{cases} (x/x_{\max})^\alpha & \text{if } x < x_{\max} \\ 1 & \text{otherwise,} \end{cases}$$

with $x_{\max} = 100$ and $\alpha = 3/4$. Word pairs with a co-occurrence below $x_{\max}$ are downweighted (by a slightly concave function). It is interest-

ing that a similar fractional power scaling was found to give the best performance in Mikolov, Chen, et al. (2013).

Levy, Goldberg, and Dagan (2015) point out that if we were to fix

$$b_w = log\text{freq}(w) \text{ and}$$

$$b_c = log\text{freq}(c),$$

this would be almost equivalent to factorizing the PMI matrix shifted by $log(|D|)$, where $|D|$ is the vocabulary size. However, GloVe learns these parameters, giving an extra degree of freedom over SVD and SGNS. (Unlike Arora et al. (2015)'s RandWalk model, which has a linear relation between the squared norms of the word vectors and the logarithm of the word frequencies.)

Pennington, Socher, and Manning (2014) compare their method to word2vec mathematically and in performance in their sections 3.1 and 4.7, respectively. The quantitative comparison is complicated by many parameters that have a strong effect on performance. They control for the main sources of variation, vector length, context window size, corpus, and vocabulary size. The most important remaining variable to control for is *training time*.

For GloVe, the relevant parameter is the number of training iterations, while for word2vec, the obvious choice would be the number of training epochs, but back then the code was restricted to a single epoch. They measure training time instead by the number of negative samples, which effectively increases the number of training words seen by the model. For the same corpus, vocabulary, window size, and training time, GloVe consistently outperforms word2vec. More interestingly from the big-picture perspective, word2vec's performance decreases if the number of negative samples increases beyond about 10.

4.2.7 *Word analogies, direction, and multiplication*

Levy and Goldberg (2014b) generalize word analogies as searching for a word that maximizes a linear combination of three pairwise word similarities

$$\arg\max_{b^*}(\text{sim}(b^*, b - a + a^*)) = \arg\max_{b^*}(\cos(b^*, b - a + a^*))$$

$$= \arg\max_{b^*}(\cos(b^*, b) - \cos(b^*, a) + \cos(b^*, a^*))$$

(e.g. $b = \text{king}, a = \text{man}, a^* = \text{woman}, b^* = \text{queen}$), and show that the linear representation of lexical properties is not restricted to neural word embeddings: a similar amount of relational similarities can be recovered from traditional distributional word representations. Calling the original additive objective 3CosAdd, they introduce PairDirection, which requires only the direction of $a^* - a$ to be preserved by $b^* - b$, and the multiplicative variant 3CosMul

$$\arg\max_{b^*} \frac{\cos(b^*, b) \cdot \cos(b^*, a^*)}{\cos(b^*, a) + \varepsilon}.$$

$\varepsilon = 0.001$ is used to prevent division by zero. Though it was not mentioned in the paper, Mikolov, Yih, and Zweig (2013) used PairDirection for solving the semantic analogies of the SemEval task, and 3CosAdd for solving the syntactic analogies.

PairDirection performs very well on multiple choice tasks, yet very poorly on full vocabulary searches. The difference is attributed to that PairDirection is likely to find candidates b^* that have the same relation to b as reflected by $a - a^*$ but these candidates are not necessarily similar to b. In the *queen* example, PairDirection may return feminine entities, but not necessarily royal ones. The motivation for 3CosMul is to avoid the "soft-or" behavior of linear objectives, i.e. that they allow one sufficiently large term to dominate the expression.

4.2.8 *Improving PPMI-SVD with neural lessons*

Levy, Goldberg, and Dagan (2015) improve traditional distributional similarity models with lessons learned from word embeddings. We will build on this line of research especially in Chapter 6. Their experiments reveal that much of the performance gains of word embeddings are due to certain system design choices and hyper-parameter optimizations. By making the hyper-parameters explicit, the authors show how they can be adapted and transferred into the traditional count-based approach. Changing the setting of a single hyper-parameter yields more than switching to a better algorithm or training on a larger corpus.

For historical reasons (Baroni, Dinu, and Kruszewski 2014), they refer to PPMI and SVD as "count-based" and to SGNS and GloVe as "neural" or "prediction-based". The following hyper-parameters can be transferred from word2vec and GloVe to count-based methods:

4.2.8.1 *Pre-processing Hyper-parameters*

Words can be weighted according to their distance from the focus word. In traditional count-based methods, it is less common, but also explored (Sahlgren (2006), Section 4.1.4.2). GloVe uses $1, 1/2, 1/3, \ldots$, and word2vec $w/w, w - 1/w, \ldots$. What seem important is the *dynamic context window*: word2vec implements its weighting scheme by uniformly sampling the actual window size between 1 and L.

Subsampling is for diluting very frequent words. Mikolov, Chen, et al. (2013) randomly remove words that are more frequent than some threshold t. While word2vec's code implements a slightly different formula, Levy, Goldberg, and Dagan followed the formula presented in the original paper (equation 2). Subsampling in word2vec is dirty in

the sense that the removal of tokens is done before the corpus is processed into word-context pairs. Levy, Goldberg, and Dagan found the impact of dirty and clean subsampling comparable, and report dirty.

Finally, word2vec removes some rare words before creating context windows, but Levy, Goldberg, and Dagan's experiments showed that the effect of this was small.

4.2.8.2 *Association Metric Hyper-parameters*

The authors define Shifted PMI as

$$SPPMI(w, c) = max(PMI(w, c) - log(\mathrm{k}), \mathrm{o})$$

k has two distinct functions:

- to better estimate the distribution of negative examples: a higher k means more data and better estimation, and
- it affects the probability of observing a positive example: a higher k means that negative examples are more probable.

Shifted PPMI captures only the second aspect of k. They experiment with three values of k: 1, 5, 15.

Finally, in word2vec, negative examples (negative contexts, Section 4.2.3) are sampled according to a *smoothed* unigram distribution. Smoothing alleviates PMI's bias towards rare words.

4.2.8.3 *Post-processing Hyper-parameters*

When word vectors are used in some downstream task (an intrinsic test or a real application), *context vectors* c are often *added* to focus vectors w. This was originally motivated as an ensemble method. While this addition does not apply to PPMI, it is interesting that the authors provide a different interpretation of its effect: it adds first-order similarity terms to the second-order similarity function. Second-order similarity $w_x \cdot w_y, c_x \cdot c_y$ measures the extent to which the two words are replaceable based on their tendencies to appear in similar contexts, and are the manifestation of Z. S. Harris (1954)'s distributional hypothesis. First-order similarity $w_x \cdot c_y$, on the other hand, is the tendency of one word to appear in the context of the other.

Recall that truncated Singular Value Decomposition (SVD) is a common method of dimensionality reduction, which finds the optimal rank d factorization with respect to L_2 loss. SVD was popularized in NLP via Latent Semantic Analysis (LSA, Deerwester, Dumais, and Harshman (1990), Section 4.1.3). The word-context matrix M is factorized as

$$M = U \cdot \Sigma \cdot V$$

where U and V are orthonormal and Σ is a diagonal matrix of eigenvalues. The representations are obtained as $W_{SVD} = U_d \cdot \Sigma_d$ for words and $C_{SVD} = V_d$ for contexts.

In the SVD-based factorization, the context matrix C_{SVD} is orthonormal while the word matrix W_{SVD} is not. The factorization by SGNS's is much more "symmetric": neither W_{w2v} nor C_{w2v} is orthonormal, and there is no bias to either of the matrices in the training objective. Symmetry can be achieved in SVD by *weighting the eigenvalue matrix* Σ_d with the exponent p, what has a significant effect on performance, and should be tuned. The final hyper-parameter of any vector space language model is whether rows and/or columns are normalized.

4.2.8.4 *Low-dimensional embeddings and isotropy*

Arora et al. (2016) emphasize that $\langle v_w, v_{w'} \rangle \approx PMI(w, w')$ was only true if there were no dimension constraints, but, in practice, low-dimensional embedding are used. They argue that the low dimensionality of word embeddings plays a key role. In previous papers, the model is agnostic about the dimension of the embeddings, and the superiority of low-dimensional embeddings is an empirical finding (starting with Deerwester, Dumais, and Harshman (1990)). Arora et al.'s theoretical analysis makes the key assumption that the set of all word vectors (which are latent variables of the generative model) are spatially isotropic, i.e. they have no preferred direction in space. Having n vectors be isotropic in d dimensions requires $d \ll n$. This is related to the emergence of the "relations = lines" phenomenon (Section 4.2.4).

4.2.9 *What's in a similarity score?*

The basic evaluation for static word embeddings has been word similarity, but the method has many shortcomings. Now we summarize Avraham and Goldberg (2016) to illustrate these. Avraham and Goldberg redesign the annotation task to achieve higher inter-rater agreement, and propose a performance measure which takes the reliability of each annotation decision in the dataset into account.

Datasets for Word Similarity Evaluation have been standardly used with rank correlation (Spearman's ρ). Hill, Reichart, and Korhonen (2015) pointed out that in some datasets, associated but dissimilar words, e.g. $\langle$*singer*, *microphone*$\rangle$, ranked high, sometimes even above pairs of similar words. Hill, Reichart, and Korhonen also found a clear preference for hyponym-hypernym pairs, e.g. $\langle$*cat*, *pet*$\rangle$ and $\langle$*winter*, *season*$\rangle$ over cohyponyms pairs like $\langle$*cat*, *dog*$\rangle$ (and, less outrageously, over antonyms pairs $\langle$*winter*, *summer*$\rangle$).

Avraham and Goldberg summarize the problems as follows:

- The rating scales are vulnerable to a variety of biases. This problem was earlier addressed by asking the annotators to rank each

pair in comparison to 50 randomly selected pairs, but that resulted in a daunting annotation task.

- Different relations are rated on the same scale. A difference of 1.8 similarity scores can testify to anything from no difference, e.g.

  $$\text{sim}(\text{smart, dumb}) = 0.55, \text{sim}(\text{winter, summer}) = 2.38,$$

  to true superiority of one pair, e.g.

  $$\text{sim}(\text{cab, taxi}) = 9.2, \text{sim}(\text{cab, car}) = 7.42..$$

- Different target words are rated on the same scale. Even within pairs in a targeted relation, there are ill-defined comparisons, e.g.: $\langle$*cat*, *pet*$\rangle$ vs. $\langle$*winter*, *season*$\rangle$. Pairs which share the target are much more natural to compare, e.g. the comparison $\langle$*cat*, *pet*$\rangle$ vs. $\langle$*cat*, *animal*$\rangle$ is natural. Penalizing a model for preferring $\langle$*cat*, *pet*$\rangle$ over $\langle$*winter*, *season*$\rangle$ or vice versa impairs the evaluation reliability.

- The evaluation measure does not consider the annotation decisions' reliability. Reliability should be determined by the agreement of the annotators.

They publish two datasets of Hebrew nouns with the following features:

- The annotation task is an explicit ranking task: each pair is directly compared with a subset of the other pairs, but, unlike in earlier work, with only a few carefully selected pairs, following the principles above.

- Only pairs in a single preferred relation type (hyponym-hypernym in one dataset, and cohyponym in the other one) are presented to the annotators, what spares the annotators the effort of considering the type of the similarity, and lets them concentrate on the strength of the similarity.

- Any pair is compared only with pairs sharing the same target word.

- The dataset includes a reliability indicator with a probabilistic interpretation.

4.2.10 *Retrofitting vectors to semantic lexicons*

The two main topics of this thesis are semantic networks (relational representations of lexical meaning) and neural word embeddings. The original goal of both was to model associations in the human mind that

make linguistic processing possible. Early research in computational linguistics was based on manual implementation of expert knowledge, and hand-crafted tools remain useful even today. Since the nineties, computers have become able to learn from text corpora of increasing size, and in recent years, artificial neural networks became the state of the art in many computational applications, but their interpretability remains poor. In this section, we investigate methods of injecting knowledge from semantic networks to (static) word embeddings.

Works before Faruqui et al. (2015) either augmented the co-occurrence matrix in a relation-specific way, or changed the objective of the word vector training algorithm to include some relational knowledge. The latter involves enhancing word2vec to include more similarity knowledge or word relational knowledge and or latent semantic analysis for antonym specific polarity induction or multi-relational knowledge. These methods are limited to particular vector models. Faruqui et al. introduced a graph-based learning technique. The training objective includes an additional term for new vectors to be similar to the vectors of related word types. Relations are taken from semantic lexicons such as WordNet (Section 2.4.3), FrameNet (Section 2.4.4), and the Paraphrase Database.

Besides the English GloVe (Section 4.2.6), skip-gram with hierarchical softmax (Section 4.2.4), and the multi-prototype model of Huang et al. (2012, Section 8.3), the experiments involve Multilingual Vectors by Faruqui and Dyer (2014), who learned vectors by first performing SVD on text in different languages, then applying canonical correlation analysis on pairs of vectors for words that align in parallel corpora. These vectors were trained on the WMT-2011 news corpus for English, French, German and Spanish.

The resulting representations were evaluated for their semantic and syntactic aspects in extrinsic sentiment analysis task, Word Similarity, Syntactic Relations by Mikolov, Synonym Selection (TOEFL), and phrase and sentence level Sentiment Analysis (Socher et al. 2013).

Mrkšić et al. (2016) present a counter-fitting method that injects both antonymy and synonymy constraints into vector space representations improving the vectors' capability for judging semantic similarity. The method gave new a state of the art performance on the SimLex-999 dataset and was demonstrated in the downstream task of dialogue state tracking (where the task is updating the system's distribution over user goals as the conversation progresses and new information becomes available), resulting in robust improvements across domains.

Word representations coalesce semantic similarity and conceptual association (Hill, Reichart, and Korhonen 2015). Furthermore, even methods that can distinguish similarity from association (e.g., based on syntactic co-occurrences) will generally fail to tell synonyms from antonyms (Mohammad, Dorr, and Hirst 2008). Distinguishing antonymy from similarity is critical for the dialogue state tracking (DST) task, more specifically the restaurant domain, where systems should not

recommend an "expensive pub in the south" when asked for a "cheap bar in the east". Counter-fitting, is a lightweight post-processing procedure in the spirit of the retrofitting introduced in the previous subsubsection.

Mrkšić et al. (2017) introduce Attract-Repel which jointly injects mono- and cross-lingual synonymy and antonymy in word embeddings, yielding semantically specialised[5] cross-lingual vector spaces. In practice, semantic transfer goes from high to lower-resource languages. Their evaluation obtains SOTA on SimLex semantic similarity datasets in six languages and in DST across multiple languages. Their multilingual DST models bring further performance improvements.

Mrkšić et al. call the retrofitting approach, i.e. when vectors are refined to satisfy constraints extracted from a lexicons such as WordNet, *semantic specialization*. Mrkšić et al. deploy the Attract-Repel algorithm in a multilingual setting, taking semantic relations from BabelNet and exploiting information from high-resource languages to improve the lower-resourced ones. They train their cross-lingual vector spaces jointly, which brings benefits in the form of positive semantic transfer.

Mrkšić et al. demonstrate their efficacy both in intrinsic and downstream tasks. The former includes SOTA results on the four languages in the Multilingual SimLex-999 dataset and in lower-resource languages Hebrew and Croatian, where Mrkšić et al. collect evaluation datasets, and show that cross-lingual specialization significantly improves word vector quality.

Their downstream applications are motivated by improving the lexical coverage of supervised models. Mrkšić et al. consider again DST. Incorporating their specialised vectors into a SOTA neural network model for DST improves performance on English dialogues. In a multilingual spirit, Mrkšić et al. produce new Italian and German DST datasets, where Attract-Repel-specialised vectors leads to even stronger gains, and they train a single model that performs DST in all three languages, in each case outperforming the monolingual model.

The retrofitting models discussed so far specialize only the vectors of words from the constraints. Glavaš and Vulić (2018) use the external lexico-semantic relations to train an explicit retrofitting model (ExRf), a deep feedforward neural architecture, which learns a global specialization function and specializes the vectors of words unobserved in the whole training data. ExRf is applicable to arbitrary embeddings. The authors also specialize vector spaces of new languages (i.e. unseen in the training) by coupling ExRf with shared multilingual distributional vector spaces. Glavaš and Vulić's proposal unifies the two prominent ways for external constraints: joint specialization models, which integrate the constraints into the distributional learning objective, and post-processing models, which fine-tune distributional vectors retroactively.

5 They use British spelling, and we keep it, because this is a term.

In general, the latter outperform the former, and they can be applied to arbitrary distributional spaces but vectors of all unseen words remain intact. Their evaluate the model in intrinsic word similarity evaluation (on the standard benchmarks SimLex-999 (Hill, Reichart, and Korhonen 2015) and SimVerb-3500 (Gerz et al. 2016)) and two downstream tasks – lexical simplification and dialog state tracking.

4.2.11 *Sub-word embeddings for rich morphology*

The next important step in the history of word embeddings is sub-word level modeling, which we now discuss with an emphasis on rich morphology, keeping in mind that sub-word level modeling solves other kinds of out-of-vocabulary problems, like proper nouns, as well.

As we will see for the case of Hungarian in Section 7.4, for morphologically rich languages, word embeddings provide less consistent semantic representations due to higher variance in word forms and often less constrained word order, which further increases variance. In this section, we focus on two families of solutions proposed.

4.2.11.1 *Gluten-free word embeddings*

In Nemeskey (2017)'s retrospection, „The most common solution in the literature is to break up the words into smaller segments (Hirsimäki et al. 2005; Afify et al. 2006; Botha and Blunsom 2014)." For Hungarian, the idea was introduced in the context of statistical machine translation (László Tihanyi, personal communication). More specifically, in the context of Hungarian static word embeddings, it was proposed by Siklósi and Novák (2016) and Novák and Novák (2018a).[6] Following Borbély, Kornai, et al. (2016) and Nemeskey (2017), we will call the method *deglutination*, and the models *deglutinated, deglutinized*, or simply *gluten-free*. The name refers to that languages like Finnish, Hungarian, or Turkish are called agglutinative, because they mark grammatical (syntactic) relations by gluing inflectional suffixes to the words. In the gluten-free method, we split compositional derivational and inflectional suffixes from the stem. The suffixes are represented by their morphological analysis (i.e. different allomorphs of the same morpheme are represented by the same symbol).

On the practical level, deglutination uses a classical NLP pipeline with a rule-based morphological tagger, which lists all the possible morphological analyses of each word in a linguistically principled formalism, and a POS disambiguator (tagger), which selects the analysis which is relevant in the context. This thesis uses gluten-free embeddings in two experiments. One of our cross-lingual word-sense inductions experiments (Section 8.5.5), which is based on Makrai and Lipp (2018), uses the general purpose morphological annotation by

6 The former is in Hungarian.

jelmondatával	→	<jelmondat> <poss> <casins>
akartak	→	<akar> <past> <plur>
érdekelheti	→	érdekel [/V] [_Mod/V] [Prs.Def.3Sg]
köszönhetően	→	köszönhető [/Adj] [_Manner/Adv]

Table 7: Deglutination in the Kornai–Rebrus annotation (top pane, example by Nemeskey (2017)) and with emMorph (bottom pane).

Rebrus, Kornai, and Varga (2012) and the `emLam` free-access gluten-free corpus (Nemeskey 2017), while in our experiments in the context of analogical questions (Section 7.4), which were conducted directly for the purposes of the present thesis, we used Webcorpus 2.0 (Nemeskey 2020), which is tagged with the emMorph morphological analyzer (Novák 2014; Novák, Siklósi, and Oravecz 2016). (A detailed discussion of the `emMorph` formalism is given in Hungarian by Novák, Rebrus, and Ludányi (2017)). Ideally, non-compositional derivational suffixes like *-hető* 'able' in *köszönhető . . . -nak/-nek* 'be due to . . . ', lit. *thank-able to* (bottom pane in Table 7, PAT is thankable to AGT, i.e. we may thank AGT for PAT) should remain on the stem, while compositional derivational and inflectional suffixes like the possessive or *-het* 'can' in *érdekelheti* '(he/she) *can* be interested in (it)' should be put in the suffix series.

4.2.11.2 *Character n-grams, lemmatization, and stemming*

In Döbrössy et al. 2019[7], we explored and evaluated several simple subword unit based embedding strategies – character n-grams, lemmatization provided by an NLP-pipeline, and segments obtained in unsupervised learning (Morfessor) – to boost the semantic consistency of Hungarian word vectors in the analogical benchmark that will be introduced in Section 7.4.1 in this thesis. The effect of changing embedding dimension and context window size is also considered. Morphological analysis based lemmatization is found to be the best strategy to improve embeddings' semantic accuracy, whereas representation by character n-grams is consistently counterproductive in this regard (Figures 9 to 12).

4.2.12 *The offset is naked*

The basic way of evaluating static word embeddings has been intrinsic evaluations, namely similarities and analogies. Both methods have serious shortcomings – we illustrated this for similarities in Section 4.2.9. Now we turn to a critical reflection on what have been called the vector offset method, relational similarity, or word analogies.

7 The paper was based on the BSc thesis of Bálint Döbrössy, advised by György Szaszák, and refereed by Makrai.

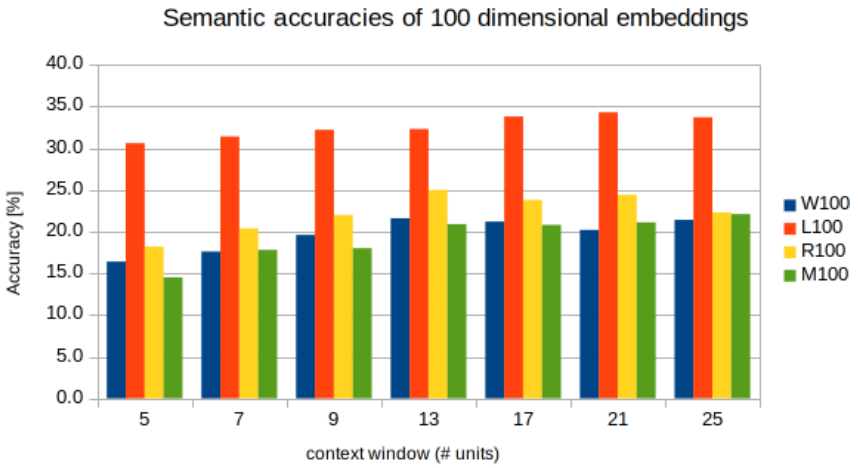

Figure 9: Semantic accuracy of Hungarian 100 dimensional embeddings as obtained by Döbrössy et al. (2019) with different preprocessings: word forms with no preprocessings (W) as the baseline, lemmas (L) obtained with the magyarlánc toolkit (Zsibrita, Vincze, and Farkas 2013), and two strategies based on Morfessor (Virpioja et al. 2013): taking all morfs (M) or only the root (R) of each word. During testing in analogical questions, query words are also spitted to segments, and their vectors are computed as the sum of the segments' vectors.

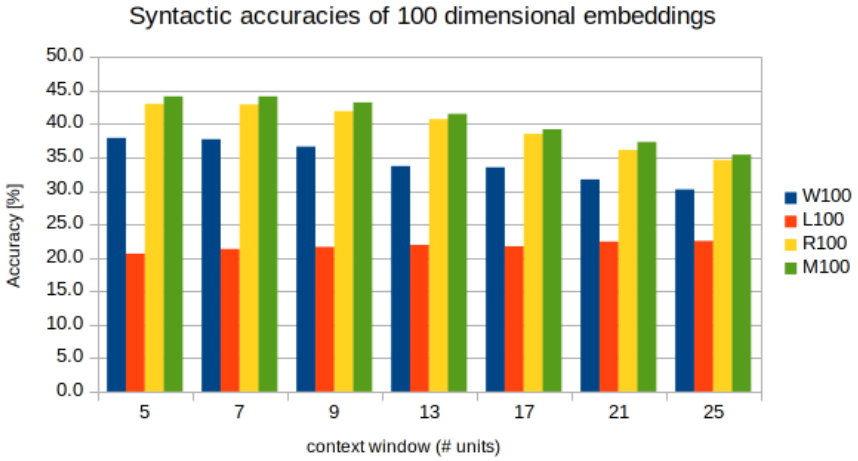

Figure 10: Syntactic accuracy of Hungarian 100 dimensional embeddings with different strategies.

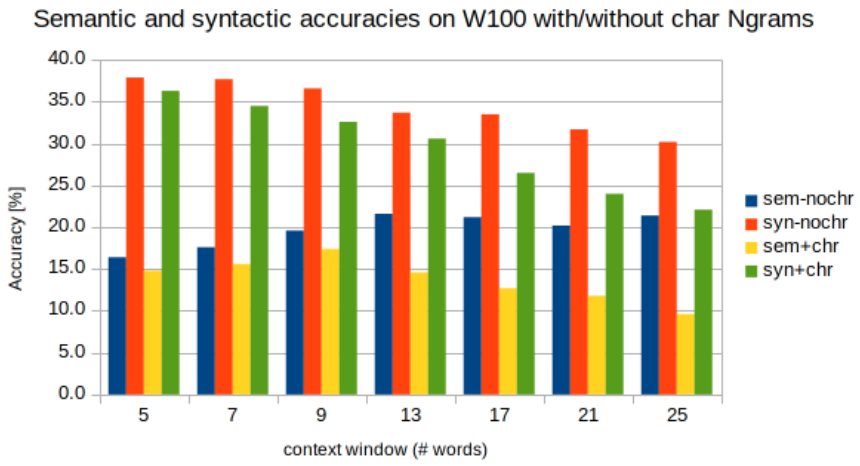

Figure 11: Semantic and syntactic accuracy of Hungarian 100 dimensional word embeddings with character n-grams (chr) and in the original way (nochr).

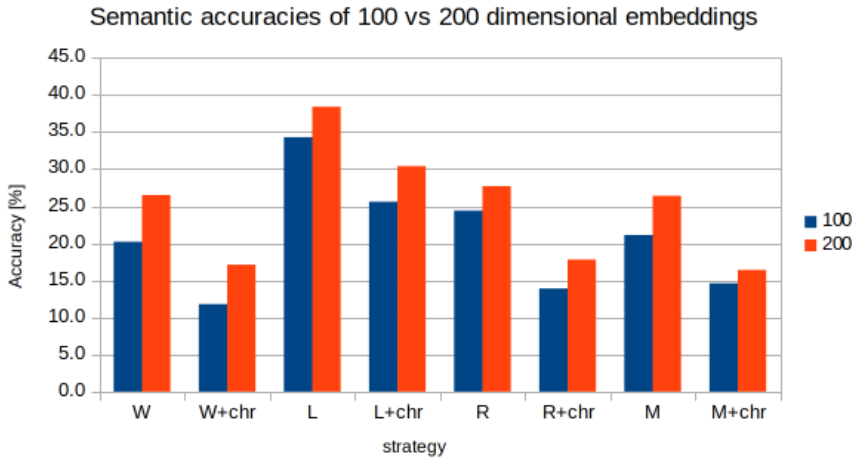

Figure 12: Semantic accuracy of Hungarian 100 and 200 dimensional embeddings with different strategies; context window covers 21 units.

Levy et al. (2015) argue that supervised methods for hypernymy extraction are actually memorizing whether the hypernym candidate is a "prototypical hypernym", i.e. a category, irrespective of the word to be categorized. They compare four compositions for representing (x, y) (e.g. x = cat, y = animal) as a feature vector: besides the standard concatenation $x \oplus y$ and difference $y - x$, they use the diagnostic "only x", and "only y". The finding is that models just learn whether y is a likely "category" word – a prototypical hypernym – and, to a lesser extent, whether x is a likely "instance" word. This extends to other inference relations, such as meronymy. To test the hypothesis, the authors manipulate the test pairs by inserting mismatched pairs, e.g. *(banana, animal)*.

The word embeddings they use include interpretable PPMI-based ones, which enable them to look for prototypical hypernym contexts. Besides dataset-specific contexts like *psychosomatic* `-1` (*word* $\pm$ `i` denotes the context where the ith word to to right/left is *word*), they find domain-independent indicators of category, e.g. *any* `-1`, *every* `-1`, and *kinds* `-2`, and even relics of the Hearst patterns in all datasets: *other* `-1`, *such* `+1`, *including* `+1`, etc., and their analogons, e.g. *such* `-2`.

Linzen (2016) notes that in analogical tasks

$$x = a^* - a + b,$$

if a^* and a are very similar to each other (as *scream* and *screaming* are likely to be) the nearest word to x may simply be the nearest neighbor of b. If in a given set of analogies the nearest neighbor of b tends to be b^*, the answer may be correct regardless of the consistency of the offsets. He proposes new baselines that perform the task without using the offset $a^* - a$, and measures how the performance is affected by *reversing* the direction of each analogy problem (which should not affect its accuracy).

4.2.13 *Theoretical critique of vector analogy*

Rogers, Drozd, and Li (2017) criticize the vector analogy method on theoretical grounds. Given the vital role that analogical reasoning plays *in human cognition*, automated analogical reasoning could become a game-changer in many fields. The method is already used in many downstream NLP tasks, such as splitting compounds, semantic search, and cross-language relational search. One way to explain the current limitations is to attribute them to the imperfections of the models and/or the corpora. With this view, in a perfect VSM, any linguistic relation should work. The alternative explored by Rogers, Drozd, and Li is that there are both theoretical and mathematical issues with analogical reasoning with word vectors and 3CosAdd (see Section 4.2.7).

In the authors' view, the most fundamental term is not analogy, but *relational similarity*, i.e. that pairs of words may hold similar relations.

We speak of similarity rather than identity: instances of a single relation may still have significant variability in how characteristic they are of that class.

"Classical" analogical reasoning follows roughly this template: objects X and Y share properties a, b, and c; therefore, they may also share the property d. For example, both Earth and Mars orbit the Sun, have at least one moon, revolve on axis, and are subject to gravity; therefore, if Earth supports life, so could Mars. The NLP move from relational similarity to analogy follows the use of the term by Turney (2006).

Analogy was once rejected in generative linguistics as a mechanism for language acquisition through discovery, although now it is making a comeback. It has been criticized for ambiguity, guesswork and puzzle-like nature.

The paper has been referred to as *Mikolov cheated!*, because they point out that 3CosAdd, as initially formulated by Mikolov, Yih, and Zweig (2013), "dishonestly" excludes a, a_* and b from among potential $b*$s.

The authors present a series of experiments performed with the BATS dataset, which has more relations and is more difficult than the original Google test. BATS is balanced across derivational and inflectional morphology, lexicographic and encyclopedic semantics (10 relations of each type). They explain lower performance on *derivational morphology questions* as opposed to inflectional or encyclopedic semantics: *man* and *woman* are reasonably similar distributionally, as they combine with many of the same verbs: both men and women sit and sleep, but the same could not be said of words derived with prefixes that change POS.

Another, purely logical problem is exemplified by *snow: white :: sugar: ?white*, where, in the dishonest setting, the correct answer is a priori excluded. In BATS data, this factor affects several semantic categories, including country:language, thing:color, animal:young, and animal:shelter.

Rogers, Drozd, and Li hypothesize that the more *crowded a particular region* is, the more difficult it should be to hit a particular target. Estimating density as the similarity to the 5th neighbor, they get the counter-intuitive results that denser neighborhoods actually yield higher scores.

They consider LRCos, a method based on *supervised* learning from a set of word pairs. The model learns a representation of the target class with a supervised classifier. The question is this: what word is the closest to king, but belongs to the "women" class? The accuracy of LRCos is much higher than the top-1 3CosAdd or 3CosMul, and its "honest" version performs just as well as the "dishonest" one.

4.2.14 *Frequency effects in cosine similarity*

Faruqui et al. (2016) review the main problems with word similarity evaluations, and they discuss frequency effects in cosine similarity (besides the subjectivity of the task; the confusion of semantic and task-specific similarity; the lack of standardized splits and overfitting; the low correlation with extrinsic evaluation, e.g. that in text classification, parsing, or sentiment analysis; and the absence of statistical significance).

Vectors of frequent words are longer as they are updated more often during training (Turian, Ratinov, and Bengio 2010). In Faruqui et al.'s view, ideally the relatively small number of frequent words should be evenly distributed through the space, while rare words should cluster around related, but more frequent words.

However, vector-spaces contain hubs, i.e. vectors that are close to a large number of other vectors in the space. In word vector-spaces, this manifests in words that have high cosine similarity with a large number of other words (Dinu, Lazaridou, and Baroni 2015), as we will discuss in Sections 7.5.3 and 8.4.1. Schnabel et al. (2015) further refine this hubness problem to show a power-law relationship between the frequency-rank r of a word (i.e. the rank of a word in vocabulary of the corpus sorted in decreasing order of frequency) and the frequency-rank of its neighbors: the average rank a of the 1000 nearest neighbors of a word follows: $a \approx 1000r^{0.17}$.

The last problem Faruqui et al. discuss is related to the main problem with word embeddings of the type investigated in this section: the inability to account for polysemy. As we will see in Chapter 8, there has been progress on obtaining multiple vectors per word-type to account for different word-senses, but the practical advantage of word embeddings with more but fixed vectors to account for different senses remained modest (Li and Jurafsky 2015), and in most applications, the real solution is contextualized word representations provided by deep language models, which brought a new paradigm in NLP, to which will now turn.

4.3 ATTENTION AND DEEP LANGUAGE MODELS

The contributions of this thesis are based on static word embeddings, i.e. the kind discussed so far, but we would like to put our investigation in the context of the advances of the past few years. Deep neural networks defined a new state-of-the-art in many areas of NLP.

Deep neural networks and *deep learning* mean machine learning of a model that consists of layers from the input layer through hidden layers to the output layer, and calculates higher and higher level features. Deep learning brought its first breakthroughs in speech technology (Dahl et al. 2011) and computer vision (Krizhevsky and Sutskever

2012). The ImageNet moment of NLP, as Ruder (2018) called it, arrived in 2018.

> Pretraining entire models to learn both low and high level features has been practiced for years by the computer vision (CV) community. Most often, this is done by learning to classify images on the large ImageNet dataset. ULMFiT, ELMo, and the OpenAI transformer have now brought the NLP community close to having an "ImageNet for language" – that is, a task that enables models to learn higher-level nuances of language, similarly to how ImageNet has enabled training of CV models that learn general-purpose features of images. (Ruder 2018)

Up to this point of the thesis, we were chronological and didactic. Main contribution chapters will be similar, even self-contained in many cases. This section provides, however, just some flashes for the reader somewhat familiar with deep learning of language. Those with less background in machine learning may skip to the foreground part, to Chapter 5. Where citations are omitted, they can be found in the corresponding paper we just summarize.

4.3.1 *Pre-trained deep models for NLP*

Qiu et al. (2020, Section 2.4.2) summarize the history of pre-trained deep NLP models as follows: McCann et al. (2017) pre-trained a deep *LSTM encoder from an attentional sequence-to-sequence model* with machine translation objective, and used the context vectors (CoVe) output by the pre-trained encoder. Peters et al. (2018) pre-trained a 2-layer LSTM encoder with a *bidirectional language model (BiLM)*, consisting of a forward LM and a backward LM. Contextual representations output by the pre-trained BiLM, ELMo (Embeddings from Language Models) brought large improvements on a broad range of tasks. Flair (Akbik, Blythe, and Vollgraf 2018) captured word meaning with contextual string embeddings pre-trained with a *character-level* LM. Ramachandran, Liu, and Le (2017) significantly improved the seq2seq models by unsupervised pre-training. The weights of both the encoder and the decoder are initialized with pre-trained weights of two language models and then fine-tuned with labeled data.

ULMFiT (Universal Language Model Finetuning, Howard and Ruder (2018)) fine-tuned a *pre-trained LM* for text classification, achieving state-of-the-art results on six widely-used text classification datasets. ULMFiT training consists of three phases: pre-training LM on general-domain data; fine-tuning LM on target data; and fine-tuning on the target task. Their *fine-tuning* strategies include discriminative fine-tuning, slanted triangular learning rates, and gradual unfreezing. Since ULM-

FiT, fine-tuning has become the mainstream approach to adapt PTMs for the downstream tasks.

Very deep PTMs have shown their powerful ability in learning universal representations, including OpenAI GPT (Generative Pre-trained Transformer Radford et al. (2018)) and BERT (Bidirectional Encoder Representation from Transformer, Devlin et al. (2018)). Besides LM, an increasing number of self-supervised tasks are proposed to make the PTMs capturing more knowledge form large scale text.

4.3.2 *BERTology*

Transformer-based models are now widely used in NLP, and much work has been done to understand their inner workings. The stream of papers seems to be accelerating rather than slowing down. Here we summarize the findings of Rogers, Kovaleva, and Rumshisky (2020), who synthesize over 40 analysis studies, overview the proposed modifications and the training regime, and offer directions for further research.

4.3.2.1 *Introduction*

Transformers (Vaswani et al. 2017) took NLP by storm, offering enhanced parallelization and better modeling of long-range dependencies. The most popular model is BERT (Devlin et al. 2019), which obtained state-of-the-art results in many benchmarks, and it has been integrated in Google search, improving an estimated 10% of queries. However, this family of models has little cognitive motivation, and the size of these models limits their training and study. Rogers, Kovaleva, and Rumshisky focus on the papers investigating the types of knowledge learned by BERT, where this knowledge is represented, how it is learned, and the methods proposed to improve it.

4.3.2.2 *Overview of BERT architecture*

BERT is a stack of Transformer encoder layers with multiple *heads*, i.e. fully-connected neural networks augmented with a self-attention mechanism. For every input token in a sequence, each head computes key, value and query vectors, which are used to create a weighted representation. The outputs of all heads in the same layer are combined and run through a fully-connected layer. Each layer is wrapped with a skip connection and layer normalization

The conventional workflow is pre-training and fine-tuning. Pre-training uses two semi-supervised tasks: masked language modeling (MLM, prediction of randomly masked input tokens), and next sentence prediction (NSP, predicting if two input sentences are adjacent to each other). In fine-tuning for downstream applications, one or more fully-connected layers are typically added on top of the final encoder layer.

The representations are computed as follows: the model tokenizes the given word into wordpieces, and then combines three embedding layers (token, position, and segment). The special token [CLS] is used for classification predictions, and [SEP] separates segments of typically multi-sentence input. Two sizes fit all: base and large, varying in the number of layers, their hidden size, and number of attention heads.

4.3.2.3 *What knowledge does BERT have?*

Analysis approaches include fill-in-the-gap probes of BERT's MLM, that of self-attention weights, and probing classifiers using different BERT representations as inputs.

SYNTACTIC KNOWLEDGE Representations are hierarchical rather than linear. There is something akin to syntactic tree structure in addition to the word order information. BERT has information about parts of speech, syntactic chunks and roles. Knowledge of syntax is partial, not enough to recover the labels of distant parent nodes in the syntactic tree. The syntactic structure is not directly encoded in self-attention weights, but they can be transformed to reflect it. Dependency trees have been extracted directly from self-attention weights but without quantitative evaluation. Transformation matrices recover much of the Stanford Dependencies formalism for PennTreebank data.

BERT representations have been approximated with Tensor Product Decomposition Networks, concluding that dependency trees are the best match among five decomposition schemes, but differences are very small. BERT takes subject-predicate agreement into account in the cloze task even with distractor clauses and meaningless sentences. BERT is able to detect the presence of negative polarity items (e.g. "ever") and the words that allow their use (e.g. "whether") but not scope violations. BERT does not understand negation, and it is insensitive to malformed input: predictions were not altered even with shuffled word order, truncated sentences, or removed subjects and objects. Models are disturbed by nonsensical input (adversarial attacks).

SEMANTIC KNOWLEDGE Fewer studies were devoted to BERT's knowledge of semantics. Entity types, relations, semantic roles, and proto-roles have been detected with probing classifiers. BERT has some knowledge for *semantic roles*. We have seen in Section 4.3.5 that Ettinger (2020) shows with an MLM probing study that the model prefers incorrect fillers for semantic roles that are semantically related to the correct ones to those that are unrelated, e.g. *to tip a chef* to *to tip a robin*.

BERT struggles with representations of *numbers* (addition, number decoding, floating point numbers). The problem may be with wordpiece tokenization: numbers of similar values can be divided up into substantially different word chunks.

BERT is surprisingly brittle to *named entity* replacements: replacing names in the coreference task changes 85% of predictions. This suggests that the model does not form a generic idea of named entities, although its F1 scores on NER probing tasks are high. Fine-tuning BERT on Wikipedia entity linking "teaches" it additional entity knowledge, which suggests that it did not absorb all the relevant entity information during pre-training on Wikipedia.

WORLD KNOWLEDGE MLM has been adapted for knowledge induction by filling in the blanks, e.g. "Cats like to chase [ ]". Besides a probing study of world knowledge in BERT, evidence comes from many practitioners using BERT to extract knowledge. For some relation types, vanilla BERT is competitive with knowledge base methods. BERT generalizes well to unseen data, but we need good template sentences. There has been research on the automatic extraction and augmentation of such templates.

BERT cannot reason based on its world knowledge. It can *guess* the affordances and properties of many objects, but it has no information about their interactions. E.g. it knows that people can walk into houses, and that houses are big, but it cannot infer that houses are bigger than people. Its performance drops with the number of necessary inference steps. Some of BERT's success in factoid knowledge retrieval comes from learning stereotypical character combinations, e.g. that a person with an Italian-sounding name is Italian.

LIMITATIONS Some researchers remark that "the fact that a linguistic pattern is not observed by our probing classifier does not guarantee that it is not there, and the observation of a pattern does not tell us how it is used." A hot question is how complex a probe should be: If a more complex probe recovers more information, to what extent are we still relying on the original model? Different probing methods may lead to complementary or even contradictory results. A given method might also favor one model over another. E.g., RoBERTa trails BERT with one tree extraction method, but leads with another. The choice of linguistic formalism also matters.

We should focus on identifying what BERT actually relies on at inference time. Amnesic probing aims to specifically remove certain information, and see how it changes performance. This method has shown that e.g. language modeling does rely on part-of-speech information.

Information-theoretic probing approaches include estimating the mutual information between the learned representation and a given linguistic property. Some researchers quantify the amount of effort needed to extract some information, which is more important than the amount of information in the representation. The mathematical formalism is minimum description length needed to communicate both the probe size and the amount of data required for it to do well on a task.

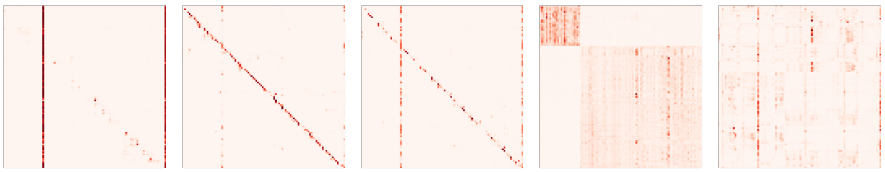

Figure 13: Typical self-attention patterns (Kovaleva et al. 2019). Both axes on every image represent BERT tokens of an input example, and colors denote absolute attention weights (darker colors stand for greater weights). The first three types are most likely associated with language model pre-training, while the last two potentially encode semantic and syntactic information.

4.3.2.4 *Localizing linguistic knowledge*

BERT EMBEDDINGS In studies of BERT, the term *embedding* refers to the output of a Transformer layer (typically, the final one). Every token contains at least some information about the *context*. Both conventional static embeddings and BERT-style embeddings can be viewed in terms of mutual information maximization.

Distilled contextualized embeddings better encode lexical semantic information, i.e. they are better at traditional word-level tasks such as word similarity. The methods to distill a contextualized representation into a static one include aggregating the information across multiple contexts, encoding "semantically bleached" sentences that rely almost exclusively on the meaning of a given word (e.g. *This is* $\langle\ \rangle$), and using contextualized embeddings to train static embeddings. Distillation to a static embedding is useful because interpretability methods for static embeddings are more diverse and mature than those available for their dynamic counterparts.

It has been studied how similar the embeddings for identical words are in every layer, reporting that later BERT layers are more context-specific. In the earlier Transformer layers, MLM forces the acquisition of contextual information at the expense of the token identity, which gets recreated in later layers. To what extent do models capture phenomena like polysemy and homonymy? BERT embeddings form distinct clusters corresponding to word senses. The model is successful at word sense disambiguation. Representations of the same word depend on the position of the sentence in which it occurs, likely due to the NSP objective, what is desirable from the linguistic point of view, and could be a promising avenue for future work.

The standard way to generate *sentence* or *text* representations for classification is to use the [CLS] token, the concatenation of token representations, or the normalized mean.

SELF-ATTENTION HEADS Several classifications of attention heads have been proposed in different studies:

- attending to the word itself, to previous/next words and to the end of the sentence,
- attending to previous/next tokens, to the [CLS], to the [SEP], to punctuation, or broadly over the sequence, or
- the five attention types in Figure 13: Vertical, Diagonal, Vertical + diagonal, Block, and Heterogeneous.

Heads with linguistic functions. The "heterogeneous" attention pattern could be linguistically interpretable, and a number of studies focused on identifying the functions of the heads.

There are BERT heads that attended significantly more than a random baseline to words in certain *syntactic* positions. Datasets and methods used in these studies differ, but there is some consistency that some heads attend to words in `obj` role more than the positional baseline. Evidence for `nsubj`, `advmod`, and `amod` varies between studies. The overall conclusion is also supported by a study in machine translation context. Even complex dependencies like `dobj` may be encoded by a combination of heads, but the corresponding work is limited to qualitative analysis.

No single head has the complete syntactic tree information, but a BERT head can directly be used for coreference classification on par with a rule-based system, what is remarkable because coreference classification requires quite a lot of syntactic knowledge. Attention weights are weak indicators of subject-verb agreement and reflexive anaphora. Instead of serving as strong pointers between related tokens, they were close to a uniform attention baseline, but there was some sensitivity to different types of distractors coherent with psycholinguistic data we saw in Section 4.3.5. Morphological information in BERT heads has not been addressed, but with a sparse attention variant in the base Transformer, some attention heads appear to merge BPE-tokenized words.

Semantic relations, core frame-semantic relations, as well as lexicographic and commonsense relations have been studied, but a head ablation study showed that heads related to some of these problems were not essential for BERT's success on GLUE tasks.

The popularity of self-attention as interpretation is due to the idea that "attention weights have a clear meaning: how much a particular word will be weighted when computing the next representation for the current word." This has been much debated. In a multi-layer model where attention is followed by a non-linear transformation, the patterns in individual heads do not provide a full picture. Many current papers are accompanied by attention visualizations, and visualization tools, but analysis is mostly qualitative, often with cherry-picked examples, and should not be interpreted as evidence.

Attention to special tokens. Most self-attention heads do not directly encode any nontrivial linguistic knowledge; at least after fine-tuning on GLUE, less than 50% of heads exhibit the "heterogeneous" pattern. Much of the heads have the vertical pattern (attnding to [CLS],

[SEP], and punctuation), what is likely related to the overparametrization issue. Norms of attention-weighted input vectors yield a more intuitive interpretation of self-attention reducing the attention to special tokens, but it is still not the case that most heads that do the "heavy lifting" are even potentially interpretable. Some work focuses on inter-word attention and simply excludes special tokens, which is a questionable method, as attention to special tokens actually matters at inference time.

The functions of special tokens are not yet well understood. [CLS] is typically viewed as an aggregated sentence-level representation – although all token representations also contain at least some sentence-level information. Some researchers experiment with encoding Wikipedia paragraphs with base BERT to consider specifically the attention to special tokens, noting that heads in early layers attend more to [CLS], in middle layers to [SEP], and in final layers to periods and commas. The function of attending to special tokens might be a kind of "no-op": a signal to ignore the head if its pattern is not applicable to the current case. While attention to special tokens increases, their importance for prediction drops. After fine-tuning, both [SEP] and [CLS] get a lot of attention, depending on the task.

BERT LAYERS BERT's input is a combination of token, segment, and positional embeddings. Lower layers have the most *linear word order* information. Knowledge of linear word order decreases around layer 4 (i.e. the middle), and that of *hierarchical sentence structure* increases, as detected by the probing tasks of predicting the index of a token, the main auxiliary verb, and the sentence subject.

There is consensus among studies with different tasks, datasets and methodologies that *syntactic information* (in general, and especially syntactic tree depth and subject-verb agreement) is the most prominent in the middle BERT layers. This must be related to that the middle layers of Transformers are overall the best-performing and the most *transferable* across tasks. There is conflicting evidence about syntactic *chunks*: Some researchers draw parallels to the order of components in a typical NLP pipeline from POS-tagging to dependency parsing to semantic role labeling; others show that lower layers were more useful for chunking, while middle layers were more useful for parsing; yet others find the opposite: both POS-tagging and chunking were performed best at the middle layers, in both BERT-base and BERT-large.

The *final layers* of BERT are the most task-specific: In pre-training, this means specificity to the MLM task, which would explain why the middle layers are more transferable. In fine-tuning, it explains why the final layers change the most.

Semantics is spread across the entire model. While most of syntactic information can be localized in a few layers, in semantic tasks, certain nontrivial examples get solved incorrectly at first but correctly at higher

layers, e.g. predicate-argument relations help to disambiguate parts of speech. This is rather to be expected: semantics permeates all language, and linguists like Goldberg (2006) debate whether meaningless structures can exist at all. What does stacking much more Transformer layers actually achieve in BERT in terms of the spread of semantic knowledge, and is that beneficial? Base and large BERTs shows the same overall pattern of cumulative score gains, only more spread out in the large BERT. This picture is disputed by other researchers, who place "surface features in lower layers, syntactic features in middle layers and semantic features in higher layers", but only one SentEval semantic task in the corresponding study actually topped at the last layer, three others peaked around the middle and then degraded by the final layers.

4.3.2.5 *Training BERT*

MODEL ARCHITECTURE CHOICES The most systematic study of BERT's architecture investigated the number of layers, heads, and model parameters, varying one option a time, and freezing the others. The number of heads was not as significant as the *number of layers*, consistently with research that found the middle layers to be the most transferable. Larger hidden representation size was consistently better, but the gains varied by setting.

IMPROVEMENTS TO THE TRAINING REGIME Regarding the batch size, large-batch training (8k examples) improves both the language model perplexity and downstream task performance. With a batch size of 32k, BERT's training time can be significantly reduced with no degradation in performance.

Embedding values of the trained [CLS] token are not centered around zero, its normalization stabilizes the training, resulting in a slight performance gain on text classification tasks. "Warm-start", i.e. training in a recursive manner, where the shallower version is trained first and then the trained parameters are copied to deeper layers, achieves 25% faster training speed with similar accuracy to the original BERT on GLUE tasks.

PRE-TRAINING BERT The original BERT is a bidirectional Transformer pre-trained on two tasks: next sentence prediction (NSP) and masked language model (MLM). Pre-training is the most expensive part of training BERT, and it would be informative to know how much benefit it provides. On some tasks, a randomly initialized and fine-tuned BERT obtains competitive or higher results than the pre-trained BERT. Most weights of pre-trained BERT are useful in fine-tuning, although there are "better" and "worse" subnetworks. One explanation is that pre-trained weights help the fine-tuned BERT find wider and flatter areas with smaller generalization error, which makes the model more

robust to overfitting. Most new models' gains are often marginal, and estimates of model stability and significance testing are very rare.

The following topics have been investigated to improve pre-training.

How to mask? There are systematic experiments with corruption rate and corrupted span length; diverse masks for training examples within an epoch; masking every token in a sequence instead of a random selection; replacing the MASK token with [UNK] token, to help the model learn a representation for unknowns that could be useful for translation; and maximizing the amount of information available to the model by conditioning on both masked and unmasked tokens, and letting the model see how many tokens are missing.

What to mask? Alternatives include full words instead of word-pieces and spans rather than single tokens (predicting how many are missing). Masking phrases and named entities improves representation of structured knowledge.

Alternatives to masking. Experiments have been performed for replacing and dropping spans; deletion, infilling, sentence permutation and document rotation; for predicting whether a token is capitalized and whether it occurs in other segments of the same document; training on different permutations of word order in the input with the objective of maximizing the probability of the original word order; and the detection of tokens that were replaced by a generator network.

NSP alternatives and additional tasks. Removing NSP does not hurt or slightly improves performance. It has been replaced with the task of predicting both the next and the previous sentences; or identifying swapped sentences. Another model includes sentence reordering and sentence distance prediction with two new tasks on two levels. On the token-level: it has to be predicted whether a token is capitalized and whether it occurs in other segments of the same document; and the segment-level tasks include sentence reordering, sentence distance prediction, and supervised discourse relation classification. In another approach, both NSP and token position embeddings have been replaced by a combination of paragraph, sentence, and token index embeddings. Utterance order prediction for multiparty dialogue has also been proposed. Rogers, Kovaleva, and Rumshisky cite cross-lingual work as well.

Approaches include combining MLM with some other tasks: simultaneous learning of seven tasks, including discourse relation classification and predicting whether a segment is relevant for IR; latent knowledge retrieval; knowledge base completion. Continual learning means sequential pre-training on a large number of tasks, each with their own loss which are then combined.

Pre-training data. Several studies explored the benefits of increasing the corpus volume; longer training; explicit linguistic information, both syntactic and semantic; using the label for a given sequence from an annotated task dataset (e.g. sentiment analysis); and learning representations for rare words separately.

The idea of explicitly supplying structured knowledge has been experimented with in different ways, including entity-enhanced models (including entity embeddings as input or adapting entity vectors to BERT representations); an additional pre-training objective of knowledge base completion; modifying the standard MLM task to mask named entities; training with MLM objective over text + linearized table data; or enhancing RoBERTa with both linguistic and factual knowledge with task-specific adapters.

FINE-TUNING BERT The *pre-training + fine-tuning* workflow is a crucial part of BERT. Pre-training is supposed to provide task-independent linguistic knowledge, while the fine-tuning process would presumably teach the model to extract information from the representation.

During fine-tuning BERT, the most changes for 3 epochs occurred in the last two layers. Those changes caused self-attention to focus on [SEP] rather than on linguistically interpretable patterns. It is understandable why fine-tuning increases the attention to [CLS], but the increase on [SEP]needs some explanation. As [SEP] may serve as "no-op" indicator, fine-tuning basically may tell BERT what to ignore. In multilingual BERT, fine-tuning affected both the top and the middle layers of the model.

Studies explored the possibilities of improving the fine-tuning of BERT by taking more layers into account: combining deeper layers with the output layer or a weighted representation of all layers; two-stage fine-tuning with an intermediate supervised training stage; adversarial token perturbations that improve the robustness of the model; or mixout regularization, which improves the stability of BERT fine-tuning even for a small number of training examples.

With larger and larger models even fine-tuning becomes expensive, but this cost has been limited by adapter modules, which have been also used for multi-task learning and cross-lingual transfer; by reusing monolingual BERT weights for cross-lingual transfer; or by extracting features from frozen representations.

Initialization can have a dramatic effect, which is not often reported: performance improvements claimed in many NLP modeling papers may be within the range of that variation. Significant variation has been reported for BERT fine-tuned on GLUE: both weight initialization and training data order contribute to the variation. Some authors propose an early-stopping technique to avoid full fine-tuning for the less-promising seeds.

4.3.2.6 *How big should BERT be?*

OVERPARAMETRIZATION Transformer-based models keep increasing in size, e.g. T5 (Raffel et al. 2020) is over 30 times larger than the base BERT. This raises concerns about the computational complexity

of self-attention, environmental issues, and reproducibility and access to research resources in academia vs. industry. Current models do not make good use of the parameters: all but a few Transformer heads can be pruned without much loss in performance, most BERT heads in the same layer show similar self-attention patterns, and most layers can be reduced to a single head.

Depending on the task, there may be harmful BERT heads/layers. For machine translation and GLUE tasks, both heads and layers could be advantageously disabled. In a structural probing classifier, 5 out of 8 probing tasks show some layers (typically the final one) to cause a drop in scores. Comparing BERT-base and BERT-large, the larger model performs better many times, but the opposite was observed for subject-verb agreement and sentence subject detection. Why does BERT end up with redundant heads and layers? It is not clear given the complexity of language, and amounts of pre-training. The reason was suggested to be the use of attention dropouts.

COMPRESSION BERT can be efficiently compressed with minimal accuracy loss. In a knowledge distillation framework, a smaller student network is trained to mimic the behavior of BERT. Variants include mimicking the activation patterns of individual portions of the teacher, and knowledge transfer at different stages (pre-training or fine-tuning). Another method is quantization of weights, which often requires compatible hardware. Other techniques include decomposing BERT's embedding matrix into smaller matrices.

PRUNING AND MODEL ANALYSIS Care has to be taken in linguistic analysis. For example, BERT has heads that seem to encode frame-semantic relations, but disabling them might not hurt downstream task performance, which suggests that this knowledge is not actually used. A study identified the functions of self-attention heads and then checked which of them survive the pruning, finding that syntactic and positional heads are the last ones to go. An approach in the opposite direction is pruning on the basis of importance scores, and interpreting the remaining "good" subnetwork. It does not seem to be the case that only the heads that potentially encode nontrivial linguistic patterns survive the pruning.

Models and methodology in these studies differ, so the evidence is inconclusive. Head and layer ablation studies have limitations: they inherently assume that certain knowledge is contained in heads/layers despite evidence of more diffuse representations spread across the full network, i.e. the gradual increase in accuracy on difficult semantic parsing tasks, and the absence of heads that do parsing "in general". Ablating individual components may harm the weight-sharing mechanism, and ablations are also problematic if information is duplicated in the network.

4.3.2.7 *Multilingual BERT*

In version 1 of the paper, Rogers, Kovaleva, and Rumshisky (2020) discussed the Multilingual BERT (mBERT), which was trained on Wikipedia in 104 languages (with a 110K wordpiece vocabulary). (The reader interested in pre-trained multilingual deep language models should also refer to Doddapaneni et al. (2021).) Languages with a lot of data were subsampled, and some were super-sampled. mBERT is surprisingly good in zero-shot transfer on many tasks, but not in language generation. It has been used to create high-quality cross-lingual word alignments, with caution for open-class parts-of-speech. Adding more languages does not seem to harm the quality of representations. mBERT transfers knowledge across some scripts, and retrieves parallel sentences, although it has been noted that this task could be solvable by simple lexical matches. The representation space shows some systematicity in between-language mappings. "Translation" is possible by shifting the representations by a so called sentences offset. However, mBERT does not learn systematic transformations of structures to accommodate a target language with different word order, e.g. SOV instead of SVO, or a different adjective/noun order.

mBERT is simply trained on a multilingual corpus, with no language IDs, but it encodes language identities. Adding the IDs in pre-training was not beneficial. It reflects at least some typological language features, and transfer between structurally similar languages works better. This implies that mBERT could not be considered as interlingua, because its representation space is structured by typological features. Cross-lingual transfer can be achieved by only retraining the input embeddings while keeping monolingual BERT weights, i.e. even monolingual models learn generalizable linguistic abstractions. Compared with English BERT, at least some of the syntactic properties hold for mBERT: MLM is aware of four types of agreement in 26 languages, and the main auxiliary of the sentence can be detected in German and Nordic languages.

There have been conflicting results whether shared word-pieces help mBERT. The simplest formalization of this question is whether performance correlates with the amount of shared vocabulary. Proposals for improving mBERT include fine-tuning on multilingual datasets by freezing the bottom layers; improving word alignment in fine-tuning; translation language modeling as an alternative pre-training objective where words are masked in parallel sentence pairs; and combining five pre-training tasks (monolingual and cross-lingual MLM, translation language modeling, cross-lingual word recovery, and paraphrase classification). The monolingual BERT has been applied directly in cross-lingual setting, by initializing the encoder part of the neural MT model with monolingual BERT.

4.3.2.8 *Directions for further research*

BERT was shown to rely on shallow heuristics in natural language inference, reading comprehension, argument reasoning comprehension, and text classification. Such heuristics can even be used to reconstruct a non-publicly-available model, suggesting a shortcut in the data. It has been realized in the past years that the development of harder *datasets that require verbal reasoning* should be as valued as modeling work. "Amnesic probing" targets what knowledge actually gets used by identifying features that are important for prediction in a given task.

4.3.3 *The geometry of word senses*

Coenen et al. (2019) discover separate semantic and syntactic subspaces in BERT representations: a fine-grained geometric representation of word senses, and syntactic representations in attention matrices and individual word embeddings. In this section, we summarize the former, i.e. their finding that BERT distinguishes word senses at a very fine level. Much of this information in encoded in a relatively low-dimensional subspace.

The operation of BERT has the following components:

- the input to BERT is based on a sequence of tokens (words or pieces of words),
- the output is a sequence of vectors, one for each input token, a contextualized embedding, and
- the internals consist of two parts. The initial embedding for each token is created by combining a pre-trained wordpiece embedding with position and segment information; and the initial sequence of embeddings is run through multiple transformer layers producing a new sequence of context embeddings at each step. In each transformer layer is a set of attention matrices, one for each attention head, and each head contains a scalar value for each pair of tokens.

Context embeddings in BERT and related models contain enough information to perform tasks in the NLP pipeline with simple classifiers (linear or small MLP models). Such single global linear transformations have been termed "structural *probes*" (Belinkov et al. 2017; Conneau et al. 2018; Hewitt and Manning 2019).

4.3.3.1 *Visualization of word senses*

Taking sentences from the introductions to English-language Wikipedia articles, Coenen et al. retrieved 1,000 sentences for individual words, and visualized the corresponding BERT-base context embeddings using UMAP. With the example of *die*, they find crisp, well-separated

clusters: the German article, 'stop living', and the game tool. Within 'stop living', there is a kind of quantitative scale, related to the number of people dying. They ask the questions whether it is possible to find quantitative corroboration that word senses are well-represented; and the seeming contradiction: whether the positions in the clusters represent syntax or semantics.

4.3.3.2 *Measurement of word sense disambiguation capability*

Coenen et al. train a simple classifier on BERT's internal representations for WSD following the procedure described by Peters et al. (2018), i.e. a nearest-neighbor classifier, considering centroids of a given word sense's BERT-base embeddings in the training data. They achieve a higher F1 score than the previous state of the art, with accuracy monotonically increasing through the layers. An even higher score was obtained using the technique in next paragraph.

4.3.3.3 *WSD in a 128-dimensional space*

Coenen et al. hypothesize a linear transformation under which distances between embeddings would better reflect their semantic relationships. They trained a probe following Hewitt and Manning (2019)'s methodology, i.e. a matrix $B \in R^{k \times m}$, testing different values for m. The loss is, roughly, defined as the difference between the average cosine similarity between embeddings of words with different senses, and that between embeddings of the same sense. In evaluation on WSD, untransformed BERT embeddings achieve a state-of-the-art accuracy rate of 71.1%. Trained probes achieve slightly improved accuracy down to $m = 128$. Regarding layers, there is only a modest improvement in accuracy for final-layer embeddings. The method more dramatically improves the performance of embeddings at earlier layers: there is much semantic information in the geometry of earlier layers. The finding offers a resolution to the seeming contradiction mentioned above: syntax and semantics reside in separate complementary subspaces.

4.3.4 *Self attention entropy and ambiguous nouns*

NMT has achieved new state-of-the-art performance in translating ambiguous words. Tang, Sennrich, and Nivre (2019) is interested in which component dominates disambiguation. They consider hidden states, and investigate the distributions of self-attention, training a classifier to predict whether a translation is correct given the representation of an ambiguous noun. They find that encoder hidden states outperform static word embeddings significantly, which indicates that encoders adequately encode relevant information for disambiguation. In contrast to encoders, the effect of decoder differs by models. Most interestingly,

attention weights and attention entropy show that self-attention can detect ambiguous nouns and distribute more attention to the context.

Tang, Sennrich, and Nivre train a classifier which is fed a representation of ambiguous nouns and a word sense (represented as the embedding of a translation candidate). The classifier has to predict whether the two representations match.

They compare word embeddings and encoder hidden states at different layers both from RNNS2S (Luong, Pham, and Manning 2015) and the Transformer (Vaswani et al. 2017). Tang, Sennrich, and Nivre find the following.

- Encoders encode lots of relevant information for WSD into hidden states, even in the first layer. The higher the encoder layer, the more relevant information is encoded.
- Forward RNNs are better than backward RNNs in modeling ambiguous nouns.
- Decoders hidden states have different effects on WSD in Transformer and RNNS2S.
- Self-attention focuses on the ambiguous nouns themselves in the first layer, and keeps extracting relevant information from the context in higher layers.
- Self-attention can recognize the ambiguous nouns and distribute more attention to the context words compared to dealing with nouns in general.

4.3.5 *Psycholinguistic diagnostics*

Ettinger (2020) introduces a suite of diagnostics drawn from psycholinguistic experiments, that allow us to ask targeted questions about the information used by LMs. The results are that BERT can generally distinguish good sentence completions from bad ones involving shared category or role reversal, albeit with less sensitivity than humans; it robustly retrieves noun hypernyms; but struggles with challenging inferences and role-based event prediction with a clear insensitivity to the contextual impacts of negation. She is conservative in the conclusion because these sets are small, and different formulations may yield different performance.

Her diagnostics target a range of linguistic capacities, drawn from psycholinguistics (but she does not test whether LMs are psycholinguistically plausible). The psycholinguistic origin of the test has advantages: it is carefully controlled to ask targeted questions about linguistic capabilities, it asks the questions by examining word predictions in context, which is natural in the LM paradigm, and it allows us to study LMs without any need for task-specific fine-tuning. As we will

see in Section 4.3.5.2, these diagnostics are chosen specifically to reveal insensitivities in predictive models. The problematic nature of the senteces is evidenced by patterns that they elicit in human brain responses, namely N400, which is a famous component of time-locked electroencephalography (EEG) signals known as event-related potentials. N400 is a negative-going deflection that peaks around 400 milliseconds post-stimulus onset. Ettinger goes beyond the syntactic focus seen in existing LM diagnostics, and target commonsense/pragmatic inference, the knowledge of semantic roles and events, category membership, and negation.

Each of Ettinger's diagnostics is set up to support tests of word prediction accuracy and sensitivity to distinctions between good and bad context completions. Ettinger focuses on the BERT model, but the diagnostics are applicable for testing any LM. She publishes a new set of targeted diagnostics for assessing linguistic capacities that shed light on strengths and weaknesses of the popular BERT model.

4.3.5.1 *Related Work*

The related work section includes work on fine-grained classification tasks to probe information in sentence embeddings, token-level and other sub-sentence level information in contextual embeddings, specific linguistic phenomena such as function words, the overall level of "understanding" (semantic similarity and entailment), and curated versions of these tasks to test for specific linguistic capabilities. The analysis of linguistic capacities of LMs has been dominated by syntactic testing.

The *internal dynamics* underlying how LMs cape syntactic information has been examined in different components of the LM and at different timesteps within the sentence, in individual units, and regarding semantic phenomena like negative polarity items. (This line of analysis is firmly rooted in the notion of detecting structural dependencies.) *Word prediction* accuracy has been applied as a test of LMs' language understanding with the LAMBDA dataset, which tests a models' ability to predict the final word of a passage, in cases where the final sentence alone is insufficient to do so. LAMBDA is not controlled to isolate and test the use of specific types of information.

The linguistic characteristics of the *BERT model itself* have also been examined. Regarding the dynamics of BERT's self-attention mechanism, probing attention heads for syntactic sensitivity found that individual heads specialize strongly for syntactic and coreference relations. The syntactic awareness in BERT has been also examined by syntactic probing at different layers and the examination of syntactic sensitivity in the self-attention mechanism. A variety of linguistic tasks have been tested at different layers. BERT has been found to exhibit very strong performance on several of the targeted syntactic evaluations.

4.3.5.2 *Leveraging psycholinguistic studies*

The fourth section in Ettinger provides background on human language processing, and explains how she uses this kind of information to choose the tests. Psychologists test human responses to words in context, in order to better understand the information that our brain uses to generate predictions. Two types of predictive human responses are relevant here.

In the *cloze* test, humans are given an incomplete sentence and tasked with filling their expected word in the blank. This is the ideal human prediction in context, not under any time pressure, so participants have the opportunity to use all available information from the context.

The brain response *N400* can be detected by measuring electrical activity at the scalp by EEG to gauge how expected a word in a context is. The electrical signal appears to be sensitive to the fit of a word in context. It correlates with the cloze in many cases, it can be predicted by LM probabilities, and, importantly, expectations reflected in the N400 sometimes deviate from the more fully-formed expectations reflected in the untimed cloze response.

Ettinger draws diagnostic tests from human studies that have revealed divergences between cloze and N400 profiles, i.e. when the N400 response suggests a level of insensitivity to certain information in computing expectations, causing a deviation from the fully-informed cloze predictions. These present particularly challenging prediction tasks, tripping up models that fail to use the full set of available information.

4.3.5.3 *Datasets*

Each of Ettinger's diagnostics support three types of testing: word prediction accuracy, sensitivity testing, and the qualitative analysis of prediction. These diagnostics are constructed to constrain the information relevant for making word predictions. In *word prediction* evaluation accuracy, Ettinger uses the most expected items from human cloze probabilities as the gold completions. In what she calls *sensitivity testing*, Ettinger compares model probabilities for good versus bad completions — specifically, those on which the N400 showed reduced human sensitivity. The question is whether LMs will show similar insensitivities. The *qualitative analysis of models' top predictions* is also informative, because these items are constructed in a controlled manner.

In all tests, the target word to be predicted falls in the final position, which fits the computational models, both left-to-right or bidirectional ones, only token probabilities in context are concerned, and the method is equally applicable to the masked LM setting of BERT and to a standard LM. Ettinger filters out items for which the expected word is not in BERT's single-word vocabulary.

The observations, which we already summarized at the beginning, are based on the following data-sets:

CPRAG-102 tests sensitivity to differences within semantic category, the name stands for Commonsense and pragmatic inference. In the example *He complained that after she kissed him, he couldn't get the red color off his face. He finally just asked her to stop wearing that lipstick/mascara.*, commonsense knowledge informs us that red color left by kisses suggests *lipstick*, and pragmatic reasoning allows us to infer that the thing to stop wearing is related to the complaint.

As in LAMBDA, the final sentence is not supporting prediction on its own, but unlike LAMBDA, these items have consistent structure. None of these items contain the target word in context, to test commonsense inference rather than coreference. The average Human cloze probabilities for expected completions is .74. A psycholinguistic study found that inappropriate completions (e.g., *mascara, bracelet*) had cloze probabilities of virtually zero, but N400 showed some expectation for completions that shared a semantic category with the expected completion (e.g., *mascara*, by relation to *lipstick*).

ROLE-88 tests event knowledge and the sensitivity to semantic role reversals, e.g. *The restaurant owner forgot which customer/waitress the waitress/customer had served.* It requires event knowledge about typical interactions between types of entities in the given roles. The authors found that although each completion (e.g., *served*) is good for only one of the noun orders and not the reverse, the N400 shows a similar level of expectation for the target completions regardless of noun order. The sensitivity test targets this distinction. Cloze probabilities show strong sensitivity to the role reversal, with average cloze difference of 0.233 between good and bad contexts.

NEG-136 tests negation along with knowledge of category membership, e.g. *A robin is (not) a bird/tree.* N400 shows more expectation for true completions in affirmative sentences, but it fails to adjust to negation: There is more expectation for false continuations.

A separate psycholinguistic experiment chose affirmative and negative sentences to be more "natural", e.g. *Most smokers find that quitting is (not) very difficult/easy.*, and contrasts these with affirmative and negative sentences chosen to be less natural *Vitamins and proteins are (not) very good/bad.*

4.3.6 *Layers and lexical content*

Wang and Kuo (2020) generate sentence representations from BERT-based word models exploiting that different layers of BERT capture different linguistic properties. The task of sentence embedding, i.e. trans-

forming a sentence to a vector, is not trivial. A common approach with BERT-based models is to average the representations obtained from the last layer or using the [CLS] token. The authors show that both are sub-optimal. They fuse information across layers to find better sentence representation: Wang and Kuo dissect BERT-based word models through a geometric analysis of the space in an unsupervised fashion.

Different layers of BERT learn different abstraction levels: intermediate layers encode the most transferable features, and higher layers are more expressive in high-level semantic information. Information fusion across layers has great potential. Wang and Kuo experiment on patterns of the isolated word representations across layers, and find that the evolution of isolated word representation patterns across layers highly correlate with word content: words of richer information have higher variation in their representations. This finding helps them define "salient" word representations and informative words for sentence embeddings.

Wang and Kuo compare their model, SBERT-WK with the following 10 (parameterized and non-parameterized) methods: the average of GloVe word embeddings; the average of FastText word embedding; the average of the last layer token representations of BERT; [CLS] embedding from BERT, originally used for next sentence prediction; the SIF model (Arora, Liang, and Ma 2017), which is a non-parameterized model, a strong baseline in textual similarity tasks; the p-mean model that incorporates multiple word embedding models; Skip-Thought; InferSent with both GloVe and FastText versions; the Universal Sentence Encoder, which is a strong parameterized sentence embedding using multiple objectives and a transformer architecture; and Sentence-BERT, which is a SOTA sentence embedding model with a Siamese network over BERT. SBERT-WK improves the performance on textual similarity tasks by a significant margin. Regarding supervised downstream tasks, SBERT-WK obtains the best result in 5 of the 9 considered tasks, and also in average. The merit of the model is in part due to its efficiency.

Part II

MAIN CONTRIBUTIONS

Our main contributions investigate lexical relations in a very broad sense: besides lexical relations proper (i.e. relations that hold between the meanings of words independent of context, e.g. hypernymy, antonymy, and causality), we include thematic and syntactic relations, word analogies, translation, and ambiguity.

The first two chapters investigate *verbs and their arguments*. Chapter 5 investigates arguments structure: we provide a thematic categorization of arguments in the `4lang` framework. Our main question is what inventory of thematic roles is needed for the formulaic definition of each word in the defining vocabulary of this multilingual and radically monosemic semantic formalism.

Still on verb arguments, but moving from the symbolic treatment of thematic roles to the distributional representation of „syntactic roles" (i.e. grammatical functions), Chapter 6 investigates the use of different automatic association scores and tensor decomposition methods in the context of collocation extraction.

The last two chapters are motivated by the question whether relations which can be captured by intuition and recorded by human labor (as witnessed by their literature in psychology and linguistics), can also be detected in data-driven distributional representations, more specifically, static word embeddings (word representations obtained with shallow neural networks).

Chapter 7 investigates several lexical relations: hypernymy (what basic category a word belongs to, e.g. *dogs* are *animals*), antonymy (opposite meaning), causality, analogy, and translation.

Our last chapter is concerned with one of the greatest problems in lexical semantics: *word ambiguity* and, more specifically, homonymy and polysemy. Static word embeddings, our main tools in the last two chapters, represent each word form with a single linear algebraic vector. Chapter 8 proposes an evaluation method for multi-sense (static) word embeddings (MSEs), where the different senses of an ambiguous word are represented with different vectors.

Az, ki tőlem elrabolna / Lelkemtől rabolna meg...
'That who stole you from me would rob me of my soul'[8]

— Béni Egressy

5

DEEP CASES

5.1 OVERVIEW

Verbs are the backbone of sentences, expressing actions, events, and relationships between entities. They denote many kinds of semantic connections. In the field of computational lexical semantics, the study of verbs and their arguments has been a central pursuit, aiming to capture the nuanced meanings and syntactic patterns associated with this part of speech.

Now, that the previous chapters provided the background in computational lexical semantics, the first two main chapters will investigate verbs and their arguments: thematic roles in a symbolic approach, and

8 This motto from the libretto of a Hungarian opera is intended here as a Hungarian pun, but we try to explain the joke: Both clauses contain the verbal stem *rabol* 'rob', a pro-dropped syntactic object (an unmarked construction in Hungarian syntax), and an ablative-marked overt argument, but there is a mismatch in the grammatical functions (surface cases):

	(pro-dropped) object	ablative ('from')
el rabolna	Object	maleficient
rabolna meg	maleficient	Object

In the first clause, the ablative is arguably oblique and the preverb *el* 'away' is adverb-like, while in the second clause, the ablative is quirky and the adverbial *meg* is a pure perfectivizer.

grammatical functions in a distributional one, respectively. This chapter, which originally appeared as Makrai (2014b) in Hungarian[1] investigates the argument linking system in the 2014 version[2] of `4lang`, the semantic network we introduced in Chapter 3. Our discussion is based both on theoretical principles, and on our experience in creating the formulaic meaning representation of each item in the defining vocabulary. We have seen in Sections 2.4.4 to 2.4.6 that the main difference between modern resources for the representation of verb argument structure is in the granularity of the argument labels. Accordingly, our main question is what inventory of thematic roles is needed for the definition of each word in the defining vocabulary of a multilingual and radically monosemic semantic formalism.

As we have already seen in Chapter 3, `4lang` is a multilingual lexicon for general human language understanding containing formal representations of word meaning in the monosemic approach to lexical semantics, which means that items are language independent concepts covering different uses of the same word, uses in different sentence patters and even in different parts-of-speech with the same meaning representation.[3] Multilinguality and abstractness of items have the effect that a simple deep case (or thematic) frame captures uses with different arity (i.e. transitive and intransitive). Deep cases denote the nodes in the graph representing the meaning of a predicate where the representation of the argument (single word, entity or phrase) has to be inserted.

`4lang` makes no clear cut between complements and adjuncts. Basically an argument is represented by a deep case whenever it is needed for building the representation of the verb. As uses of the same verb with different arities are handled in the same item, deep cases are used consequently in different verb patterns, and all possible arguments are included in the representation (Section 5.4.2.1). However, as verbs can be defined as special cases of other verbs (biting is cutting with teeth), arguments are inherited, so not every argument is listed directly in the definition of some verb. Another source of implicit arguments are constructions providing verbs with outer arguments, e.g. *paint a picture* ***for somebody***, *sleep* ***an hour***, *fly* ***the Atlantic***. Causatives (e.g. *march* ***the soldiers***) are also attributed to constructions rather than to argument structure.

The most frequent verbal deep cases (Section 5.4.2, especially Section 5.4.2.2) are agents (denoted by `AGT`), patients (`PAT`), and datives

1 Most of the definitions were written by Makrai, and, in later phases of the project, it was him who developed them, Kornai advised. There were group discussions on the deep cases, the approach described here is that of Makrai. The paper Kornai and Makrai (2013) was written by Kornai and presented by Makrai. Makrai is the only author of the Hungarian Makrai (2013).

2 See Section 5.6 for explanation.

3 The lexicon, automatically collected word forms in 50 languages, a vector space language model (embedding) computed from `4lang`, and articles can be found at http://hlt.sztaki.hu/resources/4lang/

(DAT), already familiar from Section 2.3.2. Patient plays the role of the neutral case it seems to play in many systems (Somers 1987). Following the Unaccusative Hypothesis, arguments of intransitive verbs split to agents and patients. The label "dative" is taken from Fillmore (1968), but our understanding is narrower as we mainly restrict dative to recipients in ditransitives (verbs of communication (e.g. *tell*) and transfer (e.g. *give*)). (Psychological experiencer verbs and predicative arguments will be discussed in Section 5.4.2.3.) These verbs correspond to Schank's (see Section 2.2.3) transmission types, MTRANS and PTRANS. There are three locative cases in 4lang (TO, FROM, and AT). TO is used for the abstract goal of relational nouns such as *occasion* and *need* as well (Section 5.4.3). E.g. the definition of the former is `time, =TO AT/2744`. A greater group of relational nouns require the possessive (POSS) such as *absence* and *duty*. In the theory, quirky cases (e.g. *prefer something* ***to something***) can be marked in a language dependent module.

Deep cases in 4lang are not restricted to verbs. Some grammatical features such as plural contribute to meaning. Technically, the definition of these morphemes refer to the referent with REL. Representations of productive derivational suffixes and adpositions also refer to the conceptual element they attach to with REL.

5.2 WHAT DO ARGUMENT LABELS DO?

To calculate the meaning representation of a sentence, we need to map the predicate-argument relationships. From a theoretical linguistic point of view, we have two pillars here: selection constraints and surface cases in the broadest sense (e.g. the order of phrases, case affixes and/or adpositions varying from language to language). In our opinion, selection constraints correspond to *spreading activation* (Section 2.2.2) in the dictionary, and the knowledge about surface cases is indirectly encoded by deep cases. From the point of view of deep cases, it is important that 4lang is designed to connect to each language with a language-specific module, which tells which surface cases will realize each deep case in that language. Recall from Section 3.7 that the linking module has since been partially implemented in accusative languages by colleagues. The implementation uses a dependency analyses in the framework of Universal Dependencies (UD, Nivre et al. (2016))), and associates dependency types (i.e. subject and object) to agent and patient. In this chapter we deal with deep cases, so we outline the activation spreading only briefly and in a simplified way.

Recall the definition graph, the vertices of which are concepts in the dictionary, and two of these are connected if one is included in the definition of the other (Section 3.3), e.g. 'milk' is associated with 'liquid'. If we want to know which argument of *drink* the word *milk* fills in a sentence, we should look for the shortest path (edge sequence) between the two concepts in the graph. With some luck, this passes

through the word *liquid* and largely corresponds to the representation of the phrase *drink milk*.

Let us now turn to how the role of each argument (which slot it fills) can be calculated from surface cases. The meaning representation of a term that includes a predicate with its arguments (e.g. a verb phrase) should be calculated from the following: the representation of the meaning of the predicate, that of the arguments, and the structure of all these together. In the case of `4lang`, the latter is taken care of by indicating in the meaning representation of the predicate (typically a verb) where the meaning representation of each argument should go. To do this, we need to be able to distinguish the arguments of higher arity predicates (e.g. transitive verbs). This is done with reference to the deep case of the argument. The background for our method is the common assumption (Section 2.3.2) that, at least within languages, there are regular correspondences between the semantic role (e.g. agent) and the syntactic properties of the arguments (the surface case of the argument, which sentence alternations the verb participates in), and in several cases these regularities shows up in more languages.

5.3 THE GRANULARITY OF THE CASE LABELS

In Section 2.4, we discussed modern lexical semantic resources. Focusing on verb resources, the main difference is in the granularity of the argument labels: FrameNet (Section 2.4.4) uses verb-specific tags, e.g. the Apply-heat Frame includes a Cook, Food, and a Heating Instrument.

PropBank, on the other extreme, uses very generic labels such as Arg0, Arg1, ..., among which Arg0 is generally a prototypical Agent, Arg1 is a prototypical Patient or Theme, but there are no consistent generalizations for the higher numbered arguments, e.g. Arg2 can be beneficiary, goal, source, extent or cause. There are several more general ArgM (Argument Modifier) roles that can apply to any verb, and which are similar to adjuncts, e.g. LOCation, EXTent, ADVerbial, CAUse, TeMPoral, MaNneR, and DIRection.

VerbNet (Section 2.4.5) has a granularity between FrameNet and PropBank with semantic roles like Agent, Patient, Theme, Experiencer, etc., 24 in total. `4lang` follows a monosemic approach (Section 3.1.1), i.e. we strove to make as few distinctions as possible, but we also wanted to make meaningful abstractions, what resulted in an inventory which is finer than that of PropBank, but more abstract that that of VerbNet.

As we have already mentioned, our deep cases only serve to identify which argument is which. In this context, it is perhaps worth emphasizing that the classification of arguments into deep cases is not primarily a semantic classification. In computational semantics, the fact that there is a regular difference between the meanings of the corresponding arguments could often be an argument in favor of distinguishing between two deep cases. For example, Talmy attributes the intentional differ-

ence between the verb pairs *hide/mislay, pour/spill, ...* to the exact nature of the case of the agent.

In Allen and Teng (2018)'s view, semantic roles should have consequences independent of the predicate or event. They explore three aspects: entailment from a role independent of the type that has such roles; integration with ontology (Roles should obey the typical entailments in an ontology, e.g. inheritance of properties from parents); and derivability (roles should be derivable from the definitions in dictionaries). These authors admit that only the third property allows empirical evaluation. In `4lang` such differences do not justify the introduction of a new deep case, as the meaning is fully described in the definition field of the lexical entry.

Compared to semantic classification, the other extreme is where the number of cases cannot exceed the largest number of arguments we encounter among verbs. We do not strive for this either, as we want to take advantage of regularities between the semantic role and the syntactic properties.

5.4 INDIVIDUAL RELATIONS

5.4.1 *Function morphemes*

How does `4lang` grasp simpler dependencies? On the one hand, certain inflections, such as the plural, have a conceptual meaning in the sense that in the representation of the structure containing the inflectional affix, there is an element for which the inflectional affix is responsible. Productive derivational affixes and adpositions are similar. We need to treat these relations (stem–inflectional affix, stem–derivational affix, adpositional object–adposition) uniformly already because `4lang` wants to be language-independent, and the same semantic relation is expressed differently in different languages, e.g. the meaning, which is expressed by the possessive personal suffix in Hungarian, is expressed by the possessive pronoun in English. Here, the place of the representation of a function morphemes in the representation of the more content element is always represented by the keyword `REL` (*relational, related*), which in a broader sense can be called a deep case.

5.4.2 *Verbal deep cases*

5.4.2.1 *Argument positions, alternations, open case inventory*

Turning now to the arguments of verbs, we must first clarify what we mean by an argument. Only the obligatory ones or the adjuncts as

well?[4] Are we talking about surface arguments, or the arguments for the (deep, logical) predicate corresponding to the verb in a formal semantic translation? In the first approximation, we follow the literature (Somers 1987) in representing those surface arguments by their deep case in the definition of a verb that are needed to describe the meaning. Another issue arises from the fact that, due to the abstract nature of `4lang`, we do not differentiate between the transitive use of a verb (or even that with more surface arguments) and the intransitive use of the same verb form. Deep cases are defined in such a way that the same predicate in different uses gets the same case. It follows that if a verb has a transitive use, the deep case of two participants must also be indicated. Finally, a further nuance is that when a verb can be defined as a special case of another verb and the arguments are inherited, it is not necessary to explicate them in the definition, e.g. *bite* is defined as `cut, INSTRUMENT tooth` ('cut with tooth'), and *bite* inherits the arguments of *cut*, so these are not listed.

In choosing deep cases, it is not our task to create harmony between the participants of different verb roots. Thus, for example, it is not our intention that the participants in the sentences *John sells a book to Peter* and *Peter buys a book from John* will receive the same deep cases for the two sentences.[5] Finally, we do not include *outer roles* in the verb definition, that is, the possible arguments that can be assigned to a verb by a construction that affects entire verb classes (e.g. motion verbs) or even all verbs, so in the following examples the putative argument position corresponding to the bold face phrases: *paint an image* ***for someone****, sleep* ***an hour****, fly over* ***the Atlantic Ocean***.[6] Causation is also considered such a construction. (In Hungarian, the meta-language of the paper on which this chapter is based, causation is marked by the derivational suffix *-(t)At.*)

5.4.2.2 *The core (agent, patient, dative)*

There are 744 verbs in `4lang`. Deep cases are listed in the Table 8, along with the number of words that they occur with. Unsurprisingly, the most common deep case is the agent. When writing definitions, we can decide without much difficulty which argument of a typical transitive verb is the agent (indicated by the keyword `AGT` in the dictionary). The second most common deep case in `4lang`, which we called patient (`PAT`), is often defined only as the "semantically unmarked" deep case, but since the others are relatively clearly identifiable, this is not a problem either. According to the Unaccusative Hypothesis widely accepted

4 This chapter was originally published in Hungarian, where there is a common term for arguments and adjuncts, *bővítmény* 'expansion', arguments proper are called *vonzat* 'attractee', and adjuncts proper are called *szabad bővítmény* 'free expansion'.

5 In both cases, the English subject will be an agent, and the object will be `PAT`.

6 For more on external roles, see Somers 1987, Chapter 9.

AGT	383
PAT	311
REL	81
POSS	52
DAT	30
TO	17
FROM	11
AT	2

Table 8: Each deep case with the number of predicates using them. As for the granularity of the role inventory, our system is between PropBank/AMR (Sections 2.4.6 and 2.4.8) and VerbNet (Section 2.4.5).

	object-marking	ergative 1	ergative 2	active	lexicalized active	subject-marking
Peter is writing the letter.	nom	ag	ag	ag	ag	ag
Peter is writing.	nom	nom	ag	ag	ag	ag
Peter is walking.	nom	nom	nom	ag	ag/nom	ag
Peter is ill.	nom	nom	nom	nom	ag/nom	ag

object marking	English (eng), Hungarian (hun)
ergative 1	Kabardian (kgb), Avar (ava), Adige (ady)
ergative 2	Aghul (agx), Udi (udi)
active	Bats (bbl)
lexicalized active	Georgian (kat), Dakota (dak)
subject marker	Mingrelian (xmf), Maidu (nmu)

Table 9: Arguments of intransitive verbs in different languages (Komlósy 1982). The SIL code of the languages is also indicated.

in modern syntax, the argument of an intransitive verbs can also be patient (e.g. *fall, melt*).

Komlósy (1982) summarizes how the agent and patient of intransitive and transitive verbs are classified by surface cases in different languages. Komlósy reviews a number of ergative (or active and subject-marking) languages in terms of the case of the arguments of different single-argument verbs. Table 9 shows that different languages draw the line between the two cases at different points on a scale of activity. These data suggest that in a language-independent case system we need to make finer differences than the binary `AGT` vs `PAT` partition. It is a question whether this would really improve the performance of our systems in these languages. Such experiments would exceed the bounds of the present thesis, so we'll stick with the simpler case set.

With agent and patient, we essentially follow the generative semantic tradition. We deviate more from the literature by using the *dative* (`DAT`). The name is taken from the oldest terminology of generative semantics (Heringer 1967; Fillmore 1968). Fillmore himself later separated the dative into experiencers, objects (*Object*), and goals (*Goal*). We basically use the dative only for verbs with at least three surface arguments, in other cases only based on their similarity to the former. As for their meaning, some of the three-argument verbs are the special cases of *say*, very reminiscent of Schank's (see Section 2.2.3) mental transmission MTRANS: *admit, allow, command, declare, emphasize, explain, express, forbid, grateful, say, swear, teach, thank*. Another group is related to *give*, i.e. Schank's physical transmission PTRANS: *bestow, have, help, lend, let, make offering, offer, owe, owing to, pass, pay, present, sell, show*.

5.4.2.3 *Unaccusatives, psychological experiencers, predicative arguments*

The simplest argument structures are when there is an agentive subject with no further argument, or with an optional or obligatory object (e.g. *eat*) or goal (e.g. *join*). The first deviation from these is those optionally transitive verbs where the subject of the intransitive use corresponds to the patient of the transitive use. Recall that in `4lang`, these verbs are also represented by a single item in which both participants are indicated. There are a couple of dozen pure unaccusative verbs *(intransitive verbs with a patient)*: *bath, become, belong to, bend, burn, depend, develop, die, drown, fade, faint, fall, gain, hang, hear, hope, improve, reduce, sleep, spoil, spread, think, tire*. These are represented by PAT. In `4lang`, patient in not restricted to verbs: *divorce* is a "psych-noun".

In *two-participant psych-verbs*, where this specificity is reflected in a difference between some languages (especially English and Hungarian), e.g. *the pony pleases Dave/Dave likes the pony*, we decided to represent the stimulus as a *patient* and the experiencer as a *dative*. In principle, this should also be done with psych-verbs that resemble *like* in all the languages under consideration, i.e. whose experiencer is the subject and the stimulus is the object. In this, unfortunately, the annotation is not fully consistent. On a semantic basis, we assigned both a patient and a target `TO` to *belong to* and *remember*.

Among the *three-participant* verbs, as we have already mentioned, those whose subject is the agent pose no problem: besides the well-populated classes of *give*- and *tell*-verbs, *help*, and even most of those with predicative arguments, such as *let/allow* and *regard*, can easily be represented. Table 10 shows all *predicative arguments* in the authors version[8] of the hand-edited `4lang` definitions. To select these words,

7 In addition to English and Hungarian, there is also a target case on the surface in French *(capable de)*.

8 See Section 5.6.

English	Hungarian	Latin	Polish	id	defining?	POS	def
able	képes	idoneus	w stanie	1245		A	can/1246[=AGT[=TO]][7]
appear	tűnik	pareo	wydawać sié	2450	yes	U	=DAT THINK [=PAT[=OBL]]
command	parancs	iussum	rozkaz	1941		N	speak, HAS authority, CAUSE =DAT[=PAT]
of	-ból	ex	-any	16	yes	G	material[=REL]
recognize	felismer	cognosco	rozpoznać	771	yes	V	=AGT KNOW[=PAT[=PAT]]
regard	tart vminek	arbitor	uważać za	2312		V	=AGT THINK[=PAT[=DAT]]
self	önmaga	ipse	sam	1851		N	=PAT[=AGT]
tendency	tendendcia			2987		N	=POSS[=TO[likely]]
try	próbál	tempto	próbować	1976	yes	V	=AGT WANT =AGT[=PAT]
use	használ	utor	używać	1008	yes	V	=FOR[purpose] INSTRUMENT =PAT, =AGT[=FOR]

Table 10: Predikatív bővítmények a `4lang`-ben.

we searched for words whose definition includes a predicate on a node that is labeled with a deep case. As can be seen here, we have not run into an insoluble problem with predicative arguments in general.

There are three verbs left in the defining vocabulary: *have, appear* and *seem*. Our method during the creation of the definitions was to keep the number of deep cases limited by introducing only those that had enough occurrences in the vocabulary to make an intuitive generalization possible. If we are faithful to that principle, we cannot say anything about these three verbs. If we had to, we could represent the possessor with a dative in the case of *have*, and the possession with the neutral patient (based on languages with a dative surface case).

The peculiarity of *appear* and *seem* is that two participants behave dative-like on the surface: the predicative argument and the experiencer. Which one should be considered a deep dative? =TO remains for the analysis of the other. Note that the dative is related to the goal cross-linguistically, e.g. in Urdu the goal is usually expressed with a dative (Butt 2006, Section 5.5.2.).

Thus, the two possible analyses are

- *seem:* =DAT THINK [=PAT IS-A =TO]
- *seem:* =TO THINK [=PAT IS-A =DAT]

(IS-A is written here just for the ease of presentation. The syntax implemented in definitions is =PAT[=TO] or =PAT[=DAT].)

In the first alternative, the reader may recognize the familiar configuration of roles that the experiencer is a dative, and this alternative also resembles the typical situation where the agent causes something to the patient

- *put:* =AGT CAUSE[=PAT AT =TO]

to the extent that the object of the embedded predicate is a goal. However, in order to choose between the two kinds of analysis, more languages should be considered.

To summarize analysis of psych-verbs: in principle, the stimulus is a patient, and the experiencer is a dative. For verbs with both and

experiencer and a predicative argument – of which there are only two in the basic vocabulary of `4lang`– we do not offer any generalization.

There are some further words in the defining vocabulary with dative marked arguments in Hungarian (*nehéz* 'difficult; heavy', *y* tetszik *x*-`dat` '*x* likes *y*') or German (*ähneln* 'resemble', *beitreten* 'join', *gleichen* 'equal'), but these are too sporadic to draw any generalization, so we treat them as exceptional.

5.4.2.4 *Locative cases*

There are as many as three *locative* cases in `4lang`, `TO` and `FROM` corresponding to the Fillmore Goal (*Goal*) and Source (*Source*), and the essive `AT`. In Section 2.2.6, we reviewed Hayes (1979), in whose approach "to really capture the notion of 'above', you probably have to go into analogies to do with e.g. interpersonal status: Judge's seats are raised; Heaven is high, Hell is low; to express submission, lower yourself, etc." `4lang` has gone as far as possible in abstraction: if an argument in many languages gets a surface case that is also used to express the goal of movement (specific inflectional suffixes in Hungarian, and prepositions in English), then we consider it a goal. We mean *able, accustom, add, addition, available, belong, gentle* (hu:*gyengéd*, la:*mollis*, pl:*delikatny*), *include, invite, join, law, listen, load, mix, necessary, need, occasion, put, ready, remind, sensitive, similar, skill, tendency.* The other two locative cases are the source (*accept, borrow, buy, cut off, date, derive, of, profit, remove, rent, rubber, separate, subtract, take*) and the essive location (*situated, stay*).

In the language-specific module already mentioned it is possible to mark some arguments of some verbs with surface cases, if their case is unpredictable from their deep case (*quirky case*). On the other hand, it is already clear from English, Hungarian and German that there are verbs where no generalization seems useful. In this case, we use the same `REL` keyword as for predicates with a single surface argument, e.g. *prefer to something.*

5.4.3 *Relational nouns*

Finally, consider the relationship between *relational nouns* and the word associated with them (e.g. in the case of *interest*, the stakeholder). The phenomenon that makes the noun *interest* relational is twofold. On the surface, the proportion of possessed occurrences of the word *interest* is significantly higher than among other nouns. On the other hand, which is more interesting from a semantic point of view, no matter how we want to describe the meaning of the word *interest*, we would probably refer to the "stakeholder". The grammatical relationship between the two words is possessive in most relational nouns, but we find something different in about one-tenth of the lexemes. In the case of the words *occasion* and *need*, the participant which we call the *goal* for lack

of a better word, is sublative in Hungarian (*-rA*, lit. onto) and *for* in English. In the representation of relational nouns, we use the keywords `POSS` or `TO` according to the grammatical relationship between the two words to indicate the place where the representation of the related word (the interested person and the target, respectively) goes. `TO` is the same abstract goal we encountered at verbs. Thus deep cases mediate and helps to find the semantic relationship between the two participants (the interested and the interest; the occasion and the goal). We will not handle relational nouns that are productively formed from a verb (i.e. participles), because `4lang` does not distinguish e.g. participles from the corresponding verb.

5.5 LINKING

Recall from Section 3.7 that the manual definitions of `4lang` have been applied to word and sentence similarity and entailment (Recski and Ács 2015; Recski 2016b; Recski, Borbély, and Bolevácz 2016; Recski et al. 2016; Recski 2016a; Ács, Nemeskey, and Recski 2017; Kovács and Recski 2018; Recski 2018).[9] Both the agents (resp. patient) in the manual definitions and the subjects (resp. object) in the dependency analysis have been linked with a 1 (resp. 2) arrow. These applications did not use the remaining deep cases. Specifically, no implementation tested whether the treatment of relational nouns with `POSS` and `TO` described above benefits NLP applications.

5.5.1 *Ergative languages*

Recall from Section 3.7 that Recski et al. (2016) and Kovács, Gémes, Iklódi, et al. (2022) implemented `4lang`-linking with Universal Dependencies (UD). Marneffe et al. (2021, Chapter 4.4) describes the UD treatment of ergative languages as follows:

> A more frequent analysis is to say that such syntactically ergative languages treat the intransitive core argument and the patient-like argument of transitives together as a "pivot" (Dixon 1994), which we would analyze as a subject (`nsubj`), and then the agent-like argument of transitives is also a core argument, which we would analyze as an object (`obj`). The unusual thing, then, is the reversed alignment between semantic roles and grammatical relations. This is a place where the relation subtype `:pass` can be usefully used in an

9 The members of the HLT group can no longer recall exactly which of their publications used the manual definitions. The ones cited here mostly used them, and those since 2019 probably did not, because the colleagues moved to a definitional syntax whose parser in its present state can only check the definitions but it cannot translate them to graphs.

> extended sense. If we regard it as marking not only passives but all cases where the `nsubj` does not mark the agent-like argument of the verb, then all transitive subjects in such a language are `nsubj:pass`. In addition, we can reuse the subtype `:agent`, which in other languages is optionally used for an oblique modifier denoting a demoted agent, to mark the ergative core argument as `obj:agent`.

The relations of such an analysis could be easily mapped to `4lang` deep cases.

5.6 AN OLDER AND A NEWER APPROACH

In Kornai and Makrai (2013) and Makrai, Nemeskey, and Kornai (2013), we labeled argument locations, which we later called deep cases, by names of surface cases (e.g. NOM, ACC, POSS) and the names of classes thereof (e.g. OBL). There was no justification behind this, the name (the literal used as a label) was not considered important.

In the article on deep cases itself, on which this chapter is based, we switched to thematic roles because we thought that these were more in line with the intended language-independent generalization. Our motivation was to capture language-universal regular correspondences between the semantic role (e.g. agent) and the syntactic properties of the arguments (the surface case of the argument, which sentence alternations the verb participates in). This is common in both theoretical and computational linguistics.

Finally, Kornai (2023) parted with most of the earlier deep cases. In this more minimalistic approach to linking, cross-lingual claims are basically restricted to the agent and the patient, which directly correspond to arrow 1 and arrow 2 respectively. For Kornai, it is not important that the "linkers" be named for thematic roles. There are still binary relations, and no other levels. The approach of this recent book is "explicitly formalistic, it looks for the minimum to get things done" (András Kornai, personal communication). See Kornai (2023, Sec 5.6), especially the last part ("This is of course not to deny that there are such things as datives or locatives. . . "). Some of the syntactio-semantic information is expressed with a new relation `mark_`. The definition of (mostly mental) transfer verbs will contain `"dative" mark_ person` (where `person` is unified with the beneficient). Another frequent first argument of `mark_` is the object of *to*, e.g. `"to/3600 _" mark_ act` in the definitions of *able, difficulty,* and *ready.* Kornai (2023, Chapter 8) admits that in this system there are no universal tools for unaccusative and other situations where the subject is placed in the first argument position or the subject in the second.[10]

10 The last version which is compatible with the present thesis is https://github.com/kornai/4lang/blob/1d19f167b9c0eace5bd874759860781be78f96ed/4lang.

5.7 CONCLUSION

We described how deep cases can work in a machine comprehension resource that assigns deep cases directly to rather abstract language-independent concepts. We have manually created the `4lang` definitions of the elements of a defining vocabulary. We labeled the locations of the arguments with "deep cases", thematic-role-sytle language-independent syntacto-semantic generalizations, thus proposing a deep case inventory.

Clearly, the most important deep cases are agent and patient. In the next chapter we analyze the representation of these relations with the tools of tensor decomposition.

Keywords: language resource, syntactic analysis, verb structures, Mazsola, size

— Sass (2015)

6

DECOMPOSING A TRANSITIVE VERB TENSOR

6.1 INTRODUCTION

In the previous chapter, we investigated verb argument roles in a semantic network, and the last two chapters will analyze word embeddings. This chapter connects the two topics by word embedding experiments in the verb structure domain.

Verbs have been characterized on the basis of how frequently various syntactic constituents occur in various grammatical relations to them, which is, not surprisingly, related to the meaning of the verb (Levin 1993). These selectional preferences have been analyzed with machine learning tools (Van de Cruys 2009). Verb structures include collocations, whose syntactic modifiability or semantic compositionality is reduced: their linguistic distribution may be idiosyncratic or the sense of the combination may be habitual or even fixed (Bouma 2009).

Tensors (>2-dimensional arrays) generalize matrices; while matrices contain numbers aligned in two dimensions, rows and columns, tensors

have more of these dimensions, also called *axes* or *modes*.[1] The Singular value decomposition (SVD) of a co-occurrence matrix is a natural tool to compute generalizations about the interactions between two modes, like words and documents (LSA, Landauer and Dumais (1997), Section 4.1.3), target and context words (words embeddings, Mikolov, Sutskever, et al. (2013), Levy and Goldberg (2014c), and Pennington, Socher, and Manning (2014)), or words and dependency contexts (Levy and Goldberg 2014a). Four ways of looking at SVD (in LSA) can be distinguished (Turney and Pantel 2010): the goal can be the modeling of some latent meaning, noise reduction, indirect aka high-order co-occurrences (when two words appear in similar contexts), or data sparsity reduction. Intuitively, language features multi-mode interactions: *the turntable playing the piano* is strange (Van de Cruys 2009), while the two-mode relations ⟨play, SUBJ, turntable⟩ and ⟨play, OBJ, piano⟩ are perfect. Tensor generalizations of matrix decomposition (Kolda and Bader 2009), especially *low-rank factorizations*, open the way for the analysis of such interactions.

It seems that, after intensive early research (Van de Cruys 2009; Van de Cruys, Poibeau, and Korhonen 2013; Polajnar, Rimell, and Clark 2014; Fried, Polajnar, and Clark 2015; Hashimoto and Tsuruoka 2015), results obtained with skip-gram and related word embedding methods outshone tensor methods for verb argument structure. Yet tensor decomposition remains relevant, as it is more interpretable than more recent methods, and it has developed remarkably. NLP test-beds in the domain of verb argument structure have been involved in cutting-edge scalable, noise-robust tensor works (Sharan and Valiant 2017; Bailey, Meyer, and Aeron 2018; Frandsen and Ge 2019). The data-driven linguistic understanding of word ambiguity and especially that of verb selection is still immature. Here we try to make progress in the linguistic direction by further research on the tensorial analysis of verb argument structure.

Tensor decomposition provides embedding vectors for each mode (in our case, nouns as subjects, verb, and nouns as objects) analogous to word embeddings in (shallow or deep) neural networks. In this paper, we compute different association measures between subjects, verbs, and objects, populate tensors with these measures, decompose the tensors with different algorithms, and investigate the resulting word embeddings quantitatively and qualitatively to answer the following questions. Our first four questions will be answered quantitatively in the modeling of English subject-verb-object triple similarity, while the last two questions are qualitative.

1. Which *association measure* yields the best representations? We experiment with several measures, including our novel generaliza-

1 The term *mode* is preferred when data from different modalities are fused.

tion of normalized pointwise mutual information to the higher-order (>2) case.

2. Should we include *empty argument fillers* (existentially bound subjects or objects) in our co-occurrence statistics? Ideally, including them may help generalization over the transitive and the intransitive uses of the same verb, while discarding them may help focusing on transitive structures cleanly as a separate phenomenon.

3. The two tensor decomposition algorithms, CPD and Tucker, which we will introduce in Section 6.3, have very different time-complexity: Tucker is much faster. Tensor decomposition has hyper-parameters like the decomposition rank and the frequency cutoff. Both have an effect on the memory need, especially the latter. It would be beneficial, *if the two algorithms reached the best results with similar hyper-parameters,* because then a fast parameter tuning with Tucker would also benefit CPD. Is this the case?

4. Do latent dimensions of our word embeddings reflect lexical knowledge?

5. Can the difference between each noun as a subject versus an object correspond to some intuitive difference between subjecthood and objecthood?

Section 6.2 describes the linguistically motivated association measures between subjects, verbs, and objects we apply. These measures include ones that are novel to the best of our knowledge. Section 6.3 offers an introduction to tensor decomposition. Finally, 6.4 to 6.6 describe our experiments, originally published in Makrai (2022).[2] Our code is available online.[3]

6.2 COUNTS, WEIGHTING, AND ASSOCIATIONS

Word co-occurrences form *sparse* arrays, as most words do not occur empirically with most words, and frequencies span many orders of magnitude (*Zipf* or power-law distribution, Manin (2008) and Gittens, Achlioptas, and Mahoney (2017)). Sparsity is desiderable for both cognitive/linguistic and computational reasons. In computational terms, sparsity can be regarded a way of regularization or simply a trick to fit in memory. Whatever the main motivation is, in a data-driven scenario, linguistic tensor decomposition methods have to be based on

2 The PhD candidate is grateful to Tülay Adalı, the enthusiastic lecturer at DeepLearn Summer University 2018, who drew his attention to the potentials of tensor decomposition, and to Gábor Berend, Gábor Borbély, Balázs Indig, Ágnes Kalivoda, András Kornai, Eszter Simon, Tibor Szécsényi, and anonymous reviewers for their helpful comments.

3 https://github.com/makrai/verb-tensor

sparse tensors populated with possibly more sophisticated scores than frequency. Now we turn to these weighting functions and especially to linguistically motivated association scores.

The simplest and most popular (Pennington, Socher, and Manning 2014; Sharan and Valiant 2017) choice is the logarithm of the co-occurrence frequency, $\log f(x, y, z)$. Jenatton et al. (2012) place the modeling of the $\langle \text{subject}, \text{verb}, \text{object} \rangle$ triples in the context of multi-relational learning, and apply a weighting function related to the log-bilinear model (Mnih and G. Hinton 2007; Mikolov, Chen, et al. 2013), see Table 11.

Van de Cruys (2009, 2011) and Van de Cruys, Poibeau, and Korhonen (2013), and Bailey, Meyer, and Aeron (2018) use three-mode generalizations of the information-theoretic association measure *(Positive) Pointwise Mutual Information* ((P)PMI). Positivity is related to sparse inputs: in order to attribute higher scores to actual co-occurrences than unattested ones, in the case of PMI and the lexicographic association scores introduced in the following paragraph, *positive* variants of the association measures have to be used, e.g. PPMI, which replaces negative PMI entries with zero. We discuss the two types of three-variable generalization of PPMI in Section 6.2.1: the more standard total correlation (that we still call PMI) and interaction information.

We also experiment with generalizing Log Dice (Rychlý 2008) to three axes

$$\log \frac{3f(x, y, z)}{f(x) + f(y) + f(z)} + c, \tag{1}$$

where c is chosen so that the Log-Dice values are non-negative. (While 3 in the nominator is redundant, because it is subsumed under c, we keep it in the formula to make it more reminiscent of the established 2-variable case.) The use of Log Dice as well as salience introduced in the next paragraph has, to the best of our knowledge, mainly been limited so far to lexicography.

6.2.1 *Higher-order PMI*

One would think that it's obvious that the 3-variable generalization of Pointwise Mutual Information (PMI) is

$$\log \frac{p(x, y, z)}{p(x)p(y)p(z)}, \tag{2}$$

but it turns out that this is only one of the possible generalizations. Van de Cruys (2011) introduces two pointwise association measures, whose expected values are two different multivariate generalizations of mutual information (Shannon and Weaver 1949): interaction information (McGill 1954) and total correlation (Watanabe 1960).

	corpus	shape	weighting, postprocesssing	rank
Van de Cruys (2011)	Dutch .5 B	10 K subjects × 1 K verbs × 10 K direct objects	PPMI	50 … 300
Van de Cruys (2011)	Dutch .5 B	10 K subjects × 1 K verbs × 10 K direct objects	2 variants of PMI	(no decomp)
Van de Cruys, Poibeau, and Korhonen (2013)	UKWaC 2 B	10 K subjects × 1 K verbs × 10 K objects	PMI	300
Jenatton et al. (2012)	2 M Wp articles	30 K subjects × 5 K verbs × 30 K direct objects	$\mathbb{P} = 1/(1 + \exp(-\mathbf{s}_i \cdot \mathbf{R}_j \otimes \mathbf{o}_k))$ + refinement	25, 50, 100
Sharan and Valiant (2017)	Wikipedia 1.5 B	10 K words × 10 K words × 10 K words	$\log(f + 1)$, $w_i = s_i \oplus v_i \oplus o_i$ normalized	100
Bailey, Meyer, and Aeron (2018)	.3 B from Wp	word freq cut-off = 1 000	($\pm$shifted) PPMI, w_i normalized	300

Table 11: NLP-oriented tensor decomposition work. Corpus sizes are shown in billion words. In the formulae, f denotes co-occurrence frequency.

Pointwise *interaction information* is based on the notion of conditional mutual information.[4]

$$\log \frac{p(x,y)p(x,z)p(y,z)}{p(x,y,z)p(x)p(y)p(z)} \tag{3}$$

Total correlation on the other hand quantifies the amount of information that is shared among the variables, with a pointwise variant defined by the formula in Equation (2). Following the literature (Villada Moirón 2005; Van de Cruys 2009; Van de Cruys, Poibeau, and Korhonen 2013; Bailey, Meyer, and Aeron 2018), when we speak about *(multivariate Positive) Pointwise Mutual Information* in this paper, we will mean (pointwise) total correlation.

Van de Cruys (2011) reports that in their Dutch experiments both methods are able to extract salient subject verb object triples (prototypical SVO combinations like *poll represents opinion* and fixed expressions). Narrowing the scope to the word *play*, they find that interaction information picks up on prototypical SVO combos, e.g. *orchestra plays symphony*, while the more established one (which he calls specific correlation) picks up on *play a role* and salient subjects that go with the expression.

6.2.2 *Salience and normalized PPMI*

PPMI, despite of its nice information-theoretic interpretability, is biased towards rare events (Turney and Pantel 2010; Levy et al. 2015; Zhuang et al. 2018). This motivates the Sketch Engine lexicographic software (Kilgarriff et al. 2004) to multiply vanilla (two-order) PPMI by log-frequency, to get the measure of *salience*. We apply similar modifications to every score introduced in Section 6.2 so far. We denote vanilla PPMI (Equation (2)), interaction information (Equation (3)) and Log Dice (Equation (1)) by `pmi-vanl`, `iact-vanl`, and `Dice-vanl`, respectively, and define `pmi-sali`, `iact-sali`, and `Dice-sali` as the vanilla score multiplied by $\log f(x,y,z)$.

There is a theoretically better motivated way of transforming PMI to some measure which is less biased towards rare combinations. In Bouma (2009)'s approach, *normalization* is related to boundedness. He looks for measures whose absolute value is pointwise larger than that of PMI. Entropy and negative log probability are two of those measures. The corresponding normalized measures are called *normalized mutual information (NMI)* and *normalized pointwise mutual information (NPMI)*, respectively. Both are used in the literature, e.g. the review by Sra (2018) highlights NMI, while Balogh et al. (2020) opt for NPMI, and

4 Mnemonically, the formula of the pointwise variant generalizes the 2-mode case along the inclusion and exclusion principle, except it has the numerator and the denominator swapped to ensure a proper set-theoretic measure.

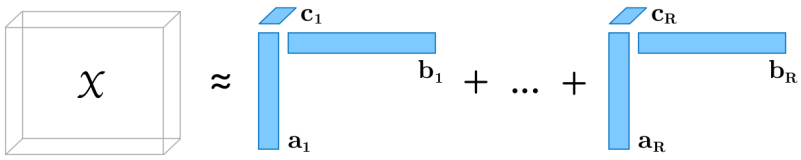

Figure 14: Canonical Polyadic Decomposition, figure from Rabanser, Shchur, and Günnemann (2017).

so do we. In our experiments, we apply this normalization to the two multi-mode generalizations of PMI, `pmi-vanl` and `iact-vanl`.

While normalized interaction information (`niact`) does not excel in our experiments, `pmi-sali` proves to be the best among the alternatives. This best result is obtained with non-negative CPD. Our best general (i.e. possibly negative) decomposition, both CPD and Tucker is obtained with tree-variable normalized PMI, i.e.

$$\frac{\log \frac{p(x,y,z)}{p(x)p(y)p(z)}}{-\log f(x,y,z)},$$

which we call `npmi` in the tables. These two measures are to the best of our knowledge the novelties of the present thesis. Empirically, when divided by $-\log p(x,y,z)$, positive interaction information and the more standard 3-mode PPMI is upper-bounded by 1 and 2, respectively.

6.3 TENSOR DECOMPOSITION

The main entry point to tensor computation is Kolda and Bader (2009), but Rabanser, Shchur, and Günnemann (2017) is also worth consulting.

There is no single generalization of the SVD concept: the two most popular extensions, Canonical Polyadic Decomposition and the more general Tucker, feature different generalized properties. Sidiropoulos et al. (2017) discuss the interpretation of these two different ways of decomposition in signal processing and machine learning points of view.

6.3.1 *Canonical Polyadic Decomposition*

Canonical Polyadic Decomposition (CPD, aka CanDecomp, `Para`llel `Fac`tor model, ParaFac, rank decomposition, or Kruskal decomposition, (Carroll and Chang 1970)) expresses a tensor as a minimum-length linear combination of rank-1 tensors. A rank-1 tensors is the tensor product of a collection of vectors, just as the dyadic product of two vectors is a 1-rank matrix, see Figure 14.

The alternating least squares algorithm (ALS, Carroll and Chang (1970) and Harshman (1970)) is an iterative method for CPD. In each iteration, all but one of the modes are fixed and the remaining one is fitted. ALS does not guarantee convergence, and even if it converges,

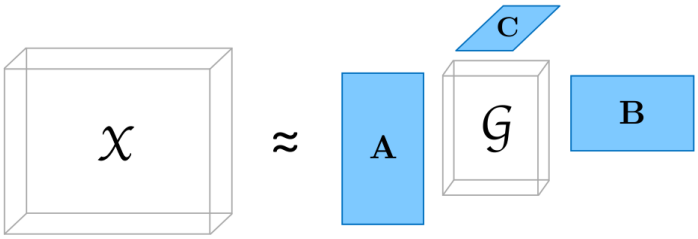

Figure 15: Tucker Decomposition, figure from Rabanser, Shchur, and Günnemann (2017).

this cannot be detected in a trivial way. Orth-ALS (Sharan and Valiant 2017) improves on ALS.

6.3.2 *Tucker decomposition*

While CPD is more popular in the computational linguistic literature, and its better parameter efficiency lends it better explanatory adequacy, we also experimented with Tucker decomposition, because it can be computed much more efficiently. Tucker decomposition (aka Higher Order SVD, Tucker (1966)) factorizes a tensor into a core tensor $\mathcal{G}$ multiplied by a matrix along each mode, see Figure 15. In the case of

$$\text{subject} \times \text{verb} \times \text{object}$$

tensors, rows of the three matrices contain embedding vectors of entities (subjects or objects) and those of verbs ("relation"), and entries of the core tensor $\mathcal{G}$ determine the levels of interactions between the latent dimensions. Tucker decomposition is not unique, because we can transform $\mathcal{G}$ without affecting the fit if we apply the inverse of that transformation to the factor matrices. Uniqueness can be improved (Kolda and Bader 2009) by imposing e.g. sparsity, making the elements small, or making the core "all-orthogonal". Other priors and constraints in tensor learning involve non-negativity and independence (Lahat, Adali, and Jutten 2015).

6.4 EXPERIMENTS

In this section, we report our experiments. After the introduction (Section 6.4.1) of the corpus that serves as the basis of our empirical investigations, Section 6.4.2 compares association measures, the two alternatives for treating missing arguments, the two decomposition algorithms, and some other hyper-parameters (the decomposition rank and the frequency cutoff) in the classical task of predicting the similarity of English subject-verb-object triples (Kartsaklis and Sadrzadeh 2014). Then in Section 6.4.3, we investigate the latent dimensions qualitatively. Section 6.4.4 compares the embedding vector of each noun as a subject versus an object, to see how differently nouns behave in the two roles.

cutoff	shape with unfilled	shape without unfilled	SVO coverage
1	(324 196, 90 606, 287 967)	(206 488, 41 075, 188 619)	1.00
10	(160 629, 37 427, 129 694)	(109 432, 19 824, 92 635)	1.00
100	(92 999, 20 937, 69 536)	(71 768, 13 907, 57 420)	1.00
1000	(44 168, 10 444, 32 359)	(40 309, 8 838, 30 280)	1.00
10000	(13 765, 5 070, 12 313)	(13 610, 4 895, 12 115)	0.97
100000	(3 474, 2 313, 4 120)	(3 463, 2 308, 4 108)	0.86
1000000	(546, 814, 981)	(545, 813, 980)	0.58
10000000	(36, 194, 87)	(35, 194, 86)	0.06

Table 12: The length of each axis, i.e. the number of subjects, verbs, and objects, at different frequency cutoffs.

6.4.1 *Experimental setting: the corpus and the similarity task*

In our experiments, we took the occurrence counts of ⟨subject, verb[5], direct object⟩ triples from the automatically dependency-parsed (Nivre et al. 2016) English corpus DepCC (Panchenko et al. 2018), irrespectively of whether there were other arguments or adjuncts. Regarding empty fillers, we investigated two alternatives: including them (represented by a fixed string) or discarding them from our statistics. `tensorly` (Kossaifi et al. 2016) was used for CPD and (general and non-negative) Tucker decomposition of tensors. For tensor population in `COO`rdinate format, we use the `sparse` Python library.

Our quantitative tests are based on a classical similarity data-set for English transitive verb structures (SVO triples) by Kartsaklis and Sadrzadeh (2014, KS14). We discussed shortcomings of this task (Section 4.2.9), we still assume it is sufficient for the present purposes. The data-set contains triples with gold (human) similarity scores. We represent SVO triples by concatenating the corresponding subject, verb, and object embedding vector and computed the Spearman correlation between the cosine similarities of the (long) vectors in each pair with the human scores.

Normalizing the vectors to unit length benefits some tasks: see Sections 4.1.6, 4.2.8.3 and 4.3.2 and especially our experiments in Section 8.4.2. The intuition behind normalization is that vector length is related to word frequency, and words with quite different frequency may have similar meaning. Motivated by this, we also experimented with normalizing the vectors. However, this did not lead to better results, similarly to Section 8.4.2.

5 *Verb* means, in Universal Dependencies terms, that the `upos` starts with `VB`.

6.4.2 *Quantitative results in transitive structure similarity*

We populated tensors with the association measures introduced in Section 6.2. The statistics were based on either including empty argument fillers (i.e. treating all arguments "optional") or excluding these occurrences. We took different cutoffs and computed non-negative or general CPD or Tucker decompositions in different ranks. Out-of-vocabulary words are represented by an all-0 vector. Table 12 shows the length of each axis, i.e. the number of subjects, verbs, and objects, at different frequency cutoffs. The last column shows what ratio of the SVO pairs is intact in the sense that all the 2×3 words are covered in the corresponding embedding. The reader may object that these cutoffs are very strict restrictions, compared to word2vec-based models where a few dozen occurrences result in perfectly usable representations. Nevertheless, it has to be borne in mind that the original motivation for using a cutoff in tensor decomposition is not to have enough samples, but to fit in the memory.

Correlations we obtain in the subject-verb-object task are shown in Table 13. The properties of the original sparse tensor (the association measure, the option whether empty fillers are included, and the frequency cutoff) are show on the left of the vertical line, while those of the decompositions (non-negative or general CPD or Tucker decompositions, and the rank of the decomposition) are shown on the right. The table shows, in addition to the best setting, each setting obtained by changing one hyper-parameter. (E.g. the second and the third entries differ from the best one only in the decompositions rank: the rank of the second one is double of the best rank, while that of the third one is the half of the best value.) The best result is obtained by non-negative CPD. The horizontal lines shows where our best general Tucker, general CPD, and non-negative Tucker decompositions – which will be shown in separate tables, Tables 14 and 15, to keep this one manageable – end up. In Tucker decompositions, we use the same rank among all axes.

We obtained the best correlation, 0.7359, from the decomposition of a tensor populated with salience-weighted PMI values, including empty fillers, and setting the frequency cutoff to 1 million, i.e. restricting the axes of the tensor to the subjects, verbs, and objects that appear at least 1 million times. This best correlation was obtained with non-negative CPD in rank 64. This correlation value is in the same range as the 0.76 Hashimoto et al. (2014) obtained with a much more complex system. Hashimoto et al.'s system used to be the state of the art, when this task was fashionable.

The table shows the correlation obtained by changing each (meta)-parameter. While the results seem to be relatively robust with respect to the decompositions *rank*, it may be interesting that when we concatenate the subject, the verb, and the object embedding vectors, 64 dimensional each, we get a vector in the famous range of a couple of

assoc measure	unfilled	cutoff	non-negative	decomp algo	rank	corr
pmi-sali	included	1 000 000	non-neg	CPD	64	0.7359
pmi-sali	included	1 000 000	non-neg	CPD	128	0.7097
pmi	included	1 000 000	non-neg	CPD	64	0.6857
pmi-sali	included	1 000 000	non-neg	CPD	32	0.6773
pmi-sali	included	*300 000*	non-neg	CPD	64	0.6630
npmi	included	1 000 000	non-neg	CPD	64	0.6602
dice-sali	included	1 000 000	non-neg	CPD	64	0.4709
pmi-sali	*excluded*	1 000 000	non-neg	CPD	64	0.4578
pmi-sali	included	1 000 000	*general*	CPD	64	0.4560
ldice	included	1 000 000	non-neg	CPD	64	0.4409
log-freq	included	1 000 000	non-neg	CPD	64	0.4322
iact-sali	included	1 000 000	non-neg	CPD	64	0.4112
niact	included	1 000 000	non-neg	CPD	64	0.4068
pmi-sali	included	*3 000 000*	non-neg	CPD	64	0.3936
iact	included	1 000 000	non-neg	CPD	64	0.3248
pmi-sali	included	1 000 000	non-neg	*tucker*	64	0.2989

Table 13: Quantitative results: correlations in the subject-verb-object triple similarity task (Kartsaklis and Sadrzadeh 2014) obtained with word embeddings of tensor decompositions.

hundreds of dimensions, which proved to work well in many different scenarios like LSA and static word embeddings.

As for our *association measures,* different weighted variants (salience, vanilla, or normalization) of PMI work the best, followed by log-Dice and log frequency. Variants of interaction information performs the worst.

The inclusion of empty fillers, the frequency cutoff, and the decomposition rank are all related to the *size of the tensors*. While we have already seen that the decomposition rank does not have a great influence on the results, if we exclude empty fillers, a more generous frequency cutoff may theoretically lead to better results than if we change only one of these two parameters. It turns out, that we can indeed get relatively good result (0.694181) this way, but with general Tucker decomposition (instead of non-negative CPD) and log-Dice (instead of salience-weighted). The cutoff is 1 million.

Non-negative decomposition is advantageous from the interpretational point of view, because in our experiments, they resulted in embedding matrices which are *sparse* in the broad sense that most coordinates are low. Figure 16 shows a histogram of the matrix elements. Note that the vertical axis, which corresponds to the histogram count in each bin, is logarithmic. The figure suggests that frequency decreases faster than exponentially as larger weights are considered. The good performance of non-negative CPD suggests that non-negativity introduces meaningful structure. Sparsity raises the hope that coordinates are in-

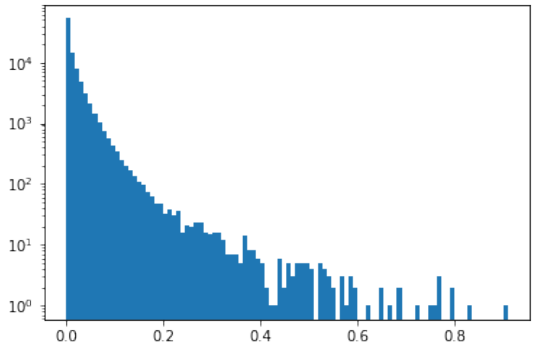

Figure 16: The histogram of the verb embedding matrix elements. Note that the vertical axis, which corresponds to the histogram count in each bin, is logarithmic. The figure suggests that frequency decreases faster than exponentially as larger weights are considered.

terpretable, i.e. they correspond to concepts or properties. We will see in Section 6.4.3 that they really do.

CPD has the advantage that it maps the *modes in the same space*. In our case, this is the most interesting for subjects and objects: we can compare the same noun in the two roles. We return to this in Section 6.4.4.

While our best results have been obtained with non-negative CPD, we discuss general Tucker and CPD and non-negative Tucker as well. Results with general decompositions and non-negative Tucker are shown in Table 14 and Table 15, respectively. General Tucker and CPD and non-negative Tucker all prefer normalized PMI as the association measure, disfavor interaction information, and results with log frequency and log Dice vary. General and non-negative Tucker obtains the best results with the same rank as non-negative CPD, and the two non-negative decomposition algorithms also share the value for a best cut-off. It is inconclusive whether it is advantageous to include occurrences with unfilled arguments in our statistics.

6.4.3 *Qualitative analysis of latent dimensions*

Now we investigate the latent dimensions obtained by tensor decomposition. We experimented with non-negative and general CPD and Tucker decomposition with the respective hyper-parameters that reached the best result in the SVO-similarity task.

The latent dimensions are shown in Tables 16 to 18. (Dimensions with general Tucker are degenerate, and they are omitted to save space.) Each line corresponds to a latent dimension. Dimensions are illustrated by the words with the greatest coordinates in the dimension. Blocks represent dimension triples. ∅ denotes that the corresponding grammatical function is unfilled. Some latent dimensions, like the first one in our

assoc measure	unfilled	cutoff	rank	correlation
npmi	included	100 000	64	0.7191
pmi-sali	included	100 000	64	0.7049
log-freq	included	100 000	64	0.6883
pmi	included	100 000	64	0.6759
npmi	included	30 000	64	0.6729
ldice	included	100 000	64	0.6685
ldice-sali	included	100 000	64	0.6666
npmi	included	300 000	64	0.6598
npmi	included	100 000	128	0.6540
npmi	included	100 000	32	0.6042
npmi	excluded	100 000	64	0.5207
iact-sali	included	100 000	64	0.5059
niact	included	100 000	64	0.4632
iact	included	100 000	64	0.4316

assoc measure	unfilled	cutoff	rank	correlation
npmi	excluded	300 000	256	0.6383
pmi-sali	excluded	300 000	256	0.6166
pmi	excluded	300 000	256	0.5811
npmi	excluded	1 000 000	256	0.5754
npmi	excluded	100 000	256	0.5713
npmi	excluded	300 000	512	0.5677
npmi	excluded	300 000	128	0.5290
npmi	excluded	30 000	256	0.5239
npmi	included	300 000	256	0.5070
log-freq	excluded	300 000	256	0.2465
ldice	excluded	300 000	256	0.2093
iact-sali	excluded	300 000	256	0.1280
niact	excluded	300 000	256	0.0726
iact	excluded	300 000	256	0.0615

Table 14: Results with general Tucker (top) and general CPD (bottom).

assoc measure	unfilled	cutoff	rank	correlation
npmi	excluded	1 000 000	64	0.5186
npmi	excluded	1 000 000	128	0.5102
npmi	excluded	300 000	64	0.4814
pmi	excluded	1 000 000	64	0.4563
pmi-sali	excluded	1 000 000	64	0.4387
npmi	excluded	1 000 000	32	0.3753
npmi	excluded	3000 000	64	0.3366
npmi	optional	1 000 000	64	0.2889
iact	excluded	1 000 000	64	0.0989
log-freq	excluded	1 000 000	64	0.0763
ldice	excluded	1 000 000	64	0.0698
ldice-sali	excluded	1 000 000	64	0.0619
niact	excluded	1 000 000	64	0.0454
iact-sali	excluded	1 000 000	64	0.0064

Table 15: Results with non-negative Tucker.

non-negative CPD are dominated by (the empty filler and) pronouns. In these cases we *emphasize* the first contentful filler. `-rrb-` stands for right round brackets, and its appearance may be an artifact of the corpus (i.e. parsing errors).

In the case of CPD, the dimensions are enumerated in the order as returned by the algorithm. With Tucker, the values g_{ijk} in the core tensor $\mathcal{G}$ represent the interaction between the ith latent dimension for subjects, the jth one for verbs, and the kth one for objects. We sorted the triples of SVO latent dimensions in our best non-negative and general Tucker decomposition by this interaction strength. The index of each dimension, as returned by the algorithm, is also shown in the table. E.g. the first block in non-negative Tucker shows that the strongest interaction is between the 5th latent dimension of subjects, the 10th one for verbs, and the 7th one for objects. Note that in the non-negative case, $g_{ijk} \geqslant 0$, so we do not have to take the absolute value. Dimensions obtained with the two *non-negative algorithms* seem semantically interpretable, while those from general decomposition are less convincing.

6.4.4 *Comparing subject and object vectors*

Tensor decomposition can shed light on how differently nouns behave as subjects and as objects. This question is related to symmetric factorization (Bailey, Meyer, and Aeron 2018), which imposes symmetry

dim	words
0	∅, that, which, it, *story*, he, they, who, what, one, she, work, event, -rrb-, this, you...
0	catch, attract, draw, pay, deserve, capture, gain, grab, get, receive, focus, require,...
0	attention, eye, crowd, interest, fire, visitor, audience, conclusion, breath, people, ...
1	∅, who, we, he, I, you, she, they, -rrb-, *student*, member, people, group, Center, parti...
1	attend, host, hold, organize, schedule, enjoy, join, arrange, cancel, miss, watch, pla...
1	meeting, event, conference, session, party, show, school, class, dinner, church, tour,...
2	that, which, it, this, ∅, *change*, factor, they, choice, condition, decision, issue, -rr...
2	affect, impact, influence, improve, hurt, reflect, benefit, change, damage, enhance, a...
2	ability, performance, health, outcome, life, quality, result, business, development, e...
3	file, which, page, site, that, it, book, report, section, document, collection, websit...
3	contain, include, provide, have, list, feature, display, show, comprise, present, give...
3	information, link, material, number, list, datum, name, content, statement, reference,...

Table 16: Latent dimensions with Non-negative CPD

dim	words
5	court, Court, judge, panel, official, we, he, it, authority, government, -rrb-, Board,...
10	reject, dismiss, deny, grant, hear, consider, decide, accept, throw, resolve, sustain,...
7	motion, appeal, claim, request, argument, case, challenge, application, complaint, att...
4	revenue, sale, share, price, stock, production, cost, rate, order, volume, number, fut...
3	rise, fall, increase, jump, drop, decline, climb, decrease, grow, gain, slip, represen...
1	percent, %, $, increase, point, most, rate, level, average, less, matter, value, cost,...
11	hotel, property, room, restaurant, home, Center, house, location, facility, House, are...
8	offer, boast, feature, have, provide, include, enjoy, serve, accommodate, occupy, prep...
9	room, pool, accommodation, access, facility, restaurant, variety, service, view, range...
6	board, Council, Board, Commission, Committee, member, committee, Congress, Court, cour...
2	approve, adopt, reject, pass, consider, review, endorse, propose, award, recommend, ac...
2	resolution, request, budget, plan, proposal, contract, change, application, project, i...

Table 17: Latent dimensions with Non-negative Tucker

dim	words
0	Israel, group, government, Foundation, Association, company, -rrb-, military, army, Cl...
0	launch, wage, suspend, mount, begin, run, fund, organize, sponsor, administer, carry, ...
0	campaign, attack, program, initiative, operation, strike, programme, website, effort, ...
1	user, you, application, customer, developer, visitor, client, processor, device, User,...
1	access, select, specify, upload, view, enter, edit, browse, click, create, retrieve, m...
1	file, datum, content, document, page, parameter, site, folder, node, Internet, informa...
2	device, assembly, means, structure, system, element, plate, section, interface, unit, ...
2	comprise, include, contain, have, utilize, employ, represent, say, mean, control, enab...
2	layer, element, device, tube, housing, spring, electrode, pump, plate, container, memb...
3	attorney, plaintiff, defendant, party, respondent, prosecutor, State, lawyer, governme...
3	file, receive, oppose, make, give, present, withdraw, handle, publish, drop, provide, ...
3	motion, notice, petition, appeal, response, answer, objection, charge, request, submis...

Table 18: Latent dimensions with General CPD

constraints between the embeddings of the same entities in different modes (in our case, between the embeddings of the same noun as a subject or an object). Our approach is complementary, based on that CPD maps nouns as subjects and objects in the same space.

In our experiments, we consider (non-negative) CPD decomposition with the hyper-parameters that proved best in English SVO-similarity. We computed the (unnormalized) dot product similarity between the subject and object vector of each noun, and sorted all the nouns by this similarity. The largest distance is found with *∅, he, she, they, I, device, system, that, you, it...*, while the most symmetric nouns are *doubt, reality, future, same, hope, feeling, mine, reason, consumer, plenty...* A possible explanation is that the former lemmas, especially personal pronouns (or their inflected forms), are arguably much more frequent in agentive roles than other nouns, while they are infrequent in patient roles. Words in the second group can be framed in language both as animate and as inanimate. *Future* or *hope* are not alive in the biological sense, but they are often attributed agentive roles (what can be called a metaphorical use of language, but being metaphorical does not mean that the usage is peripheral (Recski 2016b, Section 3.2)).

6.5 CONCLUSION OF THE MAIN EXPERIMENTS

Now we can answer the questions raised in Section 6.1:

1. Weighted variants of positive pointwise mutual information proved better than the considered alternatives in modeling subject-verb-object structure similarity.

2. It does not matter whether we include occurrences with unfilled arguments in our statistics. Our best results were obtained with non-negative CPD.

3. The best frequency cutoff and the decomposition rank is the same for the two non-negative decomposition algorithms, which raises the hope that these hyper-parameters of non-negative CPD can be fine-tuned based on the much faster non-negative Tucker.

4 and 5 Our experiments provided lexically interpretable latent dimensions, and our experiments with non-negative CPD suggest that the difference between subject and object embeddings can be related to animacy.

6.6 FOLLOW-UP

In this section, we report experiments, which did not appear in Makrai (2022).

# verbs	verbs
702	have, do, get, go, take, think, know, want, need, give, look, work, provide, try, ...
131	live, talk, stand, die, walk, wait, sit, stay, wonder, care, arrive, fly, gon, sleep, ...
86	kill, catch, trust, bear, email, marry, fuck, date, judge, bless, honor, forgive, beg,...
85	add, eat, produce, deliver, prepare, drink, spread, cook, burn, taste, wash, supply, ...
80	use, develop, manage, perform, complete, replace, install, connect, test, conduct, ...
80	let, reach, hit, cost, exceed, rate, approach, /, -lsb-_VBD, rank, -lsb-_VB, \, -lsb-_...
79	put, break, pull, throw, push, lay, stick, grab, touch, press, suck, kick, shake, ...
77	identify, commit, defend, repeat, expose, separate, dig, heal, dress, distinguish, ...
76	send, check, view, click, display, generate, update, access, search, store, delete, ...
65	leave, enter, visit, fill, explore, ride, clean, cross, surround, locate, clear, rent,...
59	be, come, start, happen, seem, begin, continue, appear, lead, end, occur, prove, ...
58	help, keep, bring, remind, hurt, strike, worry, blow, inspire, bother, surprise, suit,...
57	tell, ask, call, thank, please, join, contact, become, assist, hire, name, engage, ...
51	pay, spend, save, raise, determine, compare, charge, measure, adjust, predict, invest,...
46	make, see, find, love, like, hear, enjoy, remember, miss, guess, recommend, notice, ...
43	understand, discover, recognize, examine, evaluate, investigate, acknowledge, assess, ...
43	face, experience, address, fix, handle, suffer, solve, celebrate, resolve, mark, ...
39	receive, win, lose, earn, gain, extend, deserve, capture, retain, lack, exercise, ...
37	plan, fail, focus, vote, act, deal, attempt, rely, struggle, participate, benefit, ...

Table 19: Verb clusters obtained from our verb embedding vectors in an unsupervised fashion. The smallest cluster is omitted to save space.

6.6.1 *Clustering verb vectors*

Semantic *classes* of verbs like those in VerbNet (Section 2.4.5) may be induced by clustering verb embedding vectors. If clusters obtained in unsupervised fashion correspond to gold verb classes, ambiguous verbs like *play* mentioned in Section 6.1 may be detected as outliers from the clusters, as their uses are composed of occurrences corresponding to different clusters.

Our method for obtaining verb clusters consists of mapping verb embedding vectors to a lower dimensional space with UMAP (McInnes et al. 2018) and clustering them with HDBScan (McInnes, Healy, and Astels 2017), which is a hierarchical, density based clustering algorithm. Dimensionality reduction is needed because density makes little sense in hundreds of dimensions. Our choices of UMAP meta parameters are the following: We map verb embedding vectors to 16 or 32 dimensions (fine-tuned in a comparison to VerbNet, see later). In HDBScan, we set the number of neighbors to 30 and the minimum distance to 0, following the recommendations at `readthedocs`[6]. The metric in the ambient space (i.e. the original, high-dimensional one) is cosine. Minimum cluster size is 15 or 5, and the related parameter of `min_samples` is 5.

We compare non-negative and general CPD and Tucker decompositions. The parameters of the original tensor and its decompositions are set to the value with the best score in the SVO-similarity task. We

6 https://umap-learn.readthedocs.io/en/latest/clustering.html#umap-enhanced-clustering

preverb	verb			args			gloss
∅	bíz(ik)	NOM		-bAn 'in'			trust sth
(rá) 'onto'	bíz	NOM	ACC		-rA 'onto'		entrust sg to sy
meg PERFECT	bíz(ik)	NOM		-bAn 'in'			trust sy
meg PERFECT	bíz	NOM	ACC			INS	entrust sy with sg
el 'away'	bíz(za)	NOM	self-ACC				get conceited

Table 20: Argument structure variants of the Hungarian verb *bíz(ik)* based on Szécsényi (2019).

set one hyper-parameter of UMAP and HDBScan each, namely the dimension we map to and minimum cluster size, based on comparison to VerbNet classes.

In these computations we take VerbNet from the `nltk.corpus` package. In many cases, there are more class IDs associated to a verb. We take the first one, as returned by the corresponding function. Out-of-vocabulary verbs are treated as a separate class. We compare our clustering to VerbNet classes with adjusted rand score in scikit-learn (Pedregosa et al. 2011). We get the greatest score with non-negative Tucker (embeddings mapped to 16 dimensions, and minimum cluster size set to 15).

Table 19 shows the greatest clusters of English verbs. The greatest cluster, separated by a line in the table, is the one called `-1` in HDBScan. It contains points that "fall out" in the hierarchy as members of very small would-be clusters. The algorithm considers them outliers[7]. In our case, it seems that they are general verbs, especially those that we find in light verb constructions. The remaining clusters seem to be semantically coherent.

6.6.2 *Hungarian data and preverbs*

Finally, we mention pilot experiments in Hungarian, where two phenomena interfere with verb agument structure and ambiguity. Table 20, based on Szécsényi (2019), illustrates these with the verb *bíz(ik)* 'trust'. We can see that preverbs (verb particles, which can modify both the aspect and the meaning of a verb, Kalivoda (2021)) interfere with verb meaning, and the apparently incidental appearance of the suffix *-ik* (which can be argued to be related to unaccusativity) increases data sparsity. In our preliminary experiments, we built a *subject* × *preverb* × *verb* × *object* tensor from verb constructions in the data-base of the Mazsola verb argument browser (Sass 2015). In this earlier, unpublished phase of the project, we used CPD decomposition, solved by the Orth-ALS (Sharan and Valiant 2017) algorithm. For the future, we suggest introducing a mode for *-ik*. The "vocabulary" of this axis would

7 See https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html#

consist of only two choices: with or without *-ik*. The hypothesis is that this tensor would profit from denser data representation.

6.7 CONCLUSION

Tensor decompositions offers a direction orthogonal to the mainstream (Rogers, Kovaleva, and Rumshisky 2020) in the data-driven understanding of linguistic structure. We may want to learn semantic verb classes in an unsupervised fashion. If verb embedding vectors represent information like Levin's (1993) verb classes, ambiguous verbs could be identified in the form of outliers in the clustering. This line of research can be extended cross-lingually (Vulić, Mrkšić, and Korhonen 2017; Majewska et al. 2018; Sun et al. 2010).

We have generalized association measures to the higher-order case, and we show in the tensor decomposition modeling of English SVO triples that some are better than the existing alternatives. By exploring the hyper-parameters of the experiment, we have shown that the best better results (non-negative CPD and general Tucker) are obtained if include the occurrences when one of the arguments (typically the subject) is unfilled. The experiments gave lexically meaningful latent dimensions, and the non-negative CPD experiments qualitatively suggest that the difference between subject and object embeddings can be related to agentivity.

These experiments evidence that intransitive uses of verbs can be represent by the same item as their transitive use. This finely fits to our monosemic method in the previous chapter in the symbolic framework. In the remaining two chapters of the thesis, we investigate whether relations which intuitively hold between concepts can also be detected in data-driven distributional representations, more specifically, static word embeddings (word representations obtained with shallow neural networks).

Nekem szavakról szavak jutnak az eszembe és viszont.
'Words remind me of words and vice versa'

— Péter Esterházy

7

LEXICAL RELATIONS

The last two chapters in this thesis investigate how relations between words are represented in word embeddings. While classical lexical relations, such as hypernymy, antonymy, and causality, have long been studied in semantic networks, we expand our exploration to translation and word analogy as well.

Symbolic approaches rely on explicit, structured representations of relations. On the other hand, distributed word embeddings are neural network-based models that learn continuous vector representations from large corpora of text. These embeddings are capable of capturing

intricate semantic patterns, thereby offering a data-driven perspective on word relations. By systematically investigating these diverse word relations, we aim to gain a comprehensive understanding of how various types of semantic information are embedded within word representations. Our analysis seeks to shed light on the strengths and limitations of word embeddings by showing to what extent they represent different lexical relations in this broadened sense.

We start with lexical relations proper: hypernymy (what basic category a word belongs to, e.g. *dogs* are *animals*, Section 7.1), antonymy (opposite meaning, Section 7.2), and causality (Section 7.3). Then we broaden our focus to word analogies and translation (Sections 7.4 and 7.5).

This line of research is also related to semantic networks. In `4lang`, the semantic networks we introduced in Chapter 3, genus is formalized in by 0-edges like `dog` $\xrightarrow{0}$ `animal`, but much information is included in binary relations like `cow` $\xleftarrow{1}$ `make` $\xrightarrow{2}$ `milk`. The utility of word definitions depends on whether these binary relations capture the right pieces of information. Word embeddings can provide complementary information on whether a putative relation really exists.

As we already discussed in Chapter 4, the empirical support for both the syntactic properties and the meaning of a word form consists in the probabilities with that the word appears in different contexts. Contexts can be documents as in latent semantic analysis (LSA, Section 4.1.3) or other words appearing within a limited distance (window) from the word in focus. In these approaches, the corpus is represented by a matrix with rows corresponding to words and columns to contexts, with each cell containing the conditional probability of the given word in the given context. The matrix has to undergo some regularization to avoid overfitting. In LSA this is achieved by approximating the matrix as the product of special matrices.

In the last decade, deep neural networks have taken over the state of the art in many areas of artificial intelligence including vision (Krizhevsky and Sutskever 2012), speech processing (Dahl et al. 2011), and language (Peters et al. 2018), reducing the error in the respective tasks by a respectable factor. In language, the first wave of the revolution was word embeddings, word models learned by neural networks, which became very popular since Mikolov, Chen, et al. (2013) and Mikolov, Sutskever, et al. (2013). These more accurate variants of earlier VSMs map "similar" words to similar vectors in space of some hundred dimensions. Word similarity covers that in syntactic and semantic respect, and vector similarity is mostly measured by cosine similarity. In this chapter, we bulid on the finding of (Mikolov, Yih, and Zweig 2013) that embeddings reflect analogical relations – a.k.a. relational similarity (Levy and Goldberg 2014b) – like

woman − man ≈ queen − king

The first three sections investigate individual lexical relations with the tools of distributional modeling: hypernymy with sparse coding, antonymy with an embedding obtained by spectral clustering, and the geometry of causality. Our question remains whether relations which intuitively hold between concepts can also be detected in data-driven distributional representations (in the most cases, static word embeddings).

The main protagonist of Section 7.2 is the definition graph, which we already used for the analysis of the importance of each word as they define each other (Section 3.3). Theoretically, the same graph plays an important role in activation spreading, but this thesis does not make claims about the implementations of this process (the interested reader should consult Nemeskey et al. (2013)). In the follow section, it plays a third role: Makrai, Nemeskey, and Kornai (2013) used it to compute a word embedding, which we compared to some other embeddings which were famous at the time from the aspect of antonymy: we tested which subtype of antonymy is represented in each word embedding. We compared the embedding obtained from the definition graph to two word embeddings which were standard before the word2vec revolution: the Hierarchical Log-Bilinear Model (Section 4.2.2) and SENNA (Section 4.2.3). Our embeddings turned out to be more similar in this respect to variants of HLBL, (Mnih and G. E. Hinton 2009) than SENNA is – which suggests that our embedding was sound.

7.1 HYPERNYMY AS INTERACTION OF SPARSE ATTRIBUTES

The *distributional hypothesis* (Z. S. Harris 1954) says that a word can be described (in more computational terms, represented) based on how frequently it cooccurs with every other word. More specifically, the distributional *inclusion* hypothesis (Weeds and Weir 2003; Chang et al. 2018) says that hypernymy can be modeled based on that if *animal* is a hypernym of *dog*, *animal* will be grammatical in every context where *dog* is. It is less clear whether *animal* will appear in every context *at least as frequently* as *dog* does. Now we test this method for hypernym extraction with the tools of sparse coding.

Sparse vectors are vectors most of whose coordinates are zero, and non-zero coordinates ideally correspond to interpretable properties. It varies with models whether interpretability follows from the construction of the vectors, or the interpretation needs to be inferred from some latent structure. Even in the latter case, sparse representations tend to be more interpretable than less restricted ones. As far as sparse attributes (i.e. non-zero coordinates in *sparse word representations*) correspond to contexts, if follows from the distributional inclusion hypothesis discussed above that hypernymy should boil down to pointwise comparison. 'Dog' is an 'animal' if and only if it has all the properties

animals have, i.e. if all the non-zero coordinates of 'animal' are also non-zero for 'dog'.

This section originally appeared as Berend, Makrai, and Földiák (2018)[1], and describes 300-sparsans' participation in SemEval-2018 Task 9: *Hypernym Discovery*, with a system based on sparse coding and a formal concept hierarchy obtained from word embeddings. Our system took first place in subtasks *(1B) Italian* (*all* and *entities*), *(1C) Spanish entities*, and *(2B) music entities*.

7.1.1 *Introduction*

Natural language phenomena are extremely sparse by their nature, whereas continuous word embeddings employ dense representations of words. Turning these dense representations into a much sparser form can help in focusing on the most salient parts of word representations (Faruqui et al. 2015; Berend 2017; Subramanian et al. 2018).

Sparsity-based techniques often involve the coding of a large number of signals over the same dictionary (Rubinstein, Zibulevsky, and Elad 2008). Sparse, over-complete representations have been motivated in various domains as a way to increase separability, interpretability (Olshausen and Field 1997), and stability with respect to noise.

Non-negativity has also been argued to be advantageous for interpretability (Faruqui et al. 2015; Fyshe et al. 2015; Arora et al. 2016). As Subramanian et al. (2018) illustrates this in the language domain, where sparse features can be interpreted as lexical attributes, "to describe the city of Pittsburgh, one might talk about phenomena typical of the city, like erratic weather and large bridges. It is redundant and inefficient to list negative properties, like the absence of the Statue of Liberty".[2] Prior to our work, Berend (2018) utilized non-negative sparse coding for word translation by training sparse word vectors for the two languages such that coding bases correspond to each other.

Here we apply sparse feature pairs to hypernym extraction. The role of an attribute pair $\langle i, j \rangle \in \phi(q) \times \phi(h)$ (where q is the query word, h is the hypernym candidate, and $\phi(w)$ is the set of indices of non-zero components in the sparse representations of w) is similar to *interaction terms* in regression, what we will detail in Section 7.1.2.

Sparse representation is related to hypernymy in various natural ways. One of them is through *Formal concept Analysis (FCA)*. Cimiano, Hotho, and Staab (2005) already strove to acquire concept hierarchies from a text corpus with the tools of FCA. Our submissions experiment

1 Berend and Makrai worked together (both coding and writing the paper), but Berend's contribution is larger, say 2:1. Makrai created the poster and presented it. Földiák's contribution was focused on the FCA idea.

2 These representations are supposed to specify inherited default properties dierctly. E.g. the representation of a *sparrow* will contain, besides (being a) *bird*, (the capability to) *fly*. Exceptional subordinate concepts like *penguins* and *ostriches* will of course lack (the ability to) *fly*.

with an FCA tool by Endres, Földiák, and Priss (2010). We return in the next subsection to a description of formal concept lattices, and how hypernyms can be found in them.

Another natural formulation is related to *hierarchical sparse coding* (Zhao, Rocha, and Yu 2009), where trees describe the order in which variables "enter the model" (i.e. take non-zero values). A node may take a non-zero value only if its ancestors also do: the dimensions that correspond to top level nodes should focus on "general" meaning components that are present in most words. Yogatama et al. (2015) offer an implementation that is efficient for gigaword corpora. Exploiting the correspondence between the variable tree and the hypernym hierarchy offers itself as a natural choice.

The task (Camacho-Collados et al. 2018) evaluated systems on their ability to extract hypernyms for query words in five subtasks (three languages, English, Italian, and Spanish, and two domains, medical and music). Queries were categorized as either concepts or as entities. Results were reported for each category separately as well as in combined form, thus resulting in 5×3 combinations. Our system took first place in subtasks *(1B) Italian* (*all* and *entities*), *(1C) Spanish entities*, and *(2B) music entities*. Detailed results for our system appear in Section 7.1.3. Our source code is available online[3].

7.1.1.1 *Formal concept analysis*

Formal concept Analysis (FCA) is the mathematization of *concept* and conceptual hierarchy (Ganter and Wille 2012; Endres, Földiák, and Priss 2010). In FCA terminology, a *context* is a set of *objects* $\mathcal{O}$, a set of *attributes* $\mathcal{A}$, and a binary incidence relation $\mathcal{I} \subseteq \mathcal{O} \times \mathcal{A}$ between members of $\mathcal{O}$ and $\mathcal{A}$. In our application, $\mathcal{I}$ associates a word $w \in \mathcal{O}$ to the indices of its non-zero sparse coding coordinates $i \in \mathcal{A}$. FCA finds formal *concepts*, pairs $\langle O, A \rangle$ of object sets and attribute sets $(O \subseteq \mathcal{O}, A \subseteq \mathcal{A})$ such that A consists of the shared attributes of objects in O (and no more), and O consists of the objects in $\mathcal{O}$ that have all the attributes in A (and no more). (There is a closure-operator related to each FCA context, for which O and A are closed sets if and only if $\langle O, A \rangle$ is a concept.) O is called the extent and A is the intent of the concept.[4]

3 https://github.com/begab/fca_hypernymy

4 Those who are familiar with closure operators may note the following. We can define the prime operator $'$ both for objects and attributes in a dual way: O' is defined as the set $\{a \in \mathcal{A} \mid \forall o \in O, \langle o, a \rangle \in \mathcal{I}\}$, i.e. that of the shared attributes of objects in O, and A' as $\{o \in \mathcal{O} \mid \forall a \in A, \langle o, a \rangle \in \mathcal{I}\}$ i.e. the set of the objects in $\mathcal{O}$ that have all the attributes in A. Then the double application of $'$ is a closure operation both on objects and attributes: with notation $\overline{S} = S''$, for either $S \subseteq \mathcal{O}$ or $S \subseteq \mathcal{A}$, we have $S \subseteq \overline{S}$ and $\overline{\overline{S}} = S$, and the following conditions are equivalent for all $O \subseteq \mathcal{O}$ and $A \subseteq \mathcal{A}$:

- $\langle O, A \rangle$ is a concept
- O is a closed set with respect to $O \mapsto \overline{O}$, and $A = O'$

There is an order defined in the context: if $\langle A_1, B_1 \rangle$ and $\langle A_2, B_2 \rangle$ are concepts in C, $\langle A_1, B_1 \rangle$ is a *subconcept* of $\langle A_2, B_2 \rangle$ if $A_1 \subseteq A_2$ which is equivalent to $B_1 \supseteq B_2$. The concept order forms a so called complete lattice. The smallest concept whose extent contains a word is said to *introduce* the object. If $n(w)$ denotes the node in the concept lattice that introduces w, we expect that h will be a hypernym of q if and only if $n(q) \leqslant n(h)$.

The closedness of extents and intents has an important structural consequence. Adding attributes to $\mathcal{A}$ (e.g. responses of additional neurons) will very probably grow the model. However, the original concepts will be embedded as a substructure in the larger lattice, with their ordering relationships preserved.

7.1.2 *Our approach*

Here we describe our system that is based on sparse non-negative word representations and FCA besides more traditional features.

We use the popular skip-gram (SG) approach (Mikolov, Chen, et al. 2013) to train $d = 100$ dimensional dense distributed word representations for each subcorpus. The word embeddings are trained over the text corpora provided by the shared task organizers with the default training parameters of word2vec (w2v), i.e. a window size of 10 and 25 negative samples for each positive context.

We derived *multi-token units* by relying on the `word2phrase` software accompanying the w2v toolkit. An additional source for identifying multi-token units in the training corpora was the list of potential hypernyms released for each subtask by the organizers.

Given the dense embedding matrix $W_x \in \mathbb{R}^{d \times |V_x|}$, for some subcorpus of the shared task $x \in \{1A, 1B, 1C, 2A, 2B\}$, where $|V_x|$ is the size of the vocabulary and d is set to 100. As a subsequent step, we turn W_x into *sparse word vectors* akin to Berend (2017) by solving for

$$\min_{D \in \mathcal{C}, \alpha \in \mathbb{R}_{\geqslant 0}} \|D\alpha - W_x\|_F + \lambda \|\alpha\|_1, \tag{4}$$

where $\mathcal{C}$ refers to the convex set of $\mathbb{R}^{d \times k}$ matrices consisting of d-dimensional column vectors with norm at most 1, and α contains the sparse coefficients for the elements of the vocabulary. The only difference compared to Berend (2017) is that here we ensure a non-negativity constraint over the elements of α.

For the elements of the vocabulary we ran the *formal concept analysis* tool of Endres, Földiák, and Priss (2010)[5]. In order to keep the size of the DAG outputted by the FCA algorithm manageable, we only included the query words and those hypernyms in the analysis which

- A is a closed set with respect to $A \mapsto \overline{A}$, and $O = A'$.

5 www.compsens.uni-tuebingen.de/pub/pages/personals/3/concepts.py

Core feature name	
cosine	$\frac{\mathbf{q}^\intercal \mathbf{h}}{\|\mathbf{q}\|_2 \|\mathbf{h}\|_2}$
difference	$\|\mathbf{q} - \mathbf{h}\|_2$
normRatio	$\frac{\|\mathbf{q}\|_2}{\|\mathbf{h}\|_2}$
queryBeginsWith	$Q[0] = h$
queryEndsWith	$Q[-1] = h$
hasCommonWord	$Q \cap H \neq \varnothing$
sameFirstWord	$Q[0] = H[0]$
sameLastWord	$Q[-1] = H[-1]$
logFrequencyRatio	$\log_{10} \frac{count(q)}{count(h)}$
isFrequentHypernym	$c \in MF_{50}(q.type)$
sameConcept	$n(h) = n(q)$
parent	$n(q) \prec n(h)$
child	$n(h) \prec n(q)$
overlappingBasis	$\phi(q) \cap \phi(h) \neq \varnothing$
sparseDifference$_{q \setminus h}$	$\|\phi(q) - \phi(h)\|$
sparseDifference$_{h \setminus q}$	$\|\phi(h) - \phi(q)\|$
attributePair$_{ij}$	$\langle i, j \rangle \in \phi(q) \times \phi(h)$

Table 21: The features employed in our classifier. $MF_{50}(q.type)$ refers to the set of top-50 *m*ost *f*requent hypernyms for a given query type. At submission time, this feature did not work properly.

occur in the training dataset for the corpora. As we will see in the next subsection, this restriction turns out to be very useful.

Next, we determine a handful of features for a pair of expressions (q, h) consisting of a query q and its potential hypernym h. Table 21 provides an overview of the features employed for a pair (q, h). We denote with $\mathbf{q}$ and $\mathbf{h}$ the 100-dimensional dense vectorial representations of q and h. Additionally, we denote with Q and H the sequence of tokens constituting the query and hypernym phrases. Finally, we refer to the set of basis vectors (in the FCA terminology, attributes) which are assigned non-zero weights in the reconstruction of the vectorial representation of q and h as $\phi(q)$ and $\phi(h)$. It is also considered as a feature (`isFrequentHypernym`) whether a particular candidate hypernym h belongs to the top-50 most frequent hypernyms for the category of q (i.e. concept or entity). The fact that this feature is useful signals that the test dataset is not ideal (recall Section 4.2.12). Modeling the two categories separately played an important role in the success of our systems.

Three additional features are defined for incorporating the concept lattice output by FCA. Denoting with $n(w)$ the concept that introduces w, i.e. the most specific location within the DAG for w, our features indicate whether $n(q)$ (1) coincides with that of h, (2) is the parent (immediate successor) for that of h, or (3) is the child (immediate predictions) for that of h. Parents, and even the inverse relation, proved to be more predictive than the conceptually motivated $q \leqslant h$. In Table 21, $n_1 \prec n_2$ denotes that n_1 is an immediate predecessor of n_2. We will see in post-evaluation ablation experiments, where we refer to the above three features as the *FCA* features, that they were not useful in our submissions.

The $\texttt{attributePair}_{ij}$ features above, our most important features, are indicator features for every possible interaction term between the sparse coefficients in α. That means that for a pair of words (q, h) we defined $\phi(q) \times \phi(h)$, i.e. candidates get assigned with the Cartesian product derived from the indices of the non-zero coefficients in α. Note that this feature template induces k^2 features, with k being the number of basis vectors introduced in the dictionary matrix D according to Eq. 4.

In order to rank potential hypernym candidates over the test set we trained a *logistic regression* classifier for concepts and entities utilizing the `sklearn` package (Pedregosa et al. 2011)[6] with the regularization parameter defaulting to 1.0.

For each appropriate (q, h) pair of words for which h is a hypernym of q, we generated a number of *negative samples* (q, h') (Section 4.2.3)), such that the training data does not include h' as a valid hypernym for q. For a given query q, either *concept* or *entity*, we sampled h' from those hypernyms which were included as a valid hypernym in the training data with respect to some $q' \neq q$ query phrase.

When making predictions for the hypernyms of a query, we relied on our query type sensitive logistic regression model to determine the ranking of the hypernym candidates. In our official submission, the ranking was restricted to the phrases which appeared in the training data as a proper hypernym at least once.

After the appropriate model ranked the hypernym candidates, we selected the top 15 ranked candidates and applied a *post-ranking* heuristic over them, i.e. reordered them according to their background frequency from the training corpus. Our assumption here is that more frequent words tend to refer to more general concepts and more general hypernymy relations potentially tend to be more easily detectable than more special ones.

6 scikit-learn.org

		without attribute pairs						with attribute pairs					
		MAP	MRR	P@1	P@3	P@5	P@15	MAP	MRR	P@1	P@3	P@5	P@15
1A	offic	8.6	18.0	13.0	8.9	8.2	7.9	8.9	19.4	14.9	9.3	8.6	8.1
1A	reprod	9.07	18.7	13.5	9.4	8.8	8.5	9.2	19.9	14.9	9.5	8.7	8.4
1B	offic	9.4	19.9	13.2	9.5	9.3	8.8	12.1	25.1	17.6	12.9	11.7	11.2
1B	reprod	9.2	19.5	12.8	8.9	8.9	8.7	12.8	26.7	18.9	13.6	12.4	11.9
1C	offic	12.5	25.9	16.6	13.6	12.6	11.5	17.9	37.6	27.8	19.7	17.1	16.3
1C	reprod	12.9	26.0	16.2	13.9	13.0	11.9	18.3	38.4	28.5	20.2	17.4	16.6
2A	offic	15.0	32.2	24.8	17.7	15.8	11.6	20.8	40.6	31.6	23.5	21.4	17.1
2A	reprod	15.1	32.4	24.4	18.0	16.2	11.8	21.5	43.7	35.6	25.3	21.8	17.0
2B	offic	19.1	36.7	27.2	23.0	20.1	15.4	29.5	46.4	33.0	31.9	28.9	27.7
2B	reprod	21.5	40.9	29.6	25.6	22.1	18.0	30.4	46.8	33.8	31.8	29.5	28.9

Table 22: Our submissions results: `offic`ial and those that can be `reprod`uced with the code in the project repo (with the *isFrequentHypernym* feature being turned off).

7.1.3 *Results*

7.1.3.1 *Our submissions*

Our submissions were based on $k = 200$ dimensional sparse vectors computed from unit-normed 100-dimensional dense vectors with $\lambda = .3$. The sum of the two dimensions d and k motivates our group name 300-sparsans. For training the regression model with negative samples, 50 false hypernyms were sampled for each query q in the training dataset. One of our submissions involved attribute pairs, the other not. Both submissions used the conceptually motivated but practically harmful FCA-based features.

Table 22 shows submission results. The figures that can be reproduced with the code in the project repo (`reprod`) is slightly different from our official submissions (`offic`) for two reasons: because the implementation of `isFreqHyp` contained a bug, and because of the natural randomness in negative sampling. For reproducibility, we report result without the `isFreqHyp` feature. The randomness introduced by negative sampling is now factored out by setting the random seed.

7.1.3.2 *Query type sensitive baselining*

Our submission with attribute pairs achieved first place in categories (1B) Italian (all and entities), (1C) Spanish entities, and (2B) music entities. This is in part due to our good choice of a fallback solution in the case of OOV queries: we applied a category-sensitive baseline returning the most frequent training hypernym in the corresponding query type (concept or entity). Table 23 shows how frequently we had to rely on this fallback, and Table 24 shows the corresponding pure baseline results.

	Train		Test	
1A	975(4)	0.41%	1055(4)	0.38%
1B	709(1)	0.14%	767(2)	0.26%
1C	776(2)	0.26%	625(2)	0.32%
2A	442(58)	11.60%	433(67)	13.40%
2B	366(21)	5.43%	341(17)	4.75%

(a) concept

	Train		Test	
1A	379(142)	27.26%	344(99)	22.35%
1B	249(41)	14.14%	205(26)	11.26%
1C	184(38)	17.12%	328(45)	12.06%
2A	0(0)	—	0(0)	—
2B	79(34)	30.09%	102(40)	28.17%

(b) entity

Table 23: Number of in-vocabulary (and out-of-vocabulary, OOV) queries per query type. The ratio of the latter is also shown.

	MAP	MRR	P@1	P@3	P@5	P@10
1A	9.8	22.6	19.8	10.0	9.0	8.6
1A	8.8	21.4	19.8	8.9	7.8	7.5
1B	8.9	21.2	17.1	9.1	8.3	7.9
1B	7.8	19.4	17.1	8.3	6.8	6.5
1C	16.4	33.3	24.6	17.5	16.1	14.9
1C	12.2	29.8	24.6	12.0	11.3	11.0
2A	29.0	35.9	32.6	34.3	34.2	21.7
2A	28.9	35.8	32.6	34.3	34.2	21.4
2B	40.2	58.8	50.6	44.6	40.3	35.5
2B	33.3	51.5	36.2	40.1	35.8	28.4

Table 24: Baseline results, most frequent training hypernyms. We (upper) consider the most frequent hypernym in the given query type (concept or entity). For comparison, we also show the `MFH` baseline provided by the organizers (lower) that is based on the most frequent hypernyms in general.

		candidate filtering off						candidate filtering on					
k	ns	MAP	MRR	P@1	P@3	P@5	P@15	MAP	MRR	P@1	P@3	P@5	P@15
200	50	6.5	14.9	13.1	7.4	6.1	5.5	12.1	25.4	18.9	12.9	11.6	10.9
200	all	6.9	15.8	14.1	7.6	6.3	5.8	13.0	27.1	19.9	14.2	12.5	11.8
300	50	6.9	15.8	13.9	7.6	6.4	5.9	12.1	25.7	19.5	13.0	11.5	11.0
300	all	8.0	17.8	15.4	8.9	7.4	6.8	13.5	28.0	21.1	14.5	12.9	12.3
1000	50	9.0	20.0	17.2	9.8	8.3	7.7	13.3	**28.1**	**21.3**	13.8	12.6	12.3
1000	all	11.6	26.1	22.5	12.5	10.8	10.0	**13.6**	27.2	19.4	**13.9**	**13.2**	**12.8**

Table 25: Post evaluation results on the 1A dataset investigating the effect of various hyperparameter choices. k and ns denotes the number of basis vectors and negative samples generated during training per each positive (q, h) pair. Best results obtained for each metric are marked as bold.

		MAP	MRR	P@1	P@3	P@5	P@15
off	off	10.3	21.3	15.0	10.6	10.1	9.6
off	on	10.1	21.1	14.9	10.5	9.9	9.5
on	off	**12.1**	25.4	**18.9**	12.9	**11.6**	10.9
on	on	**12.1**	**25.3**	18.7	**13.0**	**11.6**	**11.0**

Table 26: Ablation experiments, on the 1A dataset with $k = 200, ns = 50$ (and the implementation of `isFreqHyp` fixed). The first two columns indicate whether $\text{attributePair}_{ij}$ and FCA-derived features are utilized, respectively.

7.1.3.3 *Post-evaluation analysis*

After the evaluation closed, we conducted ablation experiments, the results of which are shown in Table 26. In these experiments, we investigated the contribution of the features derived from sparse attribute pairs and FCA. These ablation experiments corroborate the importance of features derived from sparse attribute pairs and reveal that turning off FCA-based features does not hurt performance at all. For this reason – even though our official shared task submission included FCA-related features – we no longer employed them in our post-evaluation experiments.

Table 25 contains the detailed behavior of our model on subtask 1A with respect to three factors, that is

1. the number of basis vectors employed during sparse coding ($k \in \{200, 300, 1000\}$),
2. the number of negative training samples per positive sample ($ns \in \{50, all\}$), and
3. candidate filtering being turned on/off.

In our original submission we generated 50 negative samples (ns) per query q during training. In our post evaluation experiments we investi-

	MAP	MRR	P@1	P@3	P@5	P@15
1A	76.1	92.2	92.2	82.3	76.4	71.6
1B	71.2	93.4	93.4	78.5	70.9	65.7
1C	81.0	95.9	95.9	87.2	81.7	76.4
2A	72.6	89.6	89.6	81.0	75.3	64.1
2B	95.4	98.8	98.8	97.3	96.0	93.7

Table 27: Test results of an oracle system which uses candidate filtering.

gated the effects of generating more negative samples, i.e. we regarded all the valid hypernyms over the training set – not being a proper hypernym for q – as h' upon the creation of the (q, h') negative training instances. This latter strategy is referenced as $ns = all$ in Table 25.

In our official submission we regarded only those hypernyms as potential candidates to rank during test time which occurred at least once as a correct hypernym in the training data. We call this strategy as candidate filtering. Historically, we applied this restriction to speed up the FCA algorithm because this way the size of the concept lattice could be made smaller. As there are valid hypernyms on the test set which never occurred in the training data, our official submission would not be able to obtain a perfect score even in theory. As ceiling analysis, Table 27 contains the best possible metrics on the test set that we could achieve when candidate filtering is applied. In our post evaluation experiments we also investigated the effects of turning this kind of filtering step off. As Table 25 illustrates, however, our scores degrade without candidate filtering.

Our post evaluation experiments in Table 25 suggest that it is advantageous to apply sparse representation of more expressive power, i.e. with a higher number of basis vectors. Generating more negative samples also provides some additional performance boost. These previous observations hold irrespective whether candidate filtering is employed or not, however, their effects are more pronounced when hypernym candidates are not filtered.

Finally, we report our post-evaluation results for all the subtasks and compare them to the official scores of the best performing systems in Table 28. It can be seen that with these enhanced results that we would won category "all" (concepts and entities mixed) in languages (1B) Italian and (1C) Spanish. Our post-evaluation system – which only differs from our participating system that it fixes the calculation of the features, does not rely on FCA-based features, and uses $k = 1000$ – would also place third in the rest of the subtasks.

	MAP	MRR	P@1	P@3	P@10	P@15
1A	13.3	28.1	21.3	13.8	12.6	12.3
1A	19.8	36.1	29.7	21.1	19.0	18.3
1B	**12.5**	24.2	14.5	**13.4**	**12.5**	**12.0**
1B	12.1	**25.1**	**17.6**	12.9	11.7	11.2
1C	**21.8**	**43.8**	**33.7**	**22.9**	**21.4**	**19.9**
1C	20.0	28.3	21.4	20.9	21.0	19.4
2A	21.9	39.5	34.2	25.5	22.6	18.5
2A	34.0	54.6	49.2	40.1	36.8	27.1
2B	31.5	43.6	29.8	30.3	30.3	31.5
2B	41.0	60.9	48.2	44.9	41.3	38.0

Table 28: Post evaluation results for the different subtasks using $k = 1000, ns = 50$ and hypernym candidate filtering. Upper: our system, lower: subtask winner.

7.1.4 *Conclusion*

In this section we experimented with the integration of sparse word representations into the task of hypernymy discovery. We strove to utilize sparse word representations in two ways, i.e. via building concept lattices using formal concept analysis and modeling the hypernymy relation with the help of interaction terms. While our former approach for deriving formal concepts from sparse word representations was not successful, the interaction terms derived from sparse word representations proved to be highly beneficial, placing first in more categories of SemEval 2018 Task 9.

7.2 ANTONYMS IN AN EMBEDDING FROM DEFINITIONS

In this section, which originally appeared as Makrai, Nemeskey, and Kornai (2013)[7], we test which putative semantic features like GENDER are captured by VSMs. We assume that the difference between two vectors, for antonyms, distills the actual property which is the opposite in each member of a pair of antonyms. So, for example, for a set of male and female words, such as ⟨king, queen⟩, ⟨actor, actress⟩, etc., the difference between words in each pair should represent the idea of gender. To test the hypothesis, we associated antonymic word pairs from the

7 Makrai classified the antonymic relation pairs, and prepared the statistical test. Nemeskey created the embedding, and finished the experiments. The function-applicational idea (mapping a deep case to a function), which gave the title of the paper, is due to Kornai. We thank Zsófia Tardos and the anonymous reviewers for useful comments.

GOOD		VERTICAL	
safe	out	raise	level
peace	war	tall	short
pleasure	pain	rise	fall
ripe	green	north	south
defend	attack	shallow	deep
conserve	waste	ascending	descending
affirmative	negative	superficial	profound
⋮	⋮	⋮	⋮

Table 29: Word pairs associated to features GOOD and VERTICAL

WordNet (Miller (1995), see Section 2.4.3) to 26 classes, e.g. END/BEGINNING, GOOD/BAD, . . . , see Table 29 and Table 31 for examples.

7.2.1 *Our method: comparison to random permutation*

The intuition to be tested is that the first member of a pair relates to the second one in the same way among all pairs associated to the same feature. For k pairs $\vec{x_i}, \vec{y_i}$ we are looking for a common vector $\vec{a}$ such that

$$\vec{x_i} - \vec{y_i} = \vec{a} \tag{5}$$

Given the noise in the embedding, it would be naive in the extreme to assume that (5) can be a strict identity. Rather, we take the mean of the offsets $x_i - y_i$, which minimizes the error

$$Err = \sum_i ||\vec{x_i} - \vec{y_i} - \vec{a}||^2 \tag{6}$$

The question is simply the following: is this error obtained with the mean vector any better than what we could expect from a bunch of random $\vec{x_i}$ and $\vec{y_i}$?

We selected 26 potentially antonymic datasets from WordNet such as the 'gender' set discussed above. For example, the 'hard' set contains the pairs *hardened/soft, hardball/softball, hardware/software, still/sparkling, hard/soft, solid/gaseous, tough/tender, liquid/gaseous, hardness/softness, hard_drug/soft_drug, hard_water/soft_water* and the 'distance' set contains the pairs *express/local, distant/close, repulsive/attractive, open/close, far/near, distribution/concentration, distributed/concentrated, expanded/contracted, ultimate/proximate, distal/proximal.* Since the sets are of different sizes, we took 100 random pairings of the words appearing on either sides of the pairs to estimate the error distribution, computing the minima of

# pairs	feature name	HLBL original				HLBL scaled				SENNA			
		Err	m	σ	r	*Err*	m	σ	r	*Err*	m	σ	r
156	good	1.92	2.29	0.032	11.6	4.15	4.94	0.0635	12.5	50.2	81.1	1.35	22.9
42	vertical	1.77	2.62	0.0617	13.8	3.82	5.63	0.168	10.8	37.3	81.2	2.78	15.8
49	in	1.94	2.62	0.0805	8.56	4.17	5.64	0.191	7.68	40.6	82.9	2.46	17.2
32	many	1.56	2.46	0.0809	11.2	3.36	5.3	0.176	11	43.8	76.9	3.01	11
65	active	1.87	2.27	0.0613	6.55	4.02	4.9	0.125	6.99	50.2	84.4	2.43	14.1
48	same	2.23	2.62	0.0684	5.63	4.82	5.64	0.14	5.84	49.1	80.8	2.85	11.1
28	end	1.68	2.49	0.124	6.52	3.62	5.34	0.321	5.36	34.7	76.7	4.53	9.25
32	sophis	2.34	2.76	0.105	4.01	5.05	5.93	0.187	4.72	43.4	78.3	2.9	12
36	time	1.97	2.41	0.0929	4.66	4.26	5.2	0.179	5.26	51.4	82.9	3.06	10.3
20	progress	1.34	1.71	0.0852	4.28	2.9	3.72	0.152	5.39	47.1	78.4	4.67	6.7
34	yes	2.3	2.7	0.0998	4.03	4.96	5.82	0.24	3.6	59.4	86.8	3.36	8.17
23	whole	1.96	2.19	0.0718	3.2	4.23	4.71	0.179	2.66	52.8	80.3	3.18	8.65
18	mental	1.86	2.14	0.0783	3.54	4.02	4.6	0.155	3.76	51.9	73.9	3.52	6.26
14	gender	1.27	1.68	0.126	3.2	2.74	3.66	0.261	3.5	19.8	57.4	5.88	6.38
12	color	1.2	1.59	0.104	3.7	2.59	3.47	0.236	3.69	46.1	70	5.91	4.04
17	strong	1.41	1.69	0.0948	2.92	3.05	3.63	0.235	2.48	49.5	74.9	3.34	7.59
16	know	1.79	2.07	0.0983	2.88	3.86	4.52	0.224	2.94	47.6	74.2	4.29	6.21
12	front	1.48	1.95	0.17	2.74	3.19	4.21	0.401	2.54	37.1	63.7	5.09	5.23
22	size	2.13	2.69	0.266	2.11	4.6	5.86	0.62	2.04	45.9	73.2	4.39	6.21
10	distance	1.6	1.76	0.0748	2.06	3.45	3.77	0.172	1.85	47.2	73.3	4.67	5.58
10	real	1.45	1.61	0.092	1.78	3.11	3.51	0.182	2.19	44.2	64.2	5.52	3.63
14	primary	2.22	2.43	0.154	1.36	4.78	5.26	0.357	1.35	59.4	80.9	4.3	5
8	single	1.57	1.82	0.19	1.32	3.38	3.83	0.32	1.4	40.3	70.7	6.48	4.69
8	sound	1.65	1.8	0.109	1.36	3.57	3.88	0.228	1.37	46.2	62.7	6.17	2.67
7	hard	1.46	1.58	0.129	0.931	3.15	3.41	0.306	0.861	42.5	60.4	8.21	2.18
10	angular	2.34	2.45	0.203	0.501	5.05	5.22	0.395	0.432	46.3	60	6.18	2.2

Table 30: Error of approximating real antonymic pairs (*Err*), mean and standard deviation (m, σ) of error with 100 random pairings, and the ratio $r = \frac{|Err-m|}{\sigma}$ for different features and embeddings

$$Err_{rand} = \sum_i ||\vec{x_i}' - \vec{y_i}' - \vec{a}||^2 \tag{7}$$

For each distribution, we computed the mean and the variance of Err_{rand}, and checked whether the error of the correct pairing, *Err* is at least 2 or 3 σs away from the mean.

Table 30 summarizes our results for three embeddings: the original and the scaled HLBL (Section 4.2.2, Mnih and G. E. Hinton (2009)) and SENNA (Section 4.2.3). The first two columns give the number of pairs considered for a feature and the name of the feature. For each of the three embeddings, we report the error *Err* of the unpermuted arrangement, the mean m and variance σ of the errors obtained under random permutations, and the ratio

$$r = \frac{|m - Err|}{\sigma}.$$

Horizontal lines divide the features to three groups: for the upper group, $r \geqslant 3$ for at least two of the three embeddings, and for the middle group $r \geqslant 2$ for at least two.

PRIMARY		ANGULAR	
leading	following	square	round
preparation	resolution	sharp	flat
precede	follow	curved	straight
intermediate	terminal	curly	straight
antecedent	subsequent	angular	rounded
precede	succeed	sharpen	soften
question	answer	angularity	roundness
⋮	⋮	⋮	⋮

Table 31: Features that fail the test

For the features above the first line we conclude that the antonymic relations are well captured by the embeddings, and for the features below the second line we assume, conservatively, that they are not. (In fact, looking at the first column of Table 30 suggests that the lack of significance at the bottom rows may be due primarily to the fact that WordNet has more antonym pairs for the features that performed well on this test than for those features that performed badly, but we did not want to start creating antonym pairs manually.) For example, the putative sets in Table 31 does not meet the criterion and get rejected.

7.2.2 *Embedding from a definition graph*

The `4lang` embedding is created in a manner that is notably different from HLBL and SENNA. Our input is a graph whose nodes are concepts, with edges running from A to B if and only if B is used in the manually written `4lang` definition (Chapter 3) of A. The base vectors are obtained by the spectral clustering method pioneered by Ng, Jordan, and Weiss (2001): the incidence matrix of the conceptual network is replaced by an affinity matrix whose ij-th element is formed by computing the cosine distance of the ith and jth row of the original matrix, and the first few (in our case, 100) eigenvectors are used as a basis.

Since the concept graph includes the entire Longman Defining Vocabulary (LDV), each LDV element w_i corresponds to a base vector b_i. For the vocabulary of the whole dictionary, we simply take the Longman definition of any word w, strip out the stopwords (we use a small list of 19 elements taken from the top of the frequency distribution), and form $V(w)$ as the sum of the b_i for the w_is that appeared in the definition of w (with multiplicity).

We performed the same computations based on this embedding as in Section 7.2.1, and the results are presented in Table 32. Judgment columns under the three embeddings in the previous section and `4lang` are highly correlated, see table 33.

# pairs	feature name	4lang *Err*	4lang *m*	4lang *σ*	4lang *r*
49	in	0.0553	0.0957	0.00551	7.33
156	good	0.0589	0.0730	0.00218	6.45
42	vertical	0.0672	0.1350	0.01360	4.98
34	yes	0.0344	0.0726	0.00786	4.86
23	whole	0.0996	0.2000	0.02120	4.74
28	end	0.0975	0.2430	0.03410	4.27
32	many	0.0516	0.0807	0.00681	4.26
14	gender	0.0820	0.2830	0.05330	3.76
36	time	0.0842	0.1210	0.00992	3.74
65	active	0.0790	0.0993	0.00553	3.68
20	progress	0.0676	0.0977	0.00847	3.56
18	mental	0.0486	0.0601	0.00329	3.51
48	same	0.0768	0.0976	0.00682	3.05
22	size	0.0299	0.0452	0.00514	2.98
16	know	0.0598	0.0794	0.00706	2.77
32	sophis	0.0665	0.0879	0.00858	2.50
12	front	0.0551	0.0756	0.01020	2.01
10	real	0.0638	0.0920	0.01420	1.98
8	single	0.0450	0.0833	0.01970	1.95
7	hard	0.0312	0.0521	0.01960	1.06
10	angular	0.0323	0.0363	0.00402	0.999
12	color	0.0564	0.0681	0.01940	0.600
8	sound	0.0565	0.0656	0.01830	0.495
17	strong	0.0693	0.0686	0.01111	0.0625
14	primary	0.0890	0.0895	0.00928	0.0529
10	distance	0.0353	0.0351	0.00456	0.0438

Table 32: The results on 4lang

	HLBL original	HLBL scaled	Senna	4lang
HLBL original	1	0.925	0.422	0.856
HLBL scaled	0.925	1	0.390	0.772
Senna	0.422	0.390	1	0.361
4lang	0.856	0.772	0.361	1

Table 33: Correlations between judgments based on different embeddings

Unsurprisingly, the strongest correlation is between the original and the scaled HLBL results. Both the original and the scaled HLBL correlate notably better with `4lang` than with SENNA, making the latter the odd one out.

The contribution of this section is that we showed that a dictionary-based embedding, when used for a purely semantic task, the analysis of antonyms, does about as well as the more standard embeddings based on cooccurrence data. Clearly, a VSM could be obtained by the same procedure from any machine-readable dictionary. Using LDOCE is computationally advantageous in that the core vocabulary is guaranteed to be very small, but finding the eigenvectors for an 80k by 80k sparse matrix would also be within CPU reach.

7.2.3 *Conclusion*

We created word embeddings from monolingual dictionary definitions and compared them with two standard embeddings (HLBL, (Mnih and G. E. Hinton 2009) and Senna (Collobert et al. 2011)) regarding which potential subrelations of antonymy is represented in them. In the three-way comparison, Senna proved to be the off man out, suggesting that the dictionary-based embedding encodes similar information as HLBL.

The vector offset method for solving analogical questions assumes that the four words (e.g. *king, queen, man*, and *woman*) form a parallelogam. In the next section we investigate causality, and find a different geometry.

7.3 CAUSALITY IN VECTOR SPACE LANGUAGE MODELS

In this section, which originally appeared as Makrai (2014a), we take a semantic relation with a rich literature in linguistics (Section 2.3.5) and philosophy and with many applications in knowledge representation: causality (see Figure 17). We are interested in the geometric function mapping the vector representation of a cause (e.g. *hurt*) to the vector representing the corresponding effect (*ache*). These results exemplify an exploratory data science approach to the computational analysis of the cognitive structures underlying linguistic understanding: To quote (Schank 1973) in a different context, inferences are generally made "to see what they can see".

We took causal word pairs from one of the most popular natural language processing resources containing lexical information of various kinds, WordNet (Miller (1995), see Section 2.4.3). The pairs are exemplified in Table 34. We took several VSMs: SENNA (Section 4.2.3), those published along with the papers Turian, Ratinov, and Bengio (2010) and Huang et al. (2012), HLBL (Section 4.2.2, Mnih and G. E. Hinton (2009)), the English Polyglot (Al-Rfou', Perozzi, and Skiena 2013), and 24 variants of the model created from `4lang`, the semantic

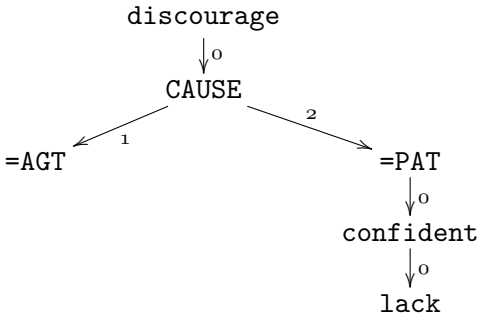

Figure 17: The definition of *discourage* in the `4lang` concept lexicon exemplifies the use of 'cause' in associative network representations of linguistics knowledge. The graph expresses that *discourage* means, that the agent (`=AGT`) causes the participant that is called patient in linguistics (`=PAT`) to lack confidence.

network we introduced in Chapter 3. (These variants differ in minor hyper-parameters, e.g. whether there is a link from the headnode of the graph back to the definiendum.) Causal pairs were projected to a 2-dimensional plane by principal component analysis, a machine learning technique often used for visualizing high-dimensional data. The visualization suggested that there is a center in the vector space of the words, that approximately fits the lines containing each causal pair, see Figure 18.

For testing the centrality property in the original, unreduced space, we took random word pairs of the same number as we have causal pairs. The point closest to all the lines fitting each pair was computed for both the real and the random sample of word pairs using a formula by Han and Bancroft (2010). The distances of the lines to the corresponding center was also computed. We formalize centrality as that the expected value of the distances is lower in the real case than in the random case. An unpaired t-test showed that this condition holds in the case of Senna ($p < 0.001$).

Some of the models created from `4lang` also show significant ($p < 0.05$) difference, but this statistical result has to be taken with caution, because of the phenomenon known as *multiple testing* (Domingos 2012).[8]

> Standard statistical tests assume that only one hypothesis is being tested, but modern learners can easily test millions before they are done. As a result what looks significant may in fact not be. [...] This problem can be combatted by correcting the significance tests to take the number of hypotheses into account [...]

8 I would like to thank Balázs Szalkai for reminding me to this problem.

give	have
show	see
encourage	hope
feed	eat
kill	die
raise	rise
⋮	⋮

Table 34: Word causes and effects in WordNet. WordNet contains semantic relations like IS-A (a chair is a furniture), instance-of (Mozart is an instance of 'composer'), antonym (cold and hot), part-of (Monday is a part of 'week') as well.

Multiplying the p values by 24 significance is lost, so unless we motivate the choice of some specific model among all `4lang` models on some independent grounds, we have to conclude that in the `4lang` models, the centrality hypothesis has to be rejected.

7.3.1 *Discussion and conclusion*

Looking for an insightful interpretation of causality in VSMs, we have found a center point $\mathbf{c}$ in the VSM SENNA with the property that the lines connecting the two members of causal word pairs run close to $\mathbf{c}$. In algebraic terms this means that

$$\mathbf{v}_{\text{effect}} \approx \lambda \mathbf{v}_{\text{cause}} + (1 - \lambda)\mathbf{c}$$

with a verb-dependent $\lambda \in \mathbb{R}$. This linear algebraic property reflects the linguistic intuition that the meaning of the effect is a combination of the meaning of the cause and a causal element.

While more sophisticated connections may exist between cause and effect vectors in a boroader family of embedding models, we focused on this easily interpretable relation.

Exploring the geometry of the causality in a 2d visualization, we formulated the hypothesis that the lines connecting causal pairs run close to a common point. We put this in the context that causes are intuitively composed of the cause and a constant causal element (lexical feature). We tested this hypothesis in various embeddings. In Senna (Collobert et al. 2011), the hypothesis holds, in the other embeddings studied it does not do so (taking the problem of multiple testing into account).

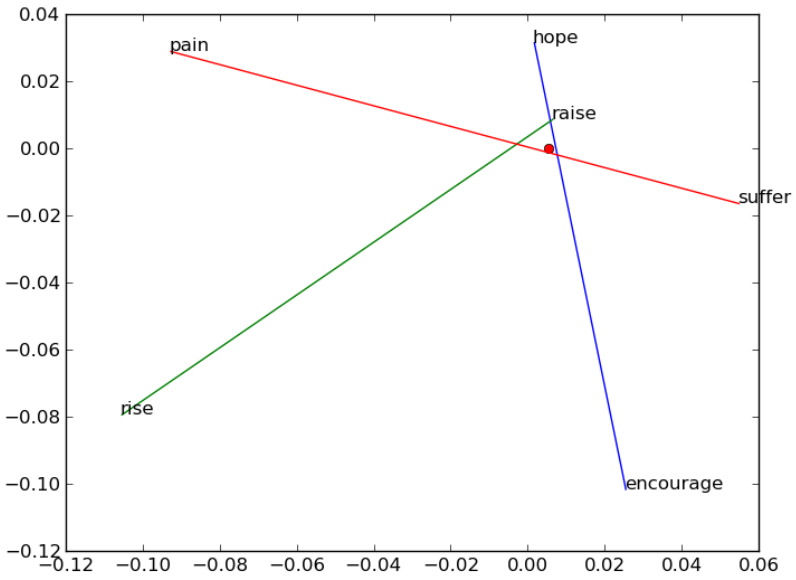

Figure 18: A 2-d visualization of causal pairs in the VSMs suggest that lines connecting causal pairs run close to a common center point.

7.4 ANALOGY AND TRANSLATION

We saw in Chapter 4, especially Section 4.2.4 that static word embeddings represent analogical relations in a simple linear algebraic form. These properties were originally discovered for English, where morphology is marginal, and syntactic roles are marked by word order. Now we ask whether word embeddings of a morphologically rich and consequently relatively free phrase-order language, Hungarian, exhibit similar structure. To do so, we published the Hungarian equivalent of the analogical benchmark. While the consequent representation of word relations is interesting for lexical theory on its own, from an engineering point of view, it is important to see to what extent such an "in vitro" or "intrinsic" task predicts the accuracy of the embedding at hand when applied to some "in vivo" or "extrinsic" task like word translation. This section revolves around these two questions.

In Section 4.2, we introduced word2vec and GloVe as the two most successful tools that create static word embeddings (a.k.a. vector space language models, VSMs) from gigaword corpora. word2vec implements the neural network style architectures skip-gram and cbow, learning parameters using each word as a training sample, while GloVe factorizes the co-occurrence matrix (or more precisely a matrix of conditional probabilities) as a whole.

In this section, most of which originally appeared as Makrai (2015), we test the two systems in two tasks. First we are interested in how analogical relations are represented in a static word embedding of a mor-

phologically rich (and accordingly, free phrase-order) language. More precisely, we test Hungarian word embeddings in a Hungarian equivalent of the popular word analogy task. We also test the gluten-free method (Section 4.2.11.1) in the analogical question benchmark.

In Section 7.4.2, we test static word embeddings in word translation between European languages including medium-resourced ones: Hungarian, Lithuanian and Slovenian. This task is related to statistical machine translation. The goal of the project to which the line of research reported here belonged, was to generate *proto-dictionaries* for European languages with fewer speakers. We collected translational word pairs between English, Hungarian, Slovenian, and Lithuanian. We took the method of Mikolov, Le, and Sutskever (2013) who train VSMs for the source and the target language from monolingual corpora, and compute word translation by learning a mapping between these supervised by a seed dictionary of a few thousand items.

The collection of word translations can be compared to the independent and simpler task of analogy. For this, we created the Hungarian equivalent of the test question set by Mikolov, Yih, and Zweig (2013) and Mikolov, Chen, et al. (2013).[9] We compare results in the two tasks, monolingual analogies and translation in Section 7.4.4.3.

It turns out that vanilla static embeddings of Hungarian trained on the semi-gigaword corpora of the time model semantic relations poorly. Section 7.4.6 investigate to what extent the gluten-free method (Section 4.2.11.1) or a larger corpus can solve this problem.

The only related work that evaluated vector models of a language other than English on word analogy in these early years of word embeddings we know is Sen and Erdogan (2014) who compared different strategies to deal with the a morphologically rich Turkish language[10]. As far as we know, the application of GloVe to word translation was a novelty of Makrai (2015).

7.4.1 *A Hungarian analogical benchmark*

Measuring the quality of VSMs in a task-independent way (a.k.a. *intrinsic evaluation*) is motivated by the idea of representation sharing. VSMs that capture something of language itself are better than ones just tailored for a single task (Section 4.2.3).

Section 4.2 (especially Sections 4.2.4 and 4.2.7 and the critical Sections 4.2.12 and 4.2.13) introduced analogical questions (also called relational similarities (Turney 2006) or linguistic regularities (Mikolov, Yih, and Zweig 2013)) as intrinsic measures of merit for vector models. This test gained remarkable popularity in the VSM community. Mikolov, Yih, and Zweig (2013) observe that analogical questions like

9 For data and else visit the project page http://corpus.nytud.hu/efnilex-vect.

10 I'm grateful to Mehmet Umut Sen for sending me the English translation the essence of Sen and Erdogan (2014).

"*good* is to *better* what *rough* is to . . . " or "*man* is to *woman* what *king* is to . . . " can be answered by basic linear algebra in neural VSMs:

$$\text{good} - \text{better} \approx \text{rough} - \mathbf{x} \tag{8}$$

$$\mathbf{x} \approx \text{rough} - \text{good} + \text{better} \tag{9}$$

In this example, the difference corresponds to the morphological relation of the comparative. So the vector nearest to the right side of Equation (9) is supposed to be *rougher*, which is really the case.

Recall that analogical benchmarks as the methods of *in vitro* evaluation have been criticized for various reasons. On the engineering level, Levy et al. (2015, summarized in our Section 4.2.12) showed that the offset method learns superficial features of individual words like being a typical hypernym/holonym, rather than relations; that in morphological questions, the offset is so short that the parallelogram search simplifies to nearest neighbor search; and that the whole idea is hacked by excluding the question words (i.e. *man, woman*, and *king* in the *queen* example) from among the answer candidates. The latter is especially problematic in the light of analogical relational pairs where the correct answer is actually one of the question words (Gladkova and Drozd (2016) and Rogers, Drozd, and Li (2017), summarized in our Section 4.2.13). Nevertheless, analogical question remain one of the basic *in vitro* evaluations of static word embeddings, so we find it advantageous to have them at hand for Hungarian.

We created a Hungarian equivalent of the analogical questions made publicly available by Mikolov, Yih, and Zweig (2013) and Mikolov, Chen, et al. (2013). More precisely, we followed the main ideas reported in Mikolov, Yih, and Zweig (2013), and targeted the sizes of the data-set accompanying Mikolov, Chen, et al. (2013).

Analogical pairs are divided to "grammatical" (i.e. morphological) and "semantic" (mostly world knowledge) ones. The morphological pairs in Mikolov, Yih, and Zweig (2013) were created in the following way:

> [We test] base/comparative/superlative forms of adjectives; singular/plural forms of common nouns; possessive/non-possessive forms of common nouns; and base, past and 3rd person present tense forms of verbs. More precisely, we tagged 267M words of newspaper text with Penn Treebank POS tags (Marcus, Santorini, and Marcinkiewicz 1993). We then selected 100 of the most frequent comparative adjectives (words labeled JJR); 100 of the most frequent plural nouns (NNS); 100 of the most frequent possessive nouns (NN POS); and 100 of the most frequent base form verbs (VB).

The Hungarian morphological pairs (Table 35) were created accordingly: For each grammatical relationship, we took the most frequent

English		Hungarian	
plural	singular	plural	singular
decrease	decreases	lesznek	lesz
describe	describes	állnak	áll
eat	eats	tudnak	tud
enhance	enhances	kapnak	kap
estimate	estimates	lehetnek	lehet
find	finds	nincsenek	nincs
generate	generates	kerülnek	kerül

Table 35: Morphological word pairs

	English		Hungarian
	# pairs	# questions	# pairs
gram1-adjective-to-adverb	32	992	40
gram2-opposite	812	29	30
gram3-comparative	37	1332	40
gram4-superlative	34	1122	40
gram5-present-participle	33	1056	40
gram6-nationality-adjective	41	1599	41
gram7-past-tense	40	1560	40
gram8-plural-noun	37	1332	40
gram9-plural-verb	30	870	40
capital-common-countries	23	506	20
capital-world	116	4524	166
city-in-state	68	2467	
county-center			19
county-district-center			175
currency	30	866	30
family	23	506	20

Table 36: Sizes of the question sets

English		Hungarian	
Athens	Greece	Budapest	Magyarország
Baghdad	Iraq	Moszkva	Oroszország
Bangkok	Thailand	London	Nagy-Britannia
Beijing	China	Berlin	Németország
Berlin	Germany	Pozsony	Szlovákia
Bern	Switzerland	Helsinki	Finnország
Cairo	Egypt	Bukarest	Románia

Table 37: Semantic word pairs

English				Hungarian			
Athens	Greece	Baghdad	Iraq	Budapest	Magyarország	Moszkva	Oroszország
Athens	Greece	Bangkok	Thailand	Budapest	Magyarország	London	Nagy-Britannia
Athens	Greece	Beijing	China	Budapest	Magyarország	Berlin	Németország
Athens	Greece	Berlin	Germany	Budapest	Magyarország	Pozsony	Szlovákia
Athens	Greece	Bern	Switzerland	Budapest	Magyarország	Helsinki	Finnország
Athens	Greece	Cairo	Egypt	Budapest	Magyarország	Bukarest	Románia

Table 38: Analogical questions

inflected forms from the Hungarian Webcorpus (Halácsy et al. 2004). The suffix in question was restricted to be the last one. See sizes in Table 36. In the case of opposite, we restricted ourselves to forms with the derivational suffix *-tlan* (and its other allomorphs) to make the task morphological rather then semantic. plural-noun includes pronouns as well.

For the semantic task (Table 37), data were taken from Wikipedia. For the capital-common-countries task, we choose the one-word capitals appearing in the Hungarian Webcorpus most frequently. The English task city-in-state contains USA cities with the states they are located in. The equivalent tasks `county-center` contains counties (*megye*) with their centers (*Bács-Kiskun – Kecskemét*), and currency contains the currencies of the most frequent countries in the Webcorpus. The `family` task targets gender distinction. We filtered for the pairs where the gender distinction is sustained in Hungarian (but dropped e.g. *he – she*, where it is not). We put some relational nouns in the possessive case (*bátyja – nővére*). We note that this category contains the royal "family" as well, e.g. the famous *king – queen*, and even *policeman – policewoman*.

Column "# questions" in Table 36 shows how many questions are formed in the English dataset we follow. In the Hungarian case, both morphological and semantic questions were created by matching every pair with every other pair resulting in e.g. $\binom{20}{2}$ questions for family. Some examples for questions are shown in Table 38.

cos >	vocab	gold	prec@1	prec@5
0.7	3803	301	68.4%	84.4%
0.6	9967	711	54.7%	74.1%
0.5	12949	958	46.6%	65.6%
0.4	13451	988	45.3%	64.0%

Table 39: Trade-off between precision and recall in Hungarian to English word translation

7.4.2 *Word translation in European languages*

For the collection of word translations, we follow the method of Mikolov, Le, and Sutskever (2013) who start with creating a VSM for the source and the target language from monolingual corpora in the magnitude of billion(s) of words. VSMs represent words in vector spaces of some hundred dimensions. The key point of the method is learning a linear mapping from the source vector space to the target space supervised by a seed dictionary of 5 000 words. Training word pairs are taken from among the most frequent ones skipping out-of-vocabulary pairs i.e. those with a source or target word unknown to the respective language model. The learned mapping is used to find a translation for each word in the source model. The computed translation is the target word with a vector closest to the image of the source word vector by the mapping. Mikolov, Le, and Sutskever report their best results when the dimension of the source model is 2–4 times the dimension of the target model, e.g. 800 → 300. The closeness (cosine similarity) between the image of the source vector and the closest target vector provides a confidence measure for the translation. We will make further use of this interlingual similarity score in Section 7.5, to filter translation pairs obtained with more traditional methods. We generate word translations in the following language pairs: Hungarian-Lithuanian, Hungarian-Slovenian, and Hungarian-English.

The measure of confidence for each translational pair (the distance of the vector computed by mapping the source word vector, and the nearest target word vector) makes some tuning between precision and recall possible (see Table 39). With a higher cosine similarity cut-off (column "cos >"), we get word translations for a smaller vocabulary ("vocab") with a higher precision, while lower cosine similarities produce a greater vocabulary with translations of a lower precision. prec@1 is the ratio of words, for which the first candidate translation coincides with that provided in the seed dictionary, prec@5 is the ratio of words with the seed translation in the first 5 candidates. These are strict metrics, as synonyms of the gold translation count as incorrect. gold is the number of words with a gold translation in the corresponding part of the test data.

language	corpus	# words
Lithuanian	webcorpus (Zséder et al. 2012)	1.4 B
Slovenian	slWaC (Ljubešić and Erjavec 2011)	1.6 B
Hungarian	Webcorpus (Halácsy et al. 2004)	0.7 B
Hungarian	HNC (Oravecz, Váradi, and Sass 2014)	0.8 B

Table 40: Corpora for medium-resourced languages. Word counts are given in billions.

While the (pre-training) model of word2vec is based on the dot product, Mikolov, Yih, and Zweig (2013) use the least squares fitting of the Euclidean distance for training the mapping, and, surprisingly, cosine similarity for translation generation. We also found this combination of distances to be the only one that works. We return to techniques related to this peculiarity in Section 8.4.2.

7.4.3 *Data*

7.4.3.1 *Corpora and vectors*

For English, we use vector models downloaded from the home pages of the tools, while for the medium-resourced languages, we train new models on the corpora in Table 40,[11] using the tokenization provided by the authors of the text collection.

The basis for our deglutination experiments is the POS-disambiguated Webcorpus 2.0 (Nemeskey 2020). To train the gluten-free embeddings, we split compositional derivational and inflectional suffixes from the stem. More precisely, accidentally we tried two variants. each-separate separates each compositional morpheme (like `[_Mod/V]` in the example of Table 7) to its own token, as proposed by Nemeskey (2020). paradigm-cell on the other hand splits the word in two, a stem (`érdekel`) and an affix series (`[/V][_Mod/V][Prs.Def.3Sg]`).

Some noun and verb forms are unmarked in Hungarian (namely the nominative case `[Nom]` and present, indefinite, 3rd person singular `[Prs.NDef.3Sg]` respectively), and, in a somewhat analogous fashion, punctuation `[Punct]` is default on abbreviations, it is nevertheless reflected in the emMorph analysis (e.g. *pl.* is analyzed as `[/Adv|Abbr][Punct]`). Thus we disregard the corresponding substrings of the analysis. Embeddings were trained with gensim (Řehůřek and Sojka 2010) in 300 dimensions and 1 epoch. The remaining hyper-parameters were set to their default values.

11 I would like to thank Vladimír Benko for information on corpora.

	efnilex12	wikt	wikt triang	OSub12	OSub13	Europarl
en-hu	83 K	47 K	+134 K	97 K	19 K	21 K
hu-lt	152 K	6 K	+21 K	11 K	9 K	27 K
hu-sl	235 K	2 K	+26 K	63 K	45 K	29 K

Table 41: Number of translational word pairs in the seed dictionaries

7.4.3.2 *Seed dictionaries*

Mikolov, Le, and Sutskever (2013) use Google translate as a seed dictionary. We experimented with three seed dictionaries: (1) efnilex12, the proto-dictionaries collected within the EFNILEX project (Héja and Takács 2012), (2) word pairs computed using wikt2dict with and without triangulation (See Ács, Pajkossy, and Kornai (2013), and, for sizes, Table 41), and (3) dictionaries from the opus collection (Europarl, OpenSubtitles2012 and OpenSubtitles2013, Tiedemann (2012))[12]. efnilex12 contains directed dictionaries (ranked by the conditional probability (of co-occurrence) of the target word conditioned on the source word). In running automatic word alignment on the corresponding parallel data set, Tiedemann (2012) used "GIZA++ (Och and Ney 2003) and the symmetrization heuristics (`grow-diag-final-and`) implemented in Moses (Koehn et al. 2007) to extract probabilistic phrase tables". These tables are routinely used in phrase-based statistical machine translation.

7.4.4 *Results*

In the remainder of Section 7.4, we will use the abbreviations d for dimension, w for window radius ($w = 15$ means that (a maximum of) 15 words are considered on both sides of the word in focus), i for number of training iterations over the corpus (epochs), m for minimum word count in the vocabulary cutoff, and n for number of negative samples (in the case of word2vec).

7.4.4.1 *Results with analogical questions*

For comparing the results with the Hungarian analogical questions to those on the English ones, we trained sgram models on the concatenation of HNC and the Hungarian Webcorpus with $d = 300, m = 5$, either negative sampling or hierarchical softmax (two techniques to avoid computing the denominator of softmax that is a sum with as many terms as there are words in the embedding, the latter is based on a self-supervised hierarchical organization of the vocabulary, see Section 4.2.2 for both), and different levels of subsampling of frequent

12 http://opus.lingfil.uu.se/

		morph		semant		total	
en, Mikolov et al (2013)	$n = 5$	61		58		60	
	$n = 15$	61		61		61	
	HS	52		59		55	
hu	$n = 5$	63.0	3419/5430	38.5	269/699	60.2	3688/6129
	$n = 15$	61.9	3359/5430	39.2	274/699	59.3	3633/6129
	HS	48.9	2653/5430	22.5	157/699	45.8	2810/6129

Table 42: Comparison of results in our Hungarian word analogies (below the line) to those of the authors of the original Mikolov, Sutskever, et al. (2013)

words, see Mikolov, Sutskever, et al. (2013) for details. In Table 42, it can be seen that, in the *morphological* questions, we (below the line) get similar results in the Hungarian equivalent as the authors of the original task (Mikolov, Sutskever, et al. (2013), above the line), while in the *semantic* questions, Hungarian results are worse, suggesting that the semantic questions are too hard. We call attention to Novák and Novák (2018, in Hungarian) and the experiments by Döbrössy et al. (2019) summarized in Section 4.2.11.2 as well.

7.4.4.2 *Proto-dictionary generation*

In this paragraph we report our results in Slovenian/Hungarian/Lithuanian to English proto-dictionary generation. We take four source embeddings: two Slovenian ones trained on slWaC, one trained on the Hungarian Webcorpus, and one on the Lithuanian webcorpus by Zséder et al. (2012), all in $d = 600$. One of the Slovenian models is a GloVe one, the other models are cbow models with $n = 15$ and $w = 10$. The target model is always `glove.840B.300d`[13] from the GloVe site, the seed dictionary is OpenSubtitles2012. Either the source (rs), the target (rt) embedding, or both (rst) was restricted to words accepted by Hunspell. In Table 43 we compare our results (below the line) to those of Mikolov, Le, and Sutskever (2013) (above the line) with slightly different hyper-parameters. The vocabulary cutoff m of the source embedding is specified for each word2vec model we trained. (The tables in this section are unfortunately not directly comparable, as some hyper-parameters may differ, but we hope that the individual message of each table is clear.)

7.4.4.3 *Comparison of results in the two tasks*

In Figure 19, we show the results of some Hungarian VSMs in the analogical and the word translation task plotted against each other. The horizontal axis shows precision in the semantic analogical questions, while the vertical axis shows precision (@5) in proto-dictionary gen-

13 http://nlp.stanford.edu/projects/glove/

	prec@1	prec@5
en → sp	33	51
sp → en	35	52
en → cz	27	47
cz → en	23	42
en → vn	10	30
vn → en	24	40
GloVe-sl → en rs	44.80	63.40
word2vec-sl → en $m = 100$ rs	41.70	60.40
word2vec-hu → en $m = 50$ rst	32.80	54.70
word2vec-lt → en $m = 100$ rt	21.20	36.50

Table 43: Results in proto-dictionary collection

source word	cos	translations			
öt	0.9101	five	six	eight	three
jó	0.8961	good	really	too	very
de	0.8957	but	though	even	just
bár	0.8955	though	but	even	because
hit	0.8904	faith	belief	salvation	truth
ugyan	0.8880	though	but	even	because
vöröshagymát	0.8878	onion	garlic	onions	tomato

Table 44: Example word translations. cos is the cosine similarity of the image of the source word vector by the learned mapping and the nearest target vector. Words in the target language are listed in the (descending) order of their similarity to the image vector.

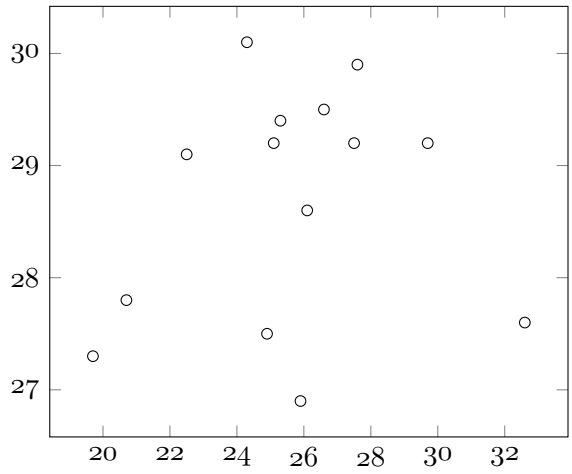

Figure 19: The precision of some of our models in monolingual (horizontal axis) vs bilingual (vertical axis) task. Unfortunately, the choice of the hyper-parameters was not systematic.

eration to the Google News model[14] restricted to words accepted by Hunspell and using seed pairs collected with wikt2dict. It can be seen that the result in the two tasks are unfortunately uncorrelated.

7.4.5 *Parameter analysis*

7.4.5.1 *Corpus*

QUALITY In Table 45, we compare how our models trained on two different corpora shine in analogical questions. The corpora are the Hungarian National Corpus v2.0.3 (Oravecz, Váradi, and Sass 2014), which is a curated corpus of Hungarian, and the Hungarian Webcorpus (Halácsy et al. 2004) that is a similarly sized webcorpus. The numbers suggest that a word embedding trained on a curated corpus represents analogical relations better, especially the semantic part, or when GloVe is used.

SIZE Table 46 shows how the performance depends on the size of the corpus. It is clear that a much larger corpus is needed to answer semantic questions.

7.4.5.2 *word2vec, LBL4word2vec and GloVe*

We compared word2vec, LBL4word2vec, and GloVe (with two parameter settings) in the analogical task. LBL4word2vec[15] implements the ideas in Mnih and Kavukcuoglu (2013), what makes it more parameter-efficient. The two parameter settings were needed for a fair comparison, because

14 https://code.google.com/p/word2vec/#Pre-trained_word_and_phrase_vectors
15 https://github.com/qunluo/LBL4word2vec

model	question type	Webcorpus		HNC	
word2vec	morphological	**54.9**	**2924**/5326	51.8	2856/5514
	semantic	8.3	40/482	**16.0**	**123**/769
	total	**51.0**	2964/5808	47.4	**2979**/6283
GloVe	morphological	47.4	2525/5326	**48.2**	**2658**/5514
	semantic	9.3	45/482	**14.4**	**111**/769
	total	**44.2**	2570/5808	44.1	**2769**/6283

Table 45: Comparison of results on two different corpora. The denominator of each fraction is the number of questions with all three words known to the vector model, while the numerator is the number of correct answers for these questions. Parameters: $d = 152$, $m = 10$, $i = 5$ in both models. For word2vec, $w = 5$ and $n = 5$ while for GloVe, $w = 3$. The different window sizes mean that these results are not suitable for comparing the models just the corpora.

	morph		sem		total	
1M	1.8	58/3256	0.0	0/84	1.7	58/3340
2M	6.1	191/3130	0.0	0/60	6.0	191/3190
10M	24.9	986/3954	7.4	8/108	24.5	994/4062
100M	55.1	2530/4594	31.4	37/118	54.5	2567/4712
754M	63.2	3486/5514	49.8	383/769	61.6	3869/6283

Table 46: The effect of corpus size

the default (recommended) values of d, w, i and m are different in the two architectures: see Table 47, where the more computation-intensive setting is in bold.

We trained two models with each architecture on HNC: a small one with the less computation-intensive one of the two default values (except for using $d = 52$ for historical reasons) and a big one with the more costly one. For the number of negative samples, which is specific for word2vec, we use the default $n = 5$. See results in Table 48. Note that GloVe could be further improved by taking the average of the two vectors, the "focus" and context vector learned by the model for each word (see Section 4.2.6).

	word2vec	GloVe
d	**100**	50
w	5	**15**
i	5	**25**
m	**5**	10

Table 47: Default values of parameters shared by word2vec and GloVe

		morph		sem		total	
small	word2vec sgram	49.0%	2703	20.3%	156	45.5%	2859
	LBL4word2vec sgram	46.6%	2567	19.4%	149	43.2%	2716
	word2vec cbow	49.9%	2751	15.7%	121	45.7%	2872
	glove	41.3%	2277	11.1%	85	37.6%	2362
big	word2vec sgram	57.8%	3186	42.0%	323	55.8%	3509
	LBL4word2vec sgram	55.5%	3058	36.3%	279	53.1%	3337
	glove	58.1%	3206	31.3%	241	54.9%	3447
	word2vec cbow	57.8%	3187	30.7%	236	54.5%	3423

Table 48: Comparison of models trained in different architectures. Rows within each model "size" are sorted by precision in the semantic task, which we consider more relevant to lexicography than morphology. The total number of questions that do not contain out-of-vocabulary words is 5514 in morphological questions and 6283 in semantic ones.

	morph		semant		total	
cbow $hs = 0, n = 5$	59.4%	**3276**/5514	24.1%	185/769	55.1%	3461/6283
cbow $hs = 1, n = 0$	49.0%	2702/5514	13.9%	107/769	44.7%	2809/6283
cbow $hs = 1, n = 5$	49.5%	2730/5514	14.3%	110/769	45.2%	2840/6283
sgram $hs = 0, n = 5$	59.1%	3261/5514	33.6%	**258**/769	56.0%	3519/6283
sgram $hs = 1, n = 0$	49.8%	2744/5514	23.1%	178/769	46.5%	2922/6283
sgram $hs = 1, n = 5$	50.4%	2781/5514	23.1%	178/769	47.1%	2959/6283

Table 49: Hierarchical softmax (HS) and negative sampling

7.4.5.3 *word2vec: Hierarchical softmax and negative samples*

Hierarchical softmax (HS) and negative sampling are alternative solutions for the partition function problem (Section 4.2.2). (We already used the latter in Section 7.1.) Nevertheless, in Makrai (2015) we also tried whether the two can be combined to get better result than with either of the techniques. A negative answer can be seen in Table 49 (HNC, $d = 100, w = 5, i = 5, m = 5$).

7.4.5.4 *proto-dictionaries: Seed dictionary*

We compare the results obtained in the proto-dictionary generation task with different English-Hungarian seed dictionaries in Table 50. The source language model is always `glove.840B.300d`, the target model is also a GloVe model trained on HNC ($d = 300, m = 1, w = 15, i = 25$). For details of the seed dictionaries see Section 7.4.3.2.

seed dictionary	prec@1	prec@5
Europarl	17.70%	34.10%
wikt triang	13.10%	25.30%
wikt	12.50%	25.40%
OpenSubtitles2012	10.30%	23.40%
efnilex12 en→hu	10.10%	23.80%

Table 50: Accuracy of proto-dictionary generation with different seed dictionaries

7.4.6 *Gluten-free embeddings for better analogical relations*

Attila Novák (personal communication) proposed to test whether gluten-free word embeddings better represent analogical relations. Indeed, Nemeskey (2017) obtained better language modeling scores (perplexity values) with the gluten-free method than with the vanilla word-level model, but, as far as we know, the method was not evaluated in word analogies. Döbrössy et al. (2019, Section 4.2.11.2) evaluated other sub-word methods.

Of course, the offset method is vacuous for most of the grammatical relations (gram1-adjective-to-adverb, gram3-comparative, gram4-superlative, gram8-plural-noun and gram9-plural-verb) as the corresponding morphemes (e.g. *-an/en* '-ly') are represented separately. Nevertheless, at least some of the derivatives in a few relations (gram2-opposite, gram5-present-participle, gram6-nationality-adjective, and gram7-past-tense) are considered non-productive or even non-compositional by emMorph, so they still allow this kind of analysis.

This experiment is somewhat retrospective: we evaluate static word embeddings for a medium-resourced language in the age of deep, contextualized language models. A couple of years ago, obtaining a clean gigaword corpus of Hungarian was difficult, but the introduction of Webcorpus 2.0 (Nemeskey 2020) led to a somewhat new situation. For that reason, we conduct experiments on increasing slices of the new webcorpus. Webcorpus 2.0 is published in the form of a few thousand files. We start with the degenerate case of a single file `2017_2018_2956.tsv.gz`, then consider the files whose name starts with `2017_2018_295`, `2017_2018_29`, or `2017_2018_2`, which mean more and more files. The size of a typical corpus ten years ago, in the golden age of static word embeddings, used to be between the last two, `2017_2018_29` and `2017_2018_2`. In addition, we repeated our experiments on all the Wikipedia files for comparison.

We saw in Section 7.4.1 that Mikolov-style analogical benchmarks, both the English ones and that proposed by Makrai (2015), consist of different relations. In the gluten-free experiments, we found that the capital oriented relations, the more limited capital-common-countries

	files	2017_2018_2956*	2017_2018_295*	wiki*	2017_2018_29*	2017_2018_2*
corpus size	sentence	66 798	879 317	11 798 744	12 058 626	133 832 368
	token	1 246 317	23 261 396	200 357 761	310 703 112	3 395 599 678
	type	9 388	149 981	920 965	1 028 343	4 730 753
all	paradigm-cell	3.33%	4.85%	30.10%	33.57%	50.56%
	each-separate	1.67%	5.05%	33.29%	31.80%	51.96%
	word-level	0.06%	7.33%	46.18%	46.97%	57.94%
semantics	paradigm-cell	10.00%	7.94%	17.10%	32.01%	33.49%
	each-separate	5.00%	7.94%	24.68%	24.68%	34.59%
	word-level	0.00%	10.00%	28.89%	34.14%	47.78%
morph.	paradigm-cell	0.00%	2.53%	37.91%	34.75%	60.80%
	each-separate	0.00%	2.88%	38.46%	37.14%	62.39%
	word-level	0.07%	6.33%	52.66%	51.78%	61.74%

Table 51: Mean accuracy scores of word-level and gluten-free (each-separate and paradigm-cell) word embeddings in word analogies. Each row corresponds to a word embedding trained on a different part of Webcorpus 2.0. The left pane shows size statistics of the corpora.

with its 20 atomic relations in Hungarian and especially the more general capital-world with its 166 countries, are instable. The non-Wikipedia corpora cover at most 6 of the "common" atomic relations (i.e. both the country and the capital), and 22 of the broader selection, which makes the results for these relations unreliable. For this reason, in the following we disregard these two relations.

The results can be seen in Table 51 and Figure 20, along with size statics of the corpora slices in the former. Surprisingly, we see that deglutination makes the representation of analogical relations worse. This may be due to the fact that deglutination decreases the effective window size. It would be possible to measure whether the smaller effective window size is really the cause: One could widen the window according to the average number of chunks in words. But our understanding is that this would not advance the matter: at best, we would only show what is already obvious, that with the same parameters (or, what is the same, equal computational demand), deglutination does not help. The semantic/world-knowledge results on the greatest corpus slice are still worse than those reported by Mikolov, Le, and Sutskever, which suggests that the Hungarian questions are more difficult.

Nevertheless, corpus size saves the day to the extent that with all methods (word-level and both variants of deglutination), scores increase with corpus size. The only exception for this is that with each-separate the 310K-word corpus results in slightly worse word embeddings than Wikipedia, which is smaller. This can be explained with the high quality of Wikipedia as a corpus. Which variant of deglutination is better changes inconsistently between corpus sizes.

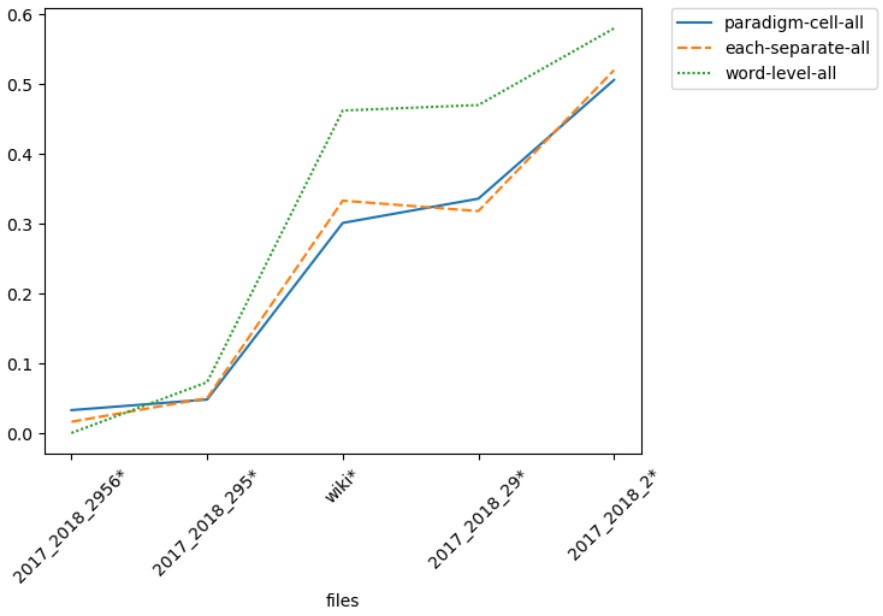

Figure 20: Mean accuracy scores of word-level and gluten-free (each-separate and paradigm-cell) word embeddings in word analogies. Points on the horizontal axis correspond to word embedding trained on a different parts of Webcorpus 2.0.

7.4.7 *Conclusion*

We have created the Hungarian equivalent of the analogical benchmark, which is one of the main evaluations test-beds of static word vectors. We trained Hungarian word embeddings and tested how they representation analogies. To the best of our knowledge, this made Hungarian the third language (after English and Turkish), where analogy was tested. The result for morphological analogies is positive, while weak in semantic ones. Besides, we extended the linear translation method to the GloVe model and applied it to medium-resourced European languages.

7.5 SMOOTHED TRIANGULATION FOR LEXICAL INDUCTION

Triangulation infers word translations in a pair of languages based on translations to other, typically better resourced ones called *pivots*. This method may introduce noise if words in the pivot are polysemous. The reliability of each triangulated translation has traditionally been estimated by the number of pivot languages (Tanaka and Umemura 1994).

As we have seen in Section 7.4.2, and will return to in Chapter 8, Mikolov, Le, and Sutskever (2013) introduced a method for generating or scoring word translations. Translation is formalized as a linear mapping between distributed vector space models (VSM) of the two languages. VSMs are trained on unlabeled monolingual data, while the mapping is learned in supervised fashion, using a seed dictionary of

some thousand word pairs. The mapping can be used to associate existing translations with a real-valued *similarity score*.

In this section, which originally appeared as Makrai (2016), we apply linear mapping to filter triangulated translations, and show that scores by the mapping are a smoother measure of merit than the number of pivots. Theoretically, smoothness can be interpreted as that there is some extra noise in triangulation that is eliminated by linear translation. The methods we use are language-independent, and the training data is easy to obtain for many languages. For the line of research reported in this section, we chose the German-Hungarian pair for evaluation, in which the filtered triangles resulting from our experiments were, to the best of our knowledge, the greatest freely available list of word translations by the time.

7.5.1 *Introduction*

Word translations arise in dictionary-like organization as well as via machine learning from corpora. The former is exemplified by Wiktionary, a crowd-sourced dictionary with editions in many languages. Ács, Pajkossy, and Kornai (2013) obtain word translations from Wiktionary with the pivot-based method, also called triangulation, that infers word translations in a pair of languages based on translations to other, typically better resourced ones called pivots. Triangulation may introduce noise if words in the pivot are polysemous. The reliability of each triangulated translation is traditionally estimated by the number of pivot languages (Tanaka and Umemura 1994).

The project reported in this section exploits human labor in Wiktionary combined with distributional information in VSMs. We train VSMs on gigaword corpora, and the linear translation mapping on direct (non-triangulated) Wiktionary pairs. This mapping is used to filter triangulated translations based on the similarity scores. The motivation is that scores by the mapping may be a smoother measure of merit than considering only the number of pivots for the triangle. We evaluate the scores against dictionaries extracted from parallel corpora (Tiedemann 2012). In running automatic word alignment on a parallel data set, Tiedemann (2012) used "GIZA++ (Och and Ney 2003) and the symmetrization heuristics (`grow-diag-final-and`) implemented in Moses (Koehn et al. 2007) to extract probabilistic phrase tables", which are standard in statistical machine translation. We show that linear translation really provides a more reliable method for triangle scoring than pivot count.

The methods we use are language-independent, and the training data is easy to obtain for many languages. We chose the German-Hungarian pair for evaluation, in which the filtered triangles resulting from our experiments are the greatest freely available list of word translations we are aware of.

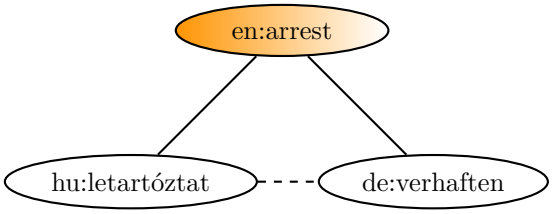

Figure 21: Triangulation

7.5.2 *Triangulation*

A method for creating dictionaries is triangulation through better resourced languages called the *pivot* (Tanaka and Umemura 1994). The idea is that if the English translation of the Hungarian word *letartóztat* is *arrest*, and the German translation of *arrest* is *verhaften*, then the German translation of *letartóztat* should be *verhaften*, see Figure 21.

Triangles are corrupted by ambiguity in the pivot word (the one in the middle): German *Dose* can be translated as *can* to English (as a synonym of *tin*), which, as a verb, translates to *tud* in Hungarian, which is unrelated to *Dose*. Saralegi, Manterola, and Vicente (2011) analyze two methods for pruning wrong triangles: one is based on exploiting the structure of the source dictionaries, and the other one is based on an estimate of distributional similarity acquired from comparable corpora. The project reported in this section is more similar to the latter in that it uses distributional information, but in the framework of neural language modeling.

7.5.3 *Linear translation*

As we already mentioned in Section 4.2.4, Mikolov, Le, and Sutskever (2013) discovered that VSMs of different languages have such similarities that a linear transformation can map representations of source language words to the representations of their translations. The reader who remembers well, even to the confidence score provided by the mapping, can safely skip the following paragraph.

The method belongs to the paradigm of supervised machine learning: specifically it makes use of a great amount of *monolingual* data i.e. gigaword corpora for training, plus a seed dictionary of some thousand words for supervision. Mikolov, Le, and Sutskever formalize translation as linear mapping $W \in \mathbb{R}^{d_2 \times d_1}$ from the source (monolingual) VSM $\mathbb{R}^{d_1}$ to the target one $\mathbb{R}^{d_2}$: the translation $z_i \in \mathbb{R}^{d_2}$ of a source word $x_i \in \mathbb{R}^{d_1}$ is approximately its image Wx_i by the mapping. The translation model is trained with linear regression on the seed dictionary

$$\min_W \sum_i ||Wx_i - z_i||^2,$$

and can be used to collect translations for the whole vocabulary (by choosing z_i to be the nearest neighbor of Wx_i) or to score a translation z coming from some other source (with the score being the distance between Wx_i and z_i).[16] In the original setting of the collection mode, evaluation is done on another thousand seed pairs.

As we introduced in Section 4.2.14, a common error in linear translation is when there are *hubs*, i.e. target words that are returned as the translation of many words, which is wrong in most of the cases. Dinu, Lazaridou, and Baroni (2015) propose a method for downplaying the importance of such target words they call *global correction*. Our experiments here use this method. We return to this problem in Section 8.4.1 in more detail.

7.5.4 *Data*

Direct and triangulated Wiktionary translations were extracted with wikt2dict (Ács, Pajkossy, and Kornai 2013)[17] that handles 43 editions of Wiktionary.

The German VSMs have been trained on SdeWaC (Baroni et al. 2009) and the Hungarian one on the concatenation of the Hungarian Webcorpus (Halácsy et al. 2004) and the Hungarian National Corpus (Oravecz, Váradi, and Sass 2014) with word2vec[18] (Mikolov, Chen, et al. 2013).[19]

For training and using the linear mapping, we forked[20] the implementation by Dinu, Lazaridou, and Baroni (2015). The German to Hungarian mapping was trained on the 5K direct word pairs that are supported by the most pivots in Wiktionary. All the triangles were scored. The Hungarian word embedding (and some glue code we wrote for this project) is freely available[21].

The scoring was evaluated against a dictionary in the OPUS project[22] that has been extracted by Tiedemann (2012) from the OpenSubtitles2013 parallel corpus, a collection of translated movie subtitles[23]. OpenSubtitles2013 contains 59 languages. The sizes of the German–Hungarian subsection are shown in Table 52.

16 Mikolov et al. use a surprising combination of vector distances, Euclidean distance in training and cosine similarity (and distance) in collection (and, respectively, scoring) of translations. This choice is theoretically unmotivated, but we (Makrai 2015) also found it to work better than more consistent combinations of metrics. However, see Xing et al. (2015) for opposing results. We return to this topic in Section 8.5.5.

17 https://github.com/juditacs/wikt2dict

18 https://code.google.com/p/word2vec/

19 The German VSM has been a continuous bag of words model in 300 dimensions (infrequent words have been cut off at 100 occurrences), the Hungarian one a 600 dimensional one (with a cut-off of 10). The choice of hyper-parameters was not fully systematic.

20 https://github.com/makrai/dinu15/

21 https://github.com/makrai/efnilex-vect

22 http://opus.lingfil.uu.se/

23 http://www.opensubtitles.org/

documents	3208
sentences	3.2 M
German tokens	23.3 M
Hungarian tokens	19.7 M
extracted word pairs	29.1 K

Table 52: The German Hungarian subsection of the OpenSubtitles2013 parallel corpus (Tiedemann 2012)

Most of our training data are general in their *domain*: web corpora (SdeWaC, the Hungarian Webcorpus), a curated corpus (the Hungarian National Corpus, as far as a corpus of 754 million words may be curated), and a crowd-sourced but otherwise standard dictionary (Wiktionary). One may ask whether the domain of the reference dictionary extracted from movie subtitles is general to an appropriate extent, or how far a problem of domain mismatch between train and test may arise. We hypothesize that the mismatch is negligible.

7.5.5 *Evaluation*

We evaluated the vector-based scoring of triangulated translational word pairs (*triangles*) in comparison with a dictionary created from the parallel corpus OpenSubtitles2013. For each (German) word, we consider as gold translations all the (Hungarian) words that are listed in the OpenSubtitles2013 dictionary as a translation. Note that the gold dictionary is automatically generated. A traditional dictionary would be more reliable, but these are difficult to obtain, so we opted for a surrogate.

For evaluation, we sort the triangles in two orders: as baseline, by the number of pivots for the triangle, and more importantly, by the score in the linear mapping (cos). Then in each order, we compute accuracy (the ratio of correct translations among the whole set of translations) on each 1000-word slice of the list (e.g. triangles 1–1000, then 1001–2000, etc.) taking OpenSubtitles2013 translations as gold.

While the overall accuracy of the linear scoring[24] (8.58%) is slightly worse than that of pivot counting (9.32%), Figure 22 suggests that in the order by cos, accuracy descends more smoothly than in the order by pivot count. (The last 22.73% of the nearly 160 K triangles is out of the vocabulary of one or both of the VSMs, so cos cannot be computed.) Now we turn to a more quantitative support of this visual analysis.

24 By accuracy, we mean the ratio of good translations among all the words that have to be translated.

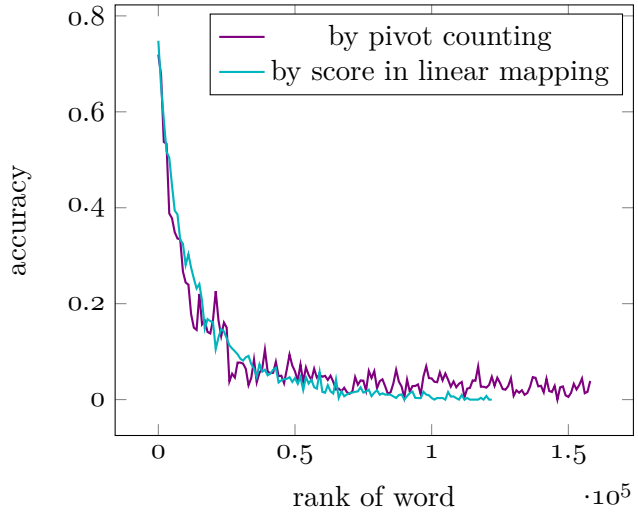

Figure 22: Accuracy curve of triangles sorted by their pivot count as baseline, or score in linear translations (cos). The later is smoother.

scoring method	exp	power law
pivot counting	6.1859e-04	5.2182e-04
linear mapping	**2.4574e-04**	**1.1789e-04**
ratio	2.51	4.42

Table 53: The mean squared error of fitting parametric curves to the accuracy values obtained by translation scoring methods. Linear mapping produces a smoother accuracy decay than pivot counting.

7.5.6 *Quantitative analysis of smoothness*

We measure the smoothness of the accuracy curves by how well they can be approximated by a function in some parametric family, see Figures 23 to 26. We tried two families with similar results. The fist family is exponential functions of the form

$$a \cdot \exp(-bx) + c,$$

where x is the index of the vocabulary slice (0 for words 0–1000, 1 for 1001–2000, etc), and a, b, and c are parameters to fit. The second family is that of power-law functions

$$a \cdot (bx + c)^k,$$

where k is another parameter to fit, and the remaining variables play similar roles as in the exponential case. The error of the fit (i. e. the lack of smoothness) is quantified as the mean squared error (MSE, not to be confused with multi-sense word embeddings, which are the topic of Chapter 8) between the two curves.

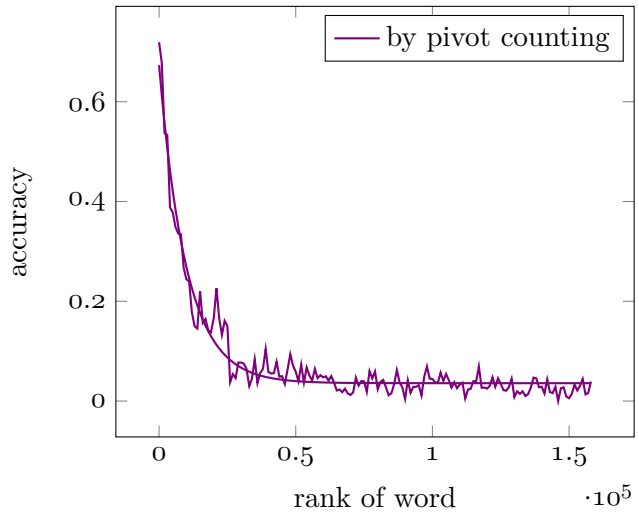

Figure 23: The accuracy curve of pivot counting approximated by an exponential function.

The MSE of the two accuracy curves (scoring translations by pivot counting or cosine score) approximated by the two families (exponential or power-law functions) are shown in Table 53. The MSE of the accuracy curve in pivot counting is 2.51 (resp. 4.42) times more than that in scoring by the linear mapping, when both curves are modeled as exponential (resp. power-law) functions. It is probably also worth mentioning that if we take the 20–30 000 words with the greatest confidence with the two methods, the accuracy is slightly better with the proposed method than in the baseline, see especially Figures 27 and 28.

7.5.7 *Conclusion*

We enhanced triangulation, a traditional method in dictionary induction, by computing a reliability measure for word pairs with linear translation, which is a smoother method than counting the triangulation pivots. We showcased the method by creating the then largest[25] freely available German-Hungarian dictionary. All the 159 K word pairs were published.

25 In 2023, when this thesis is submitted, the largest Hungarian–German word list (Schwenk et al. 2019) in OPUS consists of 265 K pairs.

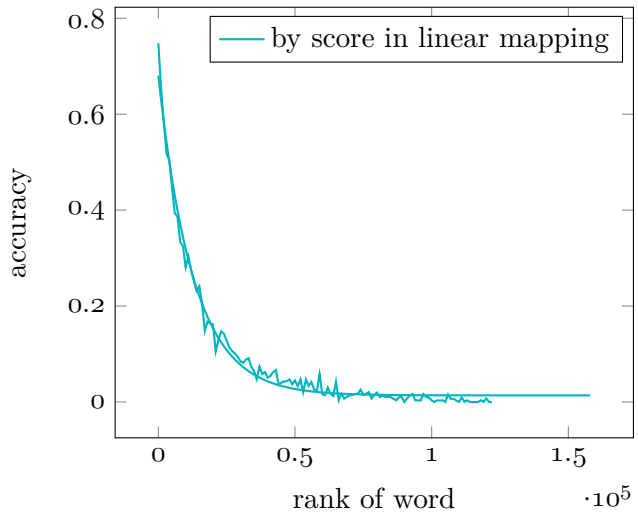

Figure 24: The accuracy curve of scores by the linear mapping approximated by an exponential function.

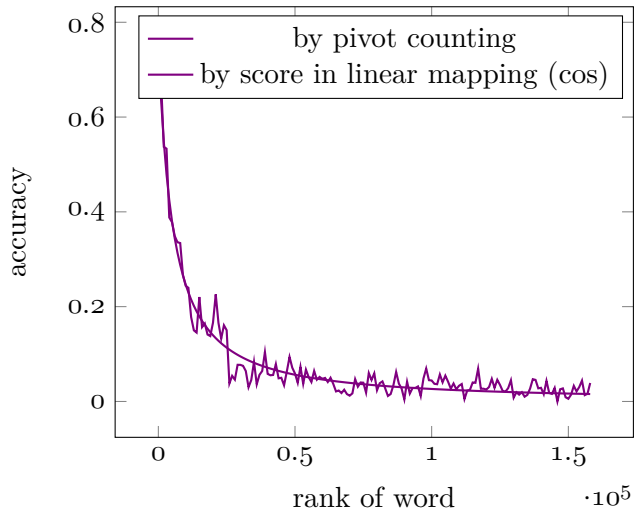

Figure 25: Accuracy curves of scores by pivot count approximated by power-law functions.

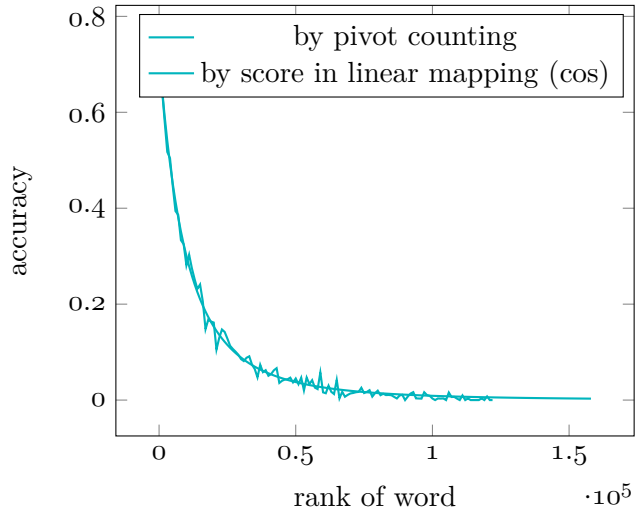

Figure 26: Accuracy curves of scores by the linear mapping approximated by power-law functions.

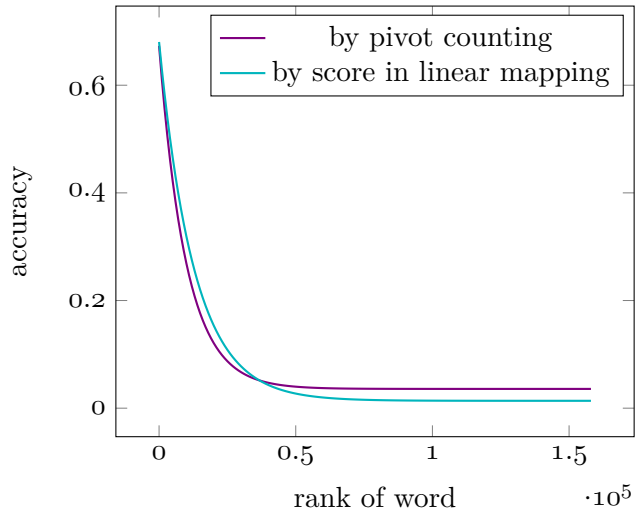

Figure 27: The exponential approximations of the accuracy curves.

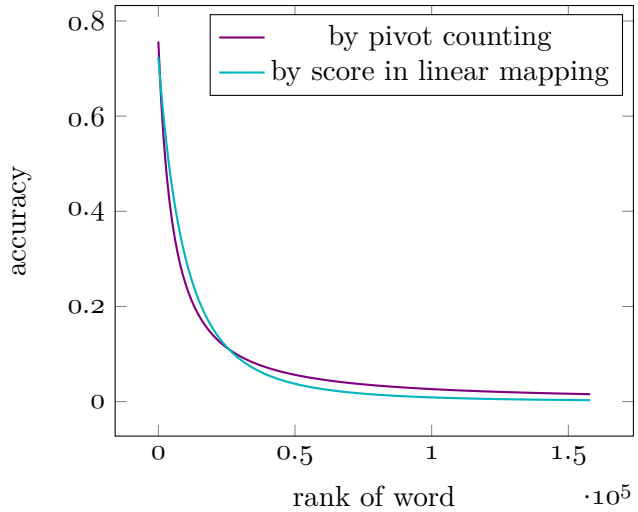

Figure 28: The power-law approximations of the accuracy curves.

The name of the song is called "Haddocks' Eyes."'
'Oh, that's the name of the song, is it?' Alice said, trying to feel interested.
'No, you don't understand,' the Knight said, looking a little vexed. 'That's what the name is called. The name really is "The Aged Aged Man."'
'Then I ought to have said "That's what the song is called"?' Alice corrected herself.
'No, you oughtn't: that's quite another thing! The song is called "Ways and Means": but that's only what it's called, you know!'
'Well, what is the song, then?' said Alice, who was by this time completely bewildered.
'I was coming to that,' the Knight said. 'The song really is "A-sitting On A Gate": and the tune's my own invention.'

— Lewis Carroll

8

CROSS-LINGUAL WORD SENSE INDUCTION

8.1 DO MULTI-SENSE EMBEDDINGS LEARN MORE SENSES?

Word ambiguity poses a significant challenge for NLP. Contextualized word representations, such as those provided by deep language models (Section 4.3), have undoubtedly revolutionized many natural language processing (NLP) tasks by capturing the context-dependent meanings of words. Nevertheless, one of the fundamental aspects of language understanding is capturing lexical semantics, i.e. the meanings and relationships of lexemes. Context-independent representations, such as static word embeddings, remain valuable for certain tasks, especially for theoretical ones involving psycholinguistics and beyond. Now we turn to this topic and its connection to multilinguality and translation.

Multi-sense word embeddings (MSEs) have modeled different meanings of word forms with different (static) vectors since before the advent of deep language models/contextualized word representations (Sec-

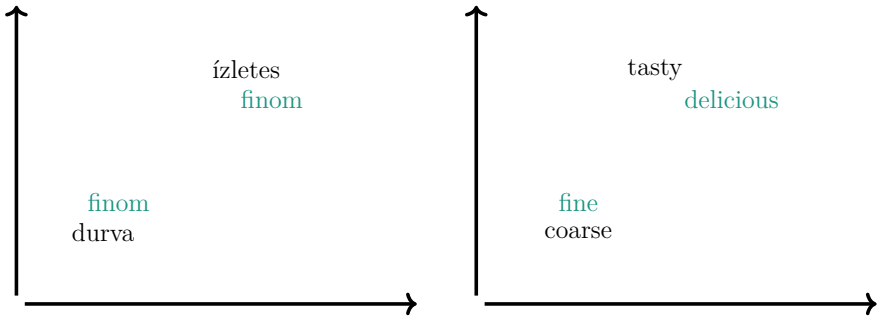

Figure 29: Linear translation of word senses. The Hungarian word *finom* is ambiguous between 'fine' and 'delicious'.

tion 4.3). In this final section of the thesis, which originally appeared as Borbély, Makrai, Nemeskey, and Kornai (2016) and Makrai and Lipp (2018)[1], we propose a method for evaluating MSEs by their degree of *semantic resolution,* measuring the detail of the sense clustering, more precisely their precision as the detectors of word ambiguity. The method exploits the principle that words may be ambiguous as far as the postulated senses translate to different words in some other language. (Specifically the method does not detect false negatives: words that are actually ambiguous, but the system attributes them a single vector.) Besides, in the context of embedding-based dictionary induction, we also test whether the orthogonality constraint and related vector preprocessing techniques help in reverse nearest neighbor search. These more technical questions receive a negative answer.

8.2 TOWARDS A LESS *DELICIOUS* INVENTORY

Word sense induction (WSI) is the task of discovering senses of words without supervision (Schütze 1998). The goal of WSI can be set at two levels. We may more modestly aim to distinguish homophony from polysemy (see Section 2.3.7). Ideally, we could even differentiate between metonymy and metaphor, two subtypes of polysemy. Approaches include multi-sense word embeddings (MSEs), i.e. vector space models of word distribution with more vectors for ambiguous words. In MSEs, each vector is supposed to correspond to a different word sense, but in practice, models frequently have different sense vectors for the same word form without an interpretable difference in meaning.

1 The 2016 paper measured the sense granularity with two methods: Section 2 was based on computer readable lexica, and Section 3 presented the multilingual method. The former was the work of Nemeskey. The latter is the joint work of Borbély and Makrai, with equal contribution.
In the 2018 paper, Veronika Lipp wrote a section on the different kinds of polysemy, which is not included in this thesis. In the remainder of this second paper, which appropriately corresponds to that from Section 8.4.1 in this thesis, Makrai went on alone to elaborate the multilingual method.

Our first publication in this topic (Borbély, Makrai, et al. 2016) appeared at the 1st Workshop on Evaluating Vector-Space Representations for NLP. In a progammatic paper of the workshop, Gladkova and Drozd (2016) called polysemy "the elephant in the room" as far as evaluating embeddings are concerned. We attacked this problem head on, by proposing a method for evaluating multi-sense word embeddings (MSEs).

We emphasize at the outset that our evaluation proposal probes an aspect of MSEs, *semantic resolution*, which is not well measured by the well-known word sense disambiguation (WSD) task that aims at classifying occurrences of a word form to different elements of a sense inventory pre-defined by some experts. Our goal is to probe the granularity of the inventory itself.

As we discussed in Section 3.1.1, the central linguistic/semantic/psychological property we wish to capture is that of a *concept*, the underlying word sense unit. To the extent standard lexicographic practice offers a reasonably robust notion (this is of course debatable, but we consider a straight correlation of 0.27 and and a frequency-effect-removed correlation of 0.60 over a large vocabulary[2] a strong indication of consistency), this is something that MSEs should aim at capturing. We expect that the inter-dictionary (inter-annotator) agreement can be improved considerably by (manual or automated) alignment of word senses in different dictionaries, to provide a more robust gold standard.

The differentiation of word senses is fraught with difficulties, especially when we wish to distinguish homophony, using the same written or spoken form to express different concepts, such as Russian *mir* 'world' and *mir* 'peace' from polysemy, where speakers feel that the two senses are very strongly connected, such as in Hungarian *nap* 'day' and *nap* 'sun'. To quote Zgusta (1971) "Of course it is a pity that we have to rely on the subjective interpretations of the speakers, but we have hardly anything else on hand". Etymology makes clear that different languages make different lump/split decisions in the conceptual space, so much so that translational relatedness can, to a remarkable extent, be used to recover the universal clustering (Youn et al. 2016).

One of the confounding factors is part of speech (POS, recall Section 3.1.2). Very often, the entire distinction is lodged in the POS, as in *divorce* (noun) and *divorce* (verb), while at other times this is less clear, compare the verbal *to bank* 'rely on a financial institution' and *to bank* 'tilt'. Clearly the former is strongly related to the nominal *bank* 'financial institution' while the semantic relation 'sloping sideways' that connects the tilting of the airplane to the side of the river is somewhat less direct, and not always perceived by the speakers. The Collins-COBUILD (CED, Sinclair (1987)) dictionary starts with the se-

2 These results are published in the same Borbély, Makrai, et al. (2016), but this thesis does not discuss them in detail, because they were conducted mainly by Dávid Nemeskey.

mantic distinctions and subordinates POS distinctions to these, while the Longman dictionary (LDOCE, Boguraev and Briscoe (1989)) starts with a POS-level split and puts the semantic split below. Of the Hungarian lexicographic sources, the Comprehensive Dictionary of Hungarian (NSZ, Ittzés (2011)) is closer to CED, while the Explanatory Dictionary of Hungarian (EKSZ, Pusztai (2003)), is closer to LDOCE in this regard.

Our method is based on the principle that words may be ambiguous to the extent to which their postulated senses translate to different words in some other language. For the translation of words, we applied the method by Mikolov, Le, and Sutskever (2013) who train a translation mapping from the source language embedding to the target as a least-squares regression supervised by a seed dictionary of the few thousand most frequent words. The translation of a source word vector is the nearest neighbor of its image by the mapping in the target space. In the multi-sense setting, we have translated from MSEs. (The target embedding remained single-sense, as that is sufficient for our purposes, and this way we restrict the less know effects of multi-sense modeling to the model under evaluation.)

Section 8.3 introduces MSEs. In section 8.4, we elaborate on the cross-lingual evaluation. Part of the evaluation task is to decide on empirical grounds whether different good translations of a word are synonyms or translations in different senses. Reverse nearest neighbor search, the orthogonality constraint on the translation mapping, and related techniques are also discussed. Section 8.5 offers experimental results with quantitative and qualitative analysis. It should be noted that our evaluation is not very strict, but rather a process of looking for something conceptually meaningful in these unsupervised MSE models. We make our Hungarian multi-sense embeddings[3] and the code for these experiments[4] available on the web.

8.3 MULTI-SENSE WORD EMBEDDINGS

Vector-space language models with more vectors for each meaning of a word originate from Reisinger and Mooney (2010). Huang et al. (2012) trained the first neural-network-based MSE. Both works use a uniform number of clusters for the subset of words that they select before training as potentially ambiguous. The first system with adaptive sense numbers and an effective open-source implementation is a modification of skip-gram (Mikolov, Sutskever, et al. 2013), *multi-sense* skip-gram by Neelakantan et al. (2014), where new senses are introduced during training by thresholding the similarity of the present context to earlier contexts.

3 https://hlt.bme.hu/en/publ/makrai17
4 https://github.com/makrai/wsi-fest

Bartunov et al. (2016) and Li and Jurafsky (2015) improve upon the heuristic thresholding by formulating text generation as a Dirichlet process. In `AdaGram` (Bartunov et al. 2016), senses may be merged as well as allocated during training. `mutli`-sense skip-gram[5] (Li and Jurafsky 2015) applies the Chinese restaurant process formalization of the Dirichlet process. `neela`, `AdaGram`, and `mutli` have a parameter each for semantics resolution (more or less senses): λ, α, and γ, respectively.

MSEs are still in the research phase: Li and Jurafsky (2015) demonstrate that, when hyper-parameters are carefully controlled for, MSEs introduce a slight performance boost in semantics-related tasks (semantic similarity for words and sentences, semantic relation identification, part-of-speech tagging), but similar improvements can also be achieved by simply increasing the dimension of a single-sense embedding.

8.4 LINEAR TRANSLATION FROM MSES

As we already discussed in Sections 7.4.2 and 7.5.3, Mikolov, Le, and Sutskever (2013) discovered that embeddings of different languages are so similar that a linear transformation can map vectors of the source language words to the vectors of their translations. The reader who remembers well, even to the confidence score provided by the mapping, can safely skip the following paragraph.

The method uses a seed dictionary of a few thousand words to learn translation as a linear mapping $W \ \mathbb{R}^{d_1} \rightarrow \mathbb{R}^{d_2}$ from the source (monolingual) embedding to the target: the translation $z_i \in \mathbb{R}^{d_2}$ of a source word $x_i \in \mathbb{R}^{d_1}$ is approximately its image Wx_i by the mapping. The translation model is trained with linear regression on the seed dictionary

$$\min_W \sum_i ||Wx_i - z_i||^2$$

and can be used to collect translations for the whole vocabulary by choosing z_i to be the nearest neighbor of Wx_i. We follow Mikolov, Le, and Sutskever (2013) in (i) using different metrics, Euclidean distance in training and cosine similarity in collection of translations, and in (ii) training the source model with approximately three times greater dimension than that of the target embedding.

In a multi-sense embedding scenario, Borbély, Kornai, Makrai, and Nemeskey (2016) take an MSE as the source model, and a single-sense embedding as target. The quality of the translation has been measured by training on the most frequent 5k word pairs and evaluating on another 1k seed pairs. Details of the data will be specified in Section 8.5.1.

5 Note the $l \leftrightarrow t$ metathesis in the name of the repo which is the only way of distinguishing it from the other two multi-sense skip-gram models.

8.4.1 *Reverse nearest neighbor search*

As we introduced in Section 4.2.14, a common problem when looking for nearest neighbors (NNs) in high-dimensional spaces (Radovanović, Nanopoulos, and Ivanović 2010; Suzuki et al. 2013; Tomašev N. 2013), and especially in embedding-based dictionary induction (Dinu, Lazaridou, and Baroni 2015; Lazaridou, Dinu, and Baroni 2015) is when there are *hubs*, data points (target words) returned as the NN (translation) of many points (Wxs), resulting in incorrect hits (translations) in most of the cases. Dinu, Lazaridou, and Baroni (2015) attack the problem with a method they call *global correction*. Here, instead of the original NN, which we will call *forward* NN search to contrast with the more sophisticated method, they first rank source words by their similarity to target words. In *reverse* nearest neighbor (rNN[6]) search, source words are translated to the target words to which they have the lowest (forward) NN rank.[7]

In reverse NN search, we restricted the vocabulary to some tens of thousands of the most frequent words. We introduced this restriction for saving memory, because the $|V_{sr}| \times |V_{tg}|$ similarity matrix has to be sorted column-wise for forward and row-wise for reverse ranking, so at some point of the computation we keep the whole integer matrix of forward NN ranks in memory. It turned out that the restriction makes the results better: a vocabulary cutoff of $2^{15} = 32\ 768$ both on the source and the target size yields slightly better results (74.3%) than the more ambitious $2^{16} = 65536$ (73.9%). This is not the case for forward NN search, where accuracy increases with vocabulary limit (but remains far below that of reverse NN).

8.4.2 *Orthogonal restriction and other tricks*

Xing et al. (2015) note that the original linear translation method is theoretically inconsistent because it is based on three different similarity measures: word2vec pre-training itself uses the dot-product of unnormalized vectors, the translation is trained based on Euclidean distance, and neighbors are queried based on cosine similarity. They make the framework more coherent by length-normalizing the embeddings, and restricting W to preserve vector length. Mathematically this means that the matrix W is orthogonal, i.e. the mapping is a rotation.

Faruqui and Dyer (2014) achieve even better results by mapping the two embeddings to a lower-dimensional bilingual space with canonical correlation analysis. Artetxe, Labaka, and Agirre (2016) analyze ele-

6 We use lowercase r in the abbreviation, to avoid confusion with recurrent neural networks.

7 If more target words have the same forward rank, Dinu, Lazaridou, and Baroni (2015) make the decision based on cosine similarity. This tie breaking has not proven useful in our experiments.

ments of these two works both theoretically and empirically, and find a combination that improves upon dictionary generation and also preserves analogies (Mikolov, Yih, and Zweig 2013) like

$$\textbf{woman} + \textbf{king} - \textbf{man} \approx \textbf{queen}$$

among the mapped points Wx_i. They find that the orthogonality constraint is key to preserve performance in analogies, and it also improves bilingual performance. In their experiments, length normalization, when followed by centering the embeddings to $\mathbf{0}$ mean, obtains further improvements in bilingual performance without hurting monolingual performance.

8.5 EXPERIMENTS

8.5.1 *Data*

We trained `neela`, `AdaGram`[8], and `mutli` models on (the original and stemmed[9] forms of) two semi-gigaword (.7–.8 B words) Hungarian corpora, the Hungarian Webcorpus (Webkorpusz, Halácsy et al. (2004)) and the Hungarian National Corpus v2.0.3 (HNC, Oravecz, Váradi, and Sass (2014)). We used (non-triangulated) Wiktionary as our seed dictionary, extracted with `wikt2dict`[10] (Ács, Pajkossy, and Kornai 2013). We tried several English embeddings as target, including the 300 dimensional skip-gram with negative sampling model `GoogleNews` released with word2vec (Section 4.2.4, Mikolov, Chen, et al. (2013))[11], and those released with GloVe Section 4.2.6, Pennington, Socher, and Manning (2014))[12]. We report the best results, which were obtained with the release GloVe embeddings trained on 840 B words in 300 dimensions.

8.5.2 *Orthogonal constraint*

We implemented the orthogonal restriction by computing the singular value decomposition

$$U\Sigma V = S_t^\top T_t$$

where S_t and T_t are the matrices consisting of the embedding vectors of the training word pairs in the source and the target space, respectively, and taking

8 I would like to thank Sergey Bartunov for help with his tool.
9 Follow-up work reported in section 8.5.5 applied a third option in preprocessing.
10 https://github.com/juditacs/wikt2dict
11 https://code.google.com/archive/p/word2vec/
12 https://nlp.stanford.edu/projects/glove/

		8192				16384				32768			
		general linear		orthogonal		general linear		orthogonal		general linear		orthogonal	
		any	disamb	any	disamb	any	disamb	any	disamb	any	disamb	any	disamb
fwd	vanilla	28.7%	2.40%	32.1%	2.40%	36.2%	3.40%	42.0%	4.70%	36.7%	4.20%	44.5%	6.00%
	normalize	28.2%	2.20%	**33.7%**	3.40%	35.1%	2.80%	**44.4%**	5.80%	36.6%	3.80%	**48.2%**	6.00%
	+ center	26.6%	2.10%	32.8%	2.90%	32.9%	2.70%	42.0%	4.50%	34.6%	3.50%	43.9%	5.50%
rev	vanilla	**53.8%**	11.85%	51.7%	11.37%	**58.3%**	11.99%	56.6%	12.59%	**74.3%**	23.60%	73.6%	22.30%
	normalize	53.3%	11.61%	50.0%	10.90%	58.0%	12.35%	56.5%	12.59%	73.7%	24.20%	72.8%	22.10%
	+ center	51.7%	11.37%	53.3%	11.14%	57.1%	11.99%	57.7%	12.35%	69.7%	22.20%	73.5%	23.00%

Table 54: Precision@10 of forward and reverse NN translations with and without the orthogonality constraint and related techniques at vocabulary cutoffs 8192 to 32768. `any` and `disamb` are explained in section 8.5.3. The source was an `AdaGram` model in 800 dimensions, $\alpha = .1$, trained on Webkorpusz with the vocabulary cut off at 8192 sense vectors.

$$W = U\mathbf{1}V$$

where $\mathbf{1}$ is the rectangular identity matrix of appropriate shape. The orthogonal approximation was implemented following a code[13] by Gábor Borbély.

Table 54 shows the effect of these factors. (Recall the evaluation protocol from Section 8.4.) Precision in forward NN search follows a similar trend to that by Xing et al. (2015) and Artetxe, Labaka, and Agirre (2016): the best combination is an orthogonal mapping between length-normalized vectors; however, centering did not help in our experiments. Reverse NNs yield much better results than the simpler method, but none of the orthogonality-related techniques give further improvement here.

8.5.3 *Results*

We evaluate MSE models in two ways, referred to as `any` and `disamb`. The method `any` has been used for tuning the (meta)parameters of the source embedding and to choose the target: a traditional, single-sense translation was trained between the first sense vector of each word form and its translations. If the training word is ambiguous in the seed dictionary, all translations were included in the training data. Exploiting the multiple sense vectors, one word can have more than one translation. Attila Novák (personal communication) suggests that restricting the training to monosemous source words would provide a cleaner signal. However, with that method, many seed word pairs would be used for training, and evaluation would have to happen in the lower-frequency domain. During test, a source word was accepted if `any` of its sense vectors had at least one good translation among its k reverse nearest neighbors (rNN@k).

13 https://github.com/hlt-bme-hu/eval-embed

	dim	α/γ	p	m	any	disamb
HNC	800	.02		100	48.5%	7.6%
`neela` Wk	300	–	2	big	54.0%	12.4%
HNC stem	800	.05		big	55.1%	10.4%
HNC	160	.05	3	200	62.2%	15.0%
`mutli` Wk	300	.25		71	62.9%	17.4%
Webkorpusz	800	.05		100	65.9%	17.4%
HNC	600	.05	5	100	68.6%	16.6%
HNC	600	.1	3	50	69.1%	18.8%
Webkorpusz	800	.1		100	73.9%	23.9%

Table 55: Our measures, `any` and `disamb`, for different MSEs. The source embedding was trained with `AdaGram`, except for when indicated otherwise (`neela`, `mutli`). The meta-parameters are *dim*ension, the resolution parameter (α in `AdaGram` and γ in `mutli`), the maximum number of *p*rototypes (sense vectors), and the vocabulary cutoff (*m*in-freq, the two models with *big* have practically no cut-off).

In `disamb`, we used the same translation matrix as in `any`, and inspected the translations of the different sense vectors to see whether the vectors really model different senses rather than synonyms. The lowest requirement for the non-synonymy of sense vectors s_1, s_2 is that the sets of corresponding good rNN@k translations are different. The ratio of words satisfying this requirement among all words with more than one sense vector is shown as `disamb` in table 55.

The values in Table 55 are low. This can in part be due to that the `neela` and the `mutli` models were trained with lower dimension than the best-performing model. This also means that results here are not comparable among these different architectures. Follow-up experiments (conducted after the paper review) are reported in section 8.5.5.

Table 56 shows the successfully disambiguated words sorted by the cosine similarity s of good rNN@1 translations of different sense vectors. This human evaluation was done by a senior computational semanticist. (We found that most of the few cases when there are more than two sense vectors with a good rNN@1 translation are due to the fact that the seed dictionary contains some non-basic translation, e.g. *kapcsolat* 'relationship, conjunction' has 'affair' among its seed translations. In these cases, we chose two sense vectors arbitrarily. When there are sense vectors with more than two rNN@k hits, the choice of the corresponding target words is also arbitrary.) Relying on s is similar to the monolingual setting of clustering the sense vectors for each word, but here we restrict our analysis to sense vectors that prove to be sensible in linear translation.

	s			covg
E	-0.04849	függő	addict, aerial	0.4
S	0.01821	alkotó	constituent, creator	0.5
S	0.05096	előzetes	preliminary, trailer	1.0
S	0.0974	kapcsolat	affair, conjunction, linkage	0.33
I	0.1361	kocsi	coach, carriage	1.0
S	0.136	futó	runner, bishop	1.0
S	0.1518	keresés	quest, scan	0.67
S	0.1574	látvány	outlook, scenery, prospect	0.6
S	0.1626	fogad	bet, greet	1.0
S	0.1873	induló	march, candidate	1.0
I	0.187	nemes	noble, peer	0.67
E	0.1934	eltérés	variance, departure	0.4
E	0.1943	alkalmazás	employ, adaptation	0.33
S	0.2016	szünet	interval, cease, recess	0.43
E	0.2032	kezdeményezés	initiation, initiative	1.0
S	0.2052	zavar	disturbance, annoy, disturb, turmoil	0.57
S	0.2054	megelőző	preceding, preventive	0.29
IE	0.2169	csomó	knot^{I}, lump^{I}, mat^{E}	1.0
E[14]	0.21	remény	outlook, promise, expectancy	0.6
S	0.2206	bemutató	exhibition, presenter	0.67
E	0.2208	egyeztetés	reconciliation, correlation	0.5
S	0.237	előadó	auditorium, lecturer	0.67
E	0.2447	nyilatkozat	profession, declaration	0.4
I	0.2494	gazda	farmer, boss	0.67
I	0.2506	kapu	gate, portal	1.0
I	0.2515	előbbi	anterior, preceding	0.67
I	0.2558	kötelezettség	engagement, obligation	0.67
E	0.265	hangulat	morale, humour	0.5
E	0.2733	követ	succeed, haunt	0.67
SE	0.276	minta	norm^{S}, formula^{E}, specimen^{S}	0.75
S	0.2807	sorozat	suite, serial, succession	1.0
S	0.2935	durva	coarse, gross	0.18
I	0.3038	köt	bind, tie	0.67
E	0.3045	egyezmény	treaty, protocol	0.67
I	0.3097	megkülönböztetés	discrimination, differentiation	0.5
I	0.309	ered	stem, originate	0.5
I	0.319	hirdet	advertise, proclaim	1.0
E	0.3212	tartós	substantial, durable	1.0
I	0.3218	ajánlattevő	bidder, supplier, contractor	0.6
I	0.3299	aláírás	signing, signature	0.67
I	0.333	bír	bear, possess	1.0
I	0.3432	áldozat	sacrifice, victim, casualty	1.0
IE	0.3486	kerület	ward^{I}, borough^{I}, perimeter^{E}	0.3
I	0.3486	utas	fare, passenger	1.0
I	0.3564	szigorú	stern, strict	0.5
I	0.3589	bűnös	sinful, guilty	0.5
I	0.3708	rendes	orderly, ordinary	0.5
I	0.3824	eladó	salesman, vendor	0.5
I	0.3861	enyhe	tender, mild, slight	0.6
I	0.3897	maradék	residue, remainder	0.33
I	0.3986	darab	chunk, fragment	0.4
E	0.4012	hiány	poverty, shortage	0.5
I	0.4093	kutatás	exploration, quest	0.5
⋮				
I	0.4138	tanítás	tuition, lesson	0.67
I	0.4196	őszinte	frank, sincere	0.67
I	0.4229	környék	neighborhood, surroundings, vicinity	0.38
I	0.4446	ítélet	judgement, sentence	0.67
I	0.4501	gyerek	childish, kid	0.67
I	0.4521	csatorna	ditch, sewer	0.4
I	0.4547	felügyelet	surveillance, inspection, supervision	0.43
E	0.4551	ritka	rare, odd	0.5
S	0.4563	szerető	fond, lover, affectionate, mistress	0.67
I	0.4608	szeretet	affection, liking	0.67
I	0.4723	vizsgálat	inquiry, examination	0.67
I	0.4853	tömeg	mob, crowd	0.5
I	0.4903	puszta	pure, plain	0.22
I	0.4904	srác	kid, lad	1.0
I	0.4911	büntetés	penalty, sentence	0.29
I	0.4971	képviselő	delegate, representative	0.67
I	0.4975	határ	boundary, border	0.67
I	0.5001	drága	precious, dear, expensive	1.0
S	0.5093	uralkodó	prince, ruler, sovereign	0.5
I	0.5097	válás	separation, divorce	0.67
I	0.5103	ügyvéd	lawyer, advocate	0.67
I	0.5167	előnyös	advantageous, profitable, favourable	1.0
I	0.5169	merev	rigid, strict	1.0
I	0.5204	nyíltan	openly, outright	1.0
I	0.5217	noha	notwithstanding, albeit	1.0
I	0.5311	hulladék	litter, garbage, rubbish	0.43
I	0.5311	szemét	litter, garbage, rubbish	0.43
I	0.5612	kielégítő	satisfying, satisfactory	1.0
E	0.5617	vicc	joke, humour	1.0
I	0.5737	szállító	supplier, vendor	1.0
I	0.5747	óvoda	nursery, daycare, kindergarten	1.0
I	0.5754	hétköznapi	mundane, everyday, ordinary	0.75
I	0.5797	anya	mum, mummy	1.0
I	0.5824	szomszédos	neighbouring, neighbour	0.4
E	0.5931	szabadság	liberty, independence	1.0
I	0.6086	lelkész	pastor, priest	0.4
I	0.6304	fogalom	notion, conception	1.0
I	0.6474	fizetés	salary, wage	0.67
I	0.6551	táj	landscape, scenery	1.0
I	0.6583	okos	clever, smart	0.67
I	0.6707	autópálya	highway, motorway	0.5
I	0.6722	tilos	prohibited, forbidden	1.0
I	0.6811	bevezető	introduction, introductory	1.0
I	0.7025	szövetség	coalition, alliance, union	0.75
I	0.7065	fáradt	exhausted, tired, weary	1.0
I	0.7066	kiállítás	exhibit, exhibition	0.67
I	0.7135	hirdetés	advert, advertisement	1.0
I	0.7147	ésszerű	rational, logical	1.0
I	0.7664	logikai	logic, logical	1.0
I	0.7757	szervez	organise, organize, arrange	1.0
I	0.8122	furcsa	strange, odd	0.4
I	0.8277	azután	afterwards, afterward	0.67
I	0.8689	megbízható	dependable, reliable	0.67

Table 56: Hungarian words with the rNN@1 translations of their sense vectors. The first column is a post-hoc annotation by András Kornai (*E* error in translation, *I* identical, *S* separate meanings), *s* is the cosine similarity of the translations, and *covg* denotes the coverage of the @1 translations over all gold (good) translations.

[14] The basic translation *hope* is missing

	any	disamb
`AdaGram`	73.3%	18.53%
`mutli` sense vectors	71.0%	19.46%
`mutli` context vectors	69.9%	20.76%

Table 57: The resolution trade-off between translation precision and sense distinctiveness. The source models are 600-dimensional Hungarian models trained on the de-glutinized version of the Hungarian National Corpus. Other meta-parameters have been set to default.

We see that most words with $s < .25$ are really ambiguous from a standard lexicographic point of view, but the translations with $s > .35$ tend to be synonyms instead.

8.5.4 *Part of speech*

The clearest case of homonymy is when unrelated senses belong to different parts-of-speech (POSs), and the translations reflect these POSs, e.g. *nő* 'woman; increase' or *vár* 'wait; castle'.[15] In purely semantic approaches, like `4lang` (see Section 3.1.2), POS-difference alone is not enough for analyzing a word as ambiguous, e.g. we see the only difference between the noun and participle senses of *alkalmazott*, 'employee; applied' as *employ*ment being the *appli*cation of people for work. Similarly, in the case of *belső* 'internal; interior', the noun refers to the part of a building described by the adjective.

More interesting are word forms with related senses in the same POS, e.g. *cikk*, 'item; article' (an article is an item in a newspaper); *eredmény*, 'score; result' (a score is a result measured by a number); *magas*, 'tall; high' (tall is used for people rather than high); or *idegen*, 'stranger, alien; foreign', where the English translations are special cases of 'unfamiliar' (person or language).

Finally we mention two cases where the relation between the two senses is more idiosyncratic, but in a monosemic approach, they will have a single representation: *beteg* means 'ill, sick; patient'. Though *ill* is a health state and *patient* is a situational role, patients of doctors are usually ill. A monosemic system is designed to give account of metaphorical relations like the one between the meanings of *világos*, 'bright; clear' as well.

8.5.5 *Comparison of* `AdaGram` *and* `mutli`

After the compilation of the 2017 edition of the Festschrift where Makrai and Lipp (2018) appeared, we trained models that enable a more fair comparison of `AdaGram` and `mutli` in terms of semantic resolution: we trained 600-dimensional models for Hungarian to have the 2:1 ratio between the source and the target dimension that has been reported to be optimal for this task (Mikolov, Le, and Sutskever 2013; Makrai 2016). This time we used the deglutinized version (Section 4.2.11.1) of the Hungarian National corpus for better morphological generalization. Recall from Section 7.4.6, that was not necessarily a good choice. The word embeddings are available online[16].

We can see in table 57[17] that there is a trade-off between the two measures, which may be interpreted to indicate that the more specific a vector is, the easier it is to translate, but if the vectors are too specific, then the translations may coincide.

While contextualized word representations of deep language models offer themselves as a starting point for the computational analysis of word ambiguity, the researcher interested specifically in the Dirichlet Process modelnig of word ambiguity may analyse the observed and inferred number of word senses as a function of word frequency in these models.

8.6 CONCLUSION

We proposed a method for measuring the precision of multi-sense word embeddings as detectors of word ambiguity. The method is based on linear translation. Investigating the effect of a couple of standard tricks of linear translation, it turns out that inverse neighbors are important, while orthogonal restriction and related techniques are not, even slightly harmful.

By comparing the two main MSE models, `AdaGram` and `mutli`, we found that taking former as the source space, we get good translations more often, however over-disambiguiation also happens more frequently than with the latter. This is in line with the intuition that the more subtle the meaning inventory, even to the extent of over-disambiguiation, the easier it is to translate.

15 We note that some POSs in Hungarian have blurred borders, e.g. it is debatable whether the nominal *önkéntes* 'voluntary; volunteer' is ambiguous for its POS.

16 https://hlt.bme.hu/en/publ/makrai17

17 There are two `mutli` models because Skip-gram and the related MSE models represent each word with two vectors, u and v in the formula $p(w_i \mid w_j) \propto \exp(u_i^\top v_j)$, that `mutli` calls *sense* versus *context* vectors, respectively.

These results display two properties, one of them remarkable.

— Levelt, Roelofs, and Meyer 1999

9

SUMMARY

Learning, and then understanding what we have learned, what more we could learn. The dual role of distributional and symbolic meaning representations can be summarized this way. The past decade and even more the past few years made it possible to train neural language models. Shallow models show already remarkable properties, and the process towards human level linguistic understanding has continued with deep language models. However, in order to understand what the model has really learned, the researcher has to investigate the representation by testing hypotheses stated in discrete terms. While the focus of present natural language processing is deep modeling, investigating the linguistic content of the model is not trivial even with shallow, static word embeddings. This thesis is a collection of works within this undertaking. Here, we summarize the most important result of the previous chapters.

9.1 PAGERANK FOR THE IMPORTANCE OF CONCEPTS

One of our contributions to this connection is related to the definition graph (Section 3.3.2), which is computed from word definitions, and can be transformed to word embeddings (see Sections 9.4.2 and 9.4.3). As a kind of feedback, the definition graph can be used to measure the importance of each defining symbol (in the case of `4lang`: concepts, binary relations, deep cases, and encyclopedic references). Indeed, in Section 3.3, **we quantified the importance each node of the semantic network plays in the recursive process of defining words by**

each other. It turned out that the greatest burden is worn by special elements in the formalism, especially deep cases (i.e. the place-holders of the representation of an argument within the representation of a function), nodes corresponding to lexical relations (e.g. the comparative *-er*), more or less contentful unary or binary predicates (e.g. *exist, want*), and special nodes in the formalism, e.g. `other`, which blocks the unification of two nodes in a definition with the same label.

9.2 THEMATIC PLACEHOLDERS OF ARGUMENTS IN 4LANG

It is not surprising that for the representation of a phrase headed by a predicate, the representation of the arguments is very important. Argument labels are needed to indicate where the representation of each argument has to be inserted. Accordingly, in Chapter 5 **we proposed a set of deep cases along with the hand-written formulaic definitions of the core vocabulary of `4lang`** (Section 3.2). Deep cases denote the nodes in the graph representing the meaning of a predicate where the representation of the argument (single word, entity or phrase) has to be inserted. In theory, the interpretation of a verb along with an argument (i.e. the satisfaction of selectional preferences) is lead by spreading activation (Section 2.2.2) in the definition graph, which is the second role of this graph in the thesis.

Our theoretical principle has been to capture syntactic-semantic regularities that appear in many languages. In the radically monosemic approach of `4lang`, the transitive and the intransitive use of the same verb is represented with the same formula, which contains both the agent (`AGT`) and the patient (`PAT`). Unaccusative verbs were attributed a deep patient. The recipient of both physical and mental transfer verbs have been represented as a deep dative. While the inventory of deep cases consists of just eight members, three of these are locative (`TO`, `FROM`, and the static `AT`). While most relational nouns are linked to the related entity with a possessive (`POSS`, e.g. *the absence of war*), our radical monosemic approach also implies that relational nouns whose arguments have similar linguistic markers as goals of verbs are attributed a deep goal (`TO`, e.g. *need for peace*). The roots of affixes, the objects of adpositions, the referents of adjectives (e.g. ***money** is necessary for war*), and some exceptional relations are marked as semantically neural relations (`REL`).

9.3 SUBJECT-VERB-OBJECT ASSOCIATION MODELING

The verb is the pivot of a sentence, and we were interested in what a purely distributional approach tells us about argument structure, especially about the subject and the object. In Chapter 6, we investigated the following questions:

- Which *association measures* are the best to characterize the co-occurrences of English verbs with their subjects and objects? The simplest testbed for distributional methods is how they predict similarity, so we chose to experiment with the comparison of English subject-verb-object triples. **We implemented** several measures, including **multiple novel generalizations of weighted positive pointwise mutual information (PPMI) to the higher-order (>2) case.** Pointwise mutual information (PMI) has two higher-order generalizations, the more popular one which is still called PMI in the literature, and interaction information. We combined both generalizations with salience (Kilgarriff et al. 2004) and normalization (Bouma 2009). The former is motivated by lexicographic practice and the latter by making the function bounded.

  Modeling the three-way interactions of English subject-verb-object triples with tensor decomposition, these weighted higher-order PPMI variants have proven better than the baselines (log frequency, vanilla PPMI, and log Dice). Specifically the best result was obtained by the non-negative Canonical Polyadic Decomposition (CPD) of a salience-weighted PMI tensor, followed by the general Tucker decomposition of a normalized PMI tensor.

- We also asked whether *empty argument fillers* (subjects or, more importantly, objects, e.g. *John drinks*) should be included in our co-occurrence statistics for better generalization over the transitive and the intransitive uses of the same verb, or they just introduce noise. Our two best results (non-negative CPD and general Tucker) suggest that **the inclusion of empty objects does benefit word representation.** This is also in line with our monosemic approach discussed in the previous chapter, i.e. that the `4lang` formulas of verbs with optional objects represent the two uses with the same formula.

- Our two tensor decomposition algorithms, CPD and Tucker, have very different time-complexity: Tucker is much faster. On the other hand, tensor decomposition has hyper-parameters like the decomposition rank and the frequency cutoff. Both are related to memory limitation, especially the latter. Nevertheless, while the cutoff is only to ensure that the decomposition fits within the memory limits, the rank is an essential parameter. Already Landauer and Dumais (1997) argued that the 300 dimensions of LSA (Section 4.1.3) are psychologically real. We found that *the two algorithms reach the best results with similar rank,* what is beneficial, because a fast parameter tuning with Tucker provides the hyper-parameter value for CPD. Specifically, we found that,

the best results are obtained with a rank of 64, either with non-negative CPD or with general Tucker.

- How does the trade-off between the three hyper-parameters related to the *size of the decomposition* (i.e. the decomposition rank, the inclusion of empty fillers, and the frequency cutoff) look like? If we exclude empty fillers, a more generous frequency cutoff may theoretically lead to better results than if we change only one of these two parameters, i.e. a larger training sample can compensate for the loss of intransitive information. It turns out, that we can indeed get relatively good Spearman correlation this way (0.6942 with a cutoff of 1 million, instead of our overall best 0.7359)[1], but with general Tucker decomposition (instead of non-negative CPD) and log-Dice (instead of salience-weighted PMI).

- Do latent dimension of our word embeddings reflect lexical knowledge? **Dimensions obtained with the two *non-negative algorithms* are** indeed **semantically interpretable** (e.g. *story catch(es) attention*), while those from general decomposition are less convincing. This is in parallel with the general motivation of non-negative representations by interpretability, what we also saw in our hypernym extraction experiments (Section 9.4.1).

- Can **the difference between each noun as a subject versus as an object correspond to** some intuitive difference between subjecthood and objecthood? Indeed, the greatest difference between the two roles is found with personal pronouns (or the missing filler, what is not surprising: an empty subject is not similar to an empty object), while the smallest is with abstract nouns like *doubt*. A possible explanation is that the former are much more frequent in **agentive roles** than other nouns, while they are infrequent in patient roles. Words in the second group can be framed in language both as if they were animate and as inanimate. *Future* or *hope* are not alive in the biological sense, but they are often attributed agentive roles. This again supports the monosemic/force-dynamic (Section 2.3.4) approach to the treatment of metaphorical word usage that we follow in `4lang`: when future shows something is essentially the same as when a teacher shows something on the blackboard.

9.4 LEXICAL RELATIONS, ANALOGY, AND TRANSLATION

The last two chapters returned to our main question, the relation between symbolic and neural representations. How are lexical relations like hypernymy, antonymy, and causality represented? Is the similarity of relations captured by the vector offset method (Section 4.2.4)

1 A correlation of 0 means independence and 1 is the theoretical maximum.

in morphologically rich languages like they are in English? Is a linear mapping of GloVe vectors suitable for word translation between medium resourced languages as word2vec vectors are between English and Spanish?

Before applying analogical tests, which benchmark the flexible processing of any kind of relation, we paid special attention to three individual relations: hypernym, antonymy, and causality.

9.4.1 *Hypernym extraction with sparse word representations*

Section 7.1 investigated hypernym with the tools of sparse coding. A variant of the distributional hypothesis, the distributional *inclusion* hypothesis (Weeds and Weir 2003; Chang et al. 2018) says that *hypernymy* can be modeled based on that if *animal* is a hypernym of *dog*, *animal* will be grammatical in every context where *dog* is. It is less clear whether *animal* will appear in every context *at least as frequently* as *dog* does.

Sparse vectors are embedding vectors most of whose coordinates are zero, and non-zero coordinates ideally correspond to interpretable properties. It varies with models whether interpretability follows from the construction of the vectors, or the interpretation needs to be inferred from some latent structure. Even in the latter case, sparse representations tend to be more interpretable than less restricted ones. As far as sparse attributes (i.e. non-zero coordinates in *sparse word representations*) correspond to contexts, it follows from the distributional inclusion hypothesis discussed above that hypernymy should boil down to pointwise comparison: ideally, *animal* is a hypernym of *dog* if and only if *dog* has all the properties *animal* has, which in turn is equivalent that all the coordinates which are non-zero in *animal* are non-zero in *dog*. Formal concept analysis (FCA) captures this same idea in strict discrete terms. **Our experiments[2] showed that while the discrete FCA** approach is brittle, and the resulting **features are counterproductive in the task, the probabilistic distributional relaxation implemented with sparse coding** gave much better results, and **enabled us to win three subtasks of SemEval-2018 Task 9.**

9.4.2 *Antonymy in classical and definition-based word embeddings*

In Section 7.2, **we took a word embedding computed from the definition graph** with spectral clustering, **and tested it in the task of representing sub-types of antonymy.** This was the third role that this graph played in this thesis. (The first one was Section 9.1, and the second one was mentioned in Section 9.2.) In this third case, we used it to compute a word embedding, which we **compared to some**

2 The ratio of contributions is Berend: Makrai = 2:1

other embeddings which were famous at the time. Specifically, **we tested[3] which subtype of antonymy is represented in each word embedding. The embedding obtained from the definition graph turned out to be similar in this respect to variants of HLBL (Section 4.2.2)** which used to be a famous word embedding set, what means that our embedding passed the sanity check. This alternative way of embedding creation enables full interpretability and control over the content of the word representations.

9.4.3 *The geometry of causal word pairs*

Our investigation of causality in Section 7.3 exemplifies exploratory computational linguistics: we started with a visual inspection of **cause-effect pairs in the word embeddings space**. The 2D plots suggested a very interesting property: **the lines connecting each cause with its effect run close to a common "center of causality". We used statistical tests to see whether the property holds** in the original space. **Senna**, a classical early word embedding for English (Section 4.2.3) **showed the property**, while many other early word embeddings and some of ours computed from the definition graph did not. (This was the third role that this graph had in the thesis.) Nevertheless, another linear algebraic formulation makes this finding cognitive linguistically appealing: we can say, at least in SENNA, that **the meaning of an effect is a combination of the meaning of the corresponding cause and a uniform causal element.**

9.4.4 *Analogy in word embeddings for a morphologically rich language*

After investigating the classical lexical relaxations of hypernym, antonymy and causality, **we created a benchmark of analogical relations for the morphologically rich Hungarian, and applied them to test static word embeddings (Section 7.4).** To our best knowliedge, Hungarian was the third language to be tested for analogy (after English and Turkish). **While for the semantic relations, the results were poor, even if we change to gluten-free embeddings, we found that morphological relations are reflected in a similarly clear fashion as in the better-resourced languages.**

3 See the specific contributions of Makrai and Nemeskey in Footnote 7 in Section 7.2

9.4.5 *Linear translation*

9.4.5.1 *Mid-resourced European languages*

In **translation from mid-resourced European languages** (also in Section 7.4), we obtained similar scores as the seminal papers "out of the box" even **with GloVe** instead of word2vec.

9.4.5.2 *Smooth filtering of triangulated translations*

Besides, in Section 7.5, **we combined linear word translation with the old method of triangulation, a.k.a. pivot-based word translation. We filtered a list of triangulated translational word pairs based on the distance** $\cos(t, Ms)$ between the embedding vector t of the word in the target language, and the vector obtained by mapping the source vector s, where M is the translation mapping. Accuracy decreases both with the number of the pivots and with the **cosine**, but we found that the latter score **is smoother than the more traditional measure, which suggests that some factor of noise is eliminated,** and good translations can be selected based on the cosine score more reliably. We demonstrated the method by **publishing the largest German-Hungarian dictionary (word pair list) of the time.**

9.5 LINEAR TRANSLATION FOR WORD SENSE INDUCTION

One of the motivations for the tensor decomposition approach to word modeling (see Section 9.3) is that three-order co-occurrences like subject-verb-object triples cannot be reduced to two-order ones (subject-verb pairs and verb-object pairs). It goes without saying that this higher-order behavior is related to that verbs are polysemous, they are used in very flexible ways, a phenomenon Pustejovsky (Section 2.3.7) calls co-composition. More generally, other major parts of speech are also often ambiguous, which is one of the greatest problems in lexical semantics.

Distributional models including static word embeddings basically represent each word form with a single embedding vector. Nowadays a discrete representation with more vectors for ambiguous words could be obtained from contextualized word representations. Nevertheless the problem could be targeted before the advent of deep language models as well, by learning discrete representations direly. These are static multi-sense word embeddings (MSE).

Word ambiguity can be divided to homonymy and polysemy. In Chapter 8, **we[4] proposed an evaluation method for MSEs as detec-**

4 The idea is joint work with Borbély and Kornai, the elaboration is individual contribution, see Footnote 1 in Chapter 8.

tors of homonymy. The method fits **in the context of embedding-based dictionary induction**, and we also analyzed the interaction between some techniques of this paradigm, especially reverse nearest neighbor search and the orthogonality constraint. We found that **reverse nearest neighbors yield much better results than the forward method** (e.g. 74.3% instead of 36.6% with the greatest vocabulary cutoff). While the orthogonality constraint helps in the forward case (48.2%), **none of the orthogonality-related techniques give further improvement** in the reverse case.

Using our two measures, `any`, which quantifies translation quality in general, and `disamb`, which measures the precision of ambiguous word detection, we compared the two SOTA MSE models, `AdaGram` and `mutli`. We found a trade-off between the two measures: the more specific a vector is, the easier it is to translate, but if the vectors are too specific, then the translations may coincide.

9.6 FINAL REMARKS

The direct message of this thesis is that the conceptual structure of word meaning can be read out from distributional models. Hypernymy is a relatively straightforward testbed, where we showed the remarkable role of sparse attribute pairs (Section 9.4.1). Causality, which would appear to lie out of the range of word-level computational methods, also shows some simple geometric structure (Section 9.4.3).

Taking a step further, this thesis proposes two very general principles for data-driven computational linguistics: First, we found that it is beneficial to treat similar phenomena with a unified mechanism, what is strongly related to `4lang`'s monosemic principle. One example is the transitive and the intransitive use of the same verb, the other one is concrete and metaphoric uses of the same word.

Our second lesson is that it is often beneficial to apply methods from some branch of computational lexicography to another one. Firstly, the definition graph is an interesting intermediary between the conceptual and the linear algebraic domain, as this kind of representation is computed from word definitions, but lends itself for distributional methods. Second, the difference between polysemy and homonymy is another field where this thesis directly translated a principle of lexicology to distributional terms. Intuitively it is clear that the distinction is related to translation: While it is possible that different uses of a polysemous word can be translated with the same word, this can happen to homonyms only by coincidence. In this thesis, we formulated the conceptual relation between translation and the types of ambiguity by applying the linear word translation to measuring the recall of multi-sense word embeddings as the detectors of word ambiguity.

The confluence of methods is evident in data science (e.g. computer vision models are applied to sound spectrograms). The preceding chapters suggest that lexical/cognitive linguistics fits in this picture of the future nicely.

Abend, Omri, and Ari Rappoport. 2013. "UCCA: A semantics-based grammatical annotation scheme." In *IWCS'13,* 1–12. (Cited on page 65).

———. 2017. "The state of the art in semantic representation." In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 77–89. https://doi.org/10.18653/v1/P17-1008. (Cited on pages 15, 62–65).

Ács, Judit, Dávid Márk Nemeskey, and Gábor Recski. 2017. "Building word embeddings from dictionary definitions." In *K + K = 120: Papers dedicated to László Kálmán and András Kornai on the occasion of their 60th birthdays,* edited by Katalin Mády Beáta Gyuris and Gábor Recski. Research Institute for Linguistics, Hungarian Academy of Sciences (RIL HAS). (Cited on pages 11, 75, 156).

Ács, Judit, Katalin Pajkossy, and András Kornai. 2013. "Building basic vocabulary across 40 languages." In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora,* 52–58. Sofia, Bulgaria: Association for Computational Linguistics. (Cited on pages 71, 76, 77, 205, 214, 216, 229).

Afify, Mohamed, Ruhi Sarikaya, Hong-Kwang Jeff Kuo, Laurent Besacier, and Yuqing Gao. 2006. "On the use of morphological analysis for dialectal Arabic speech recognition." In *INTERSPEECH,* 277–280. https://doi.org/10.21437/Interspeech.2006-87. (Cited on page 118).

Akbik, Alan, Duncan Blythe, and Roland Vollgraf. 2018. "Contextual String Embeddings for Sequence Labeling." In *Proceedings of the 27th International Conference on Computational Linguistics,* 1638–1649. Santa Fe, New Mexico, USA: Association for Computational Linguistics, August. https://www.aclweb.org/anthology/C18-1139. (Cited on page 125).

Allen, James, and Choh Man Teng. 2018. "Putting semantics into semantic roles." In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics.* https://doi.org/10.18653/v1/S18-2028. (Cited on page 150).

Andrews, Avery. 2015. "Reconciling NSM and formal semantics." *ms,* v2, jan 2015. https://doi.org/10.1080/07268602.2016.1109431. (Cited on page 76).

Antoniak, Maria, and David Mimno. 2018. "Evaluating the stability of embedding-based word similarities." *Transactions of the Association for Computational Linguistics* 6:107–119. https://doi.org/10.1162/tacl_a_00008. (Cited on page 95).

Arora, Sanjeev, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2015. "Random Walks on Context Spaces: Towards an Explanation of the Mysteries of Semantic Word Embeddings." *arXiv:1502.03520v1* 4:385–399. https://doi.org/10.1162/tacl_a_00106. (Cited on page 111).

———. 2016. "Linear Algebraic Structure of Word Senses, with Applications to Polysemy." *arXiv:1601.03764v1,* https://doi.org/10.48550/arXiv.1601.03764. (Cited on pages 114, 181).

Arora, Sanjeev, Yingyu Liang, and Tengyu Ma. 2017. "A simple but tough-to-beat baseline for sentence embeddings." In *International Conference on Learning Representations.* https://openreview.net/pdf?id=SyK00v5xx. (Cited on page 143).

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre. 2016. "Learning principled bilingual mappings of word embeddings while preserving monolingual invariance." In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.* https://doi.org/10.18653/v1/D16-1250. (Cited on pages 228, 230).

Avraham, Oded, and Yoav Goldberg. 2016. "Improving reliability of word similarity evaluation by redesigning annotation task and performance measure." *arXiv preprint arXiv:1611.03641,* https://doi.org/10.18653/v1/W16-2519. (Cited on page 114).

Bailey, Eric, Charles Meyer, and Shuchin Aeron. 2018. "Learning semantic word respresentations via tensor factorization." ArXiv:1704.02686. https://openreview.net/forum?id=B1kIr-WRb. (Cited on pages 160, 162–164, 172).

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. "The Berkeley FrameNet Project." In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1,* 86–90. ACL '98. Montreal, Quebec, Canada: Association for Computational Linguistics. https://doi.org/10.3115/980845.980860. (Cited on page 52).

Balogh, Vanda, Gábor Berend, Diochnos Dimitrios I., and György Turán. 2020. "Understanding the Semantic Content of Sparse Word Embeddings Using a Commonsense Knowledge Base." In *Proceedings of the AAAI Conference on Artificial Intelligence,* 34:7399–7406. 05. https://doi.org/https://doi.org/10.1609/aaai.v34i05.6235. (Cited on page 164).

Banarescu, Laura, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. "Abstract Meaning Representation for Sembanking." In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse,* 178–186. Sofia, Bulgaria: Association for Computational Linguistics. https://www.aclweb.org/anthology/W13-2322. (Cited on pages 57, 70).

Baroni, M., S. Bernardini, A. Ferraresi, and E. Zanchetta. 2009. "The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora." In *LREC 2009,* 3:209–226. https://doi.org/10.1007/s10579-009-9081-4. (Cited on page 216).

Baroni, M., G. Dinu, and G. Kruszewski. 2014. "Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors." In *Proceedings of ACL 2014,* 237–247. https://doi.org/10.3115/v1/P14-1023. (Cited on page 112).

Baroni, Marco, and Alessandro Lenci. 2010. "Distributional memory: A general framework for corpus-based semantics." *Computational Linguistics* 36 (4): 673–721. https://doi.org/10.1162/coli_a_00016. (Cited on page 12).

Bartunov, Sergey, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov. 2016. "Breaking Sticks and Ambiguities with Adaptive Skip-gram." *Proceedings of Machine Learning Research* 51: Artificial Intelligence and Statistics (May): 130–138. (Cited on page 227).

Basile, Valerio, Johan Bos, Kilian Evang, and Noortje Venhuizen. 2012. "Developing a large semantically annotated corpus." In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012),* 3196–3200. (Cited on page 64).

Belinkov, Yonatan, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. "What do Neural Machine Translation Models Learn about Morphology?" In *Proc. of ACL.* https://doi.org/10.18653/v1/P17-1080. (Cited on page 137).

Bengio, Yoshua, Aaron Courville, and Pascal Vincent. 2013. "Representation Learning: A Review and New Perspectives." *IEEE Trans. PAMI* 35 (8): 1798–1828. https://doi.org/10.1109/TPAMI.2013.50. (Cited on page 107).

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. "A Neural Probabilistic Language Model." *Journal of Machine Learning Research* 3:1137–1155. https://doi.org/10.1162/tacl_a_00059. (Cited on pages 106–108).

Berend, Gábor. 2017. "Sparse Coding of Neural Word Embeddings for Multilingual Sequence Labeling." *Transactions of the Association for Computational Linguistics* 5:247–261. ISSN: 2307-387X. https://doi.org/10.1162/tacl_a_00059. (Cited on pages 181, 183).

———. 2018. "Towards cross-lingual utilization of sparse word representations." In *XIV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2018),* edited by Veronika Vincze, 272–280. Szegedi Tudományegyetem Informatikai Tanszékcsoport. (Cited on page 181).

Berend, Gábor, Márton Makrai, and Péter Földiák. 2018. "300-sparsans at SemEval-2018 Task 9: Hypernymy as interaction of sparse attributes." In *Proceedings of The 12th International Workshop on Semantic Evaluation,* 928–934. New Orleans, Louisiana: Association for Computational Linguistics, June. https://doi.org/10.18653/v1/S18-1152. (Cited on page 181).

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent dirichlet allocation." *Journal of machine Learning research* 3 (Jan): 993–1022. (Cited on page 96).

Boguraev, Branimir K., and Edward J. Briscoe. 1989. *Computational Lexicography for Natural Language Processing.* Longman. (Cited on pages 14, 49, 50, 226).

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. "Enriching Word Vectors with Subword Information." *Transactions of the Association for Computational Linguistics* 5:135–146. ISSN: 2307-387X. https://doi.org/10.1162/tacl_a_00051. (Cited on page 107).

Borbély, Gábor, András Kornai, Dávid Nemeskey, and Marcus Kracht. 2016. "Denoising composition in distributional semantics." In *DSALT: Distributional Semantics and Linguistic Theory.* Poster. (Cited on page 118).

Borbély, Gábor, Márton Makrai, Dávid Márk Nemeskey, and András Kornai. 2016. "Evaluating multi-sense embeddings for semantic resolution monolingually and in word translation." In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP,* 83–89. Berlin, Germany: Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-2515. (Cited on pages 224, 225, 227).

Botha, Jan A, and Phil Blunsom. 2014. "Compositional Morphology for Word Representations and Language Modelling." In *ICML,* 1899–1907. (Cited on page 118).

Bouma, Gerlof. 2009. "Normalized (pointwise) mutual information in collocation extraction." In *GSCL 2009: International Conference of the German Society for Computational Linguistics and Language Technology.* (Cited on pages 159, 164, 237).

Brachman, R.J., and H. Levesque. 1985. *Readings in knowledge representation.* Morgan Kaufmann Publishers Inc., Los Altos, CA. (Cited on pages 11, 23, 30, 32).

Brants, Thorsten, and Alex Franz. 2006. *Web 1T 5-gram Version 1.* Philadelphia: Linguistic Data Consortium. (Cited on page 77).

Bresnan, Joan. 1978. "A realistic transformational grammar." In *Linguistic theory and psychological reality,* edited by M. Halle, J. Bresnan, and G.A. Miller. MIT Press. (Cited on page 47).

———. 2001. *Lexical-Functional Syntax.* Oxford, UK: Blackwell. (Cited on page 47).

Broder, Andrei Z. 1997. "On the resemblance and containment of documents." In *Compression and Complexity of Sequences 1997. Proceedings,* 21–29. IEEE. (Cited on page 103).

Brown, P.F., V.J. Della Pietra, P.V. de Souza, J.C. Lai, and R.L. Mercer. 1992. "Class–based n–gram models of natural language." *Computational Linguistics* 18 (4): 467–480. (Cited on page 106).

Bullon, Stephen. 2003. *Longman Dictionary of Contemporary English.* 4th ed. Longman. (Cited on page 70).

Butt, Miriam. 2006. *Theories of Case.* Cambridge University Press. https://doi.org/10.1017/CBO9781139164696. (Cited on pages 70, 154).

Buys, Jan, and Phil Blunsom. 2017. "Robust incremental neural semantic graph parsing." *arXiv preprint arXiv:1704.07092,* https://doi.org/10.18653/v1/P17-1112. (Cited on page 67).

Camacho-Collados, Jose, Claudio Delli Bovi, Luis Espinosa-Anke, Sergio Oramas, Tommaso Pasini, Enrico Santus, Vered Shwartz, Roberto Navigli, and Horacio Saggion. 2018. "SemEval-2018 Task 9: Hypernym Discovery." In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018).* New Orleans, LA, United States: Association for Computational Linguistics. https://doi.org/10.18653/v1/S18-1115. (Cited on page 182).

Carroll, J. D., and J. J. Chang. 1970. "Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition." *Psychometrika* 35:283–319. https://doi.org/10.1007/BF02310791. (Cited on page 165).

Chang, Haw-Shiuan, Ziyun Wang, Luke Vilnis, and Andrew McCallum. 2018. "Distributional inclusion vector embedding for unsupervised hypernymy detection." In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Tech- nologies, Volume 1 (Long Papers),* 485–495. Association for Computational Linguistics, June. https://doi.org/10.18653/v1/N18-1045. (Cited on pages 3, 180, 239).

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax.* MIT Press. https://doi.org/10.21236/AD0616323. (Cited on page 36).

———. 1970. "Remarks on nominalization." In *Readings in English Transformational Grammar,* edited by R. Jacobs and P. Rosenbaum, 184–221. Waltham, MA: Blaisdell. (Cited on page 69).

Church, Kenneth W., and Patrick Hanks. 1990. "Word association norms, mutual information, and lexicography." *Computational Linguistics* 16 (1): 22–29. (Cited on page 94).

Cilibrasi, Rudi, and Paul Vitányi. 2004. *Automatic Meaning Discovery Using Google.* (Cited on page 99).

Cimiano, Philipp, Andreas Hotho, and Steffen Staab. 2005. "Learning concept hierarchies from text corpora using formal concept analysis." *Journal Artificial Intelligence Research (JAIR)* 24 (1): 305–339. https://doi.org/10.1613/jair.1648. (Cited on page 181).

Coenen, Andy, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg. 2019. "Visualizing and Measuring the Geometry of BERT." *arXiv preprint arXiv:1906.02715,* (cited on pages 5, 137, 138).

Collins, A.M., and E.F. Loftus. 1975. "A spreading-activation theory of semantic processing." *Psychological Review* 82:407–428. https://doi.org/10.1037/0033-295X.82.6.407. (Cited on pages 1, 10, 17–20, 75).

Collobert, Ronan, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. "Natural Language Processing (Almost) from Scratch." *Journal of Machine Learning Research (JMLR),* (cited on pages xiv, 107, 108, 195, 197).

Conneau, Alexis, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. "What you can cram into a single \$&!#* vector: Probing sentence embeddings for linguistic properties." In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 2126–2136. Melbourne, Australia: Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1198. (Cited on page 137).

Contestabile, Monica. 2016. *The past, present and future of the PhD thesis.* https://doi.org/10.1002/wilm.10487. (Cited on page viii).

Copestake, Ann, Guy Emerson, Michael Wayne Goodman, Matic Horvat, Alexander Kuhnle, and Ewa Muszyńska. 2016. "Resources for building applications with dependency minimal recursion semantics." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16),* 1240–1247. (Cited on page 67).

Copestake, Ann, Dan Flickinger, Carl Pollard, and Ivan Sag. 2005. "Minimal Recursion Semantics: An Introduction." *Research on Language and Computation* 3:281–332. https://doi.org/10.1007/s11168-006-6327-9. (Cited on page 67).

Dahl, George E, Dong Yu, Li Deng, and Alex Acero. 2011. "Large vocabulary continuous speech recognition with context-dependent DBN-HMMs." In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on,* 4688–4691. IEEE. https://doi.org/10.1109/ICASSP.2011.5947401. (Cited on pages 124, 179).

Deerwester, Scott C., Susan T Dumais, and Richard A. Harshman. 1990. "Indexing by latent semantic analysis." *Journal of the American Society for Information Science* 41 (6): 391–407. https://doi.org/https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<3C391::AID-ASI1>3E3.0.CO;2-9. (Cited on pages 93, 113, 114).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *arXiv preprint arXiv:1810.04805* (October 11, 2018). https://doi.org/10.18653/v1/N19-1423. arXiv: 1810.04805v1 [cs.CL]. (Cited on page 126).

———. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proc. of NAACL.* (Cited on page 126).

Diederich, Paul Bernard. 1939. *The frequency of Latin words and their endings.* The University of Chicago Press. (Cited on page 77).

Dinu, Georgiana, Angeliki Lazaridou, and Marco Baroni. 2015. "Improving Zero-shot Learning by Mitigating the Hubness Problem." ICLR 2015, Workshop Track. arXiv: 1412.6568 [cs.CL]. (Cited on pages 124, 216, 228).

Döbrössy, Bálint, Márton Makrai, Balázs Tarján, and György Szaszák. 2019. "Investigating Sub-Word Embedding Strategies for the Morphologically Rich and Free Phrase-Order Hungarian." In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019),* 187–193. Florence, Italy: Association for Computational Linguistics, August. https://doi.org/10.18653/v1/W19-4321. (Cited on pages 119, 120, 206, 211).

Doddapaneni, Sumanth, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2021. "A Primer on Pretrained Multilingual Language Models." *CoRR* abs/2107.00676. arXiv: 2107.00676. https://arxiv.org/abs/2107.00676. (Cited on page 136).

Domingos, Pedro. 2012. "A few useful things to know about machine learning." In *Communications of the ACM,* 55:78–87. ACM New York, NY, USA, October. https://doi.org/10.1145/2347736.2347755. (Cited on page 196).

Dowty, David. 1991. "Thematic Proto-Roles and Argument Selection." *Language* 67 (3): 547–619. https://doi.org/10.2307/415037. (Cited on page 54).

Dumais, Susan T, George W Furnas, Thomas K Landauer, Scott Deerwester, and Richard Harshman. 1988. "Using latent semantic analysis to improve access to textual information." In *Proceedings of the SIGCHI conference on Human factors in computing systems,* 281–285. https://doi.org/10.1145/57167.57214. (Cited on page 95).

Endres, Dominik, Peter Földiák, and Uta Priss. 2010. "An Application of Formal Concept Analysis to Semantic Neural Decoding." Reviewed, *Annals of Mathematics and Artificial Intelligence* 57, nos. 3-4 (July): 233–248. https://doi.org/10.1007/s10472-010-9196-8. (Cited on pages 182, 183).

Ettinger, Allyson. 2020. "What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models." *Transactions of the Association for Computational Linguistics* 8:34–48. https://doi.org/10.1162/tacl_a_00298. (Cited on pages 127, 139–141).

Faruqui, Manaal, Jesse Dodge, Sujay Jauhar, Chris Dyer, Ed Hovy, and Noah Smith. 2015. "Retrofitting Word Vectors to Semantic Lexicons." In *Proceedings of NAACL 2015.* Best Student Paper Award. https://doi.org/10.3115/v1/N15-1184. (Cited on pages 55, 116, 181).

Faruqui, Manaal, and Chris Dyer. 2014. "Improving Vector Space Word Representations Using Multilingual Correlation." In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics,* 462–471. Gothenburg, Sweden: Association for Computational Linguistics. https://doi.org/10.3115/v1/E14-1049. (Cited on pages 116, 228).

Faruqui, Manaal, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. "Problems With Evaluation of Word Embeddings Using Word Similarity Tasks." https://doi.org/10.18653/v1/W16-2506. (Cited on page 124).

Fillmore, Charles. 1968. "The case for case." In *Universals in Linguistic Theory,* edited by E. Bach and R. Harms, 1–90. New York: Holt / Rinehart. (Cited on pages 12, 34, 42, 148, 153).

Findler, Nicholas V., ed. 1979. *Associative Networks: Representation and Use of Knowledge by Computers.* Academic Press. (Cited on page 1).

Firth, John R. 1957. "A synopsis of linguistic theory." In *Studies in linguistic analysis,* 1–32. Blackwell. (Cited on page 93).

Flickinger, Dan. 2000. "On building a more effcient grammar by exploiting types." *Natural Language Engineering* 6 (1): 15–28. https://doi.org/10.1017/S1351324900002370. (Cited on page 67).

Frandsen, Abraham, and Rong Ge. 2019. "Understanding composition of word embeddings via tensor decomposition." In *7th International Conference on Learning Representations, ICLR 2019.* ArXiv preprint arXiv:1902.00613. May. https://openreview.net/forum?id=H1eqjiCctX. (Cited on page 160).

Fried, Daniel, Tamara Polajnar, and Stephen Clark. 2015. "Low-Rank Tensors for Verbs in Compositional Distributional Semantics." In *ACL.* https://doi.org/10.3115/v1/P15-2120. (Cited on page 160).

Furnas, Susan T Dumais George W, Thomas K Landauer, Scott Deerwester, and Richard Harshman. 1988. "Using latent semantic analysis to improve access to textual information." In *Proceedings of the SIGCHI conference on Human factors in computing systems.* Citeseer. (Cited on page 95).

Fyshe, Alona, Leila Wehbe, Partha P Talukdar, Brian Murphy, and Tom M Mitchell. 2015. "A compositional and interpretable semantic space." In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 32–41. https://doi.org/10.3115/v1/N15-1004. (Cited on page 181).

Ganter, Bernhard, and Rudolf Wille. 2012. *Formal concept analysis: mathematical foundations.* Springer Science & Business Media. (Cited on page 182).

Gerz, Daniela, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. "SimVerb-3500: A Large-Scale Evaluation Set of Verb Similarity." In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing,* 2173–2182. Austin, Texas: Association for Computational Linguistics, November. https://doi.org/10.18653/v1/D16-1235. (Cited on page 118).

Gittens, Alex, Dimitris Achlioptas, and Michael W. Mahoney. 2017. "Skip-Gram – Zipf + Uniform = Vector Additivity." In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 69–76. Vancouver, Canada: Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1007. (Cited on page 161).

Gladkova, Anna, and Aleksandr Drozd. 2016. "Intrinsic Evaluations of Word Embeddings: What Can We Do Better?" In *Proc. RepEval (this volume),* edited by Omer Levy. ACL. https://doi.org/10.18653/v1/W16-2507. (Cited on pages 200, 225).

Glavaš, Goran, and Ivan Vulić. 2018. "Explicit Retrofitting of Distributional Word Vectors." In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 34–45. Melbourne, Australia: Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1004. (Cited on page 117).

Goddard, Cliff. 2002. "The search for the shared semantic core of all languages." In *Meaning and Universal Grammar – Theory and Empirical Findings,* edited by Cliff Goddard and Anna Wierzbicka, 1:5–40. Benjamins. https://doi.org/10.1075/slcs.60.07god. (Cited on page 70).

Goddard, Cliff, and Anna Wierzbicka, eds. 1994. *Semantic and Lexical Universals.* John Benjamins Publishing Company. https://doi.org/10.1075/slcs.25. (Cited on pages 13, 35, 37).

Goldberg, Adele. 2006. *Constructions at Work: The Nature of Generalization in Language.* Oxford University Press, USA. https://doi.org/10.1093/acprof:oso/9780199268511.001.0001. (Cited on page 132).

Goldberg, Yoav, and Omer Levy. 2014. "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method." *arXiv preprint arXiv:1402.3722,* (cited on page 109).

Gove, Philip Babcock, ed. 1961. *Webster's Third New International Dictionary of the English Language, Unabridged.* G. & C. Merriam. (Cited on page 71).

Grice, Paul, and Peter Strawson. 1956. "In defense of a dogma." *The Philosophical Review* 65:148–152. https://doi.org/10.2307/2182828. (Cited on page 86).

Gruber, Jeffrey Steven. 1965. "Studies in lexical relations." PhD diss., Massachusetts Institute of Technology. (Cited on pages 12, 40).

Gruber, Thomas R, et al. 1993. "A translation approach to portable ontology specifications." *Knowledge acquisition* 5 (2): 199–220. https://doi.org/10.1006/knac.1993.1008. (Cited on page 87).

Gurevych, Iryna, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M Meyer, and Christian Wirth. 2012. "UBY – A large-scale unified lexical-semantic resource based on LMF." In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics,* 580–590. (Cited on page 66).

Halácsy, Péter, András Kornai, László Németh, András Rung, István Szakadát, and Viktor Trón. 2004. "Creating open language resources for Hungarian." In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004),* 203–210. ELRA. (Cited on pages 77, 202, 204, 208, 216, 229).

Han, Lejia, and John C. Bancroft. 2010. "Nearest approaches to multiple lines in n-dimensional space." In *CREWES Research Report,* vol. 22. University of Calgary. (Cited on page 196).

Harris, Zellig. 1951. *Methods in Structural Linguistics.* University of Chicago Press. (Cited on page 93).

Harris, Zellig S. 1954. "Distributional structure." *Word* 10 (23): 146–162. https://doi.org/10.1080/00437956.1954.11659520. (Cited on pages 3, 113, 180).

Harshman, R. A. 1970. "Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis." *UCLA Working Papers in Phonetics* 16:1–84. http://publish.uwo.ca/~harshman/wpppfac0.pdf. (Cited on page 165).

Hashimoto, Kazuma, Pontus Stenetorp, Makoto Miwa, and Yoshimasa Tsuruoka. 2014. "Jointly Learning Word Representations and Composition Functions Using Predicate-Argument Structures." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing,* 1544–1555. https://doi.org/10.3115/v1/D14-1163. (Cited on page 168).

Hashimoto, Kazuma, and Yoshimasa Tsuruoka. 2015. "Learning embeddings for transitive verb disambiguation by implicit tensor factorization." In *3rd Workshop on Continuous Vector Space Models and their Compositionality.* https://doi.org/10.18653/v1/W15-4001. (Cited on page 160).

Hayes, Patrick J. 1979. "The naive physics manifesto." In *Expert Systems in the Micro-Electronic Age,* edited by D. Michie, 242–270. Edinburgh University Press. (Cited on pages 11, 24, 28, 29, 75, 155).

Hays, David G. 1964. "Dependency theory: a formalism and some observations." *Language* 40(4):511–525. https://doi.org/10.2307/411934. (Cited on page 17).

Héja, Enikő, and Dávid Takács. 2012. "An Online Dictionary Browser for Automatically Generated Bilingual Dictionaries." In *Proceedings of EURALEX2012,* 468–477. (Cited on page 205).

Heringer, T. 1967. "Wertigkeiten und nullwertige Verben im Deutschen." *Zeitschrift für Deutsche Spache* 23:13–34. (Cited on page 153).

Hewitt, John, and Christopher D Manning. 2019. "A structural probe for finding syntax in word representations." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers),* 4129–4138. (Cited on pages 137, 138).

Hill, Felix, Roi Reichart, and Anna Korhonen. 2015. "Simlex-999: Evaluating semantic models with (genuine) similarity estimation." *Computational Linguistics* 41 (4): 665–695. https://doi.org/10.1162/COLI_a_00237. (Cited on pages 91, 114, 116, 118).

Hinton, Geoffrey Everest, James Lloyd McClelland, and David Everett Rumelhart. 1986. "Distributed representations." Chap. 3 in *Parallel distributed processing: Explorations in the microstructure of cognition,* edited by James Lloyd McClelland and David Everett Rumelhart, 1:77–109. Cambridge, MA: MIT Press. (Cited on page 103).

Hirsimäki, Teemu, Mathias Creutz, Vesa Siivola, and Mikko Kurimo. 2005. "Morphologically Motivated Language Models in Speech Recognition." In *Proceedings of AKRR'05, International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning,* edited by Timo Honkela, Ville Könönen, Matti Pöllä, and Olli Simula, 121–126. Espoo, Finland: Helsinki University of Technology, Laboratory of Computer / Information Science, June. http://www.cis.hut.fi/AKRR05/papers/akrr05tuulos.pdf. (Cited on page 118).

Hobbs, J.R. 2008. "Deep Lexical Semantics." *Lecture Notes in Computer Science* 4919:183. https://doi.org/10.1162/neco.1997.9.8.1735. (Cited on pages 11, 24, 28, 29).

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 9, no. 8 (November): 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735. (Cited on page xii).

Howard, Jeremy, and Sebastian Ruder. 2018. "Universal Language Model Fine-tuning for Text Classification." In *ACL*. https://doi.org/10.18653/v1/P18-1031. (Cited on page 125).

Huang, Eric, Richard Socher, Christopher Manning, and Andrew Ng. 2012. "Improving Word Representations via Global Context and Multiple Word Prototypes." In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, 873–882. Jeju Island, Korea: Association for Computational Linguistics. (Cited on pages 116, 195, 226).

Iliev, R, M Dehghani, and E Sagi. 2014. "Automated text analysis in psychology: Methods, applications, and future developments." *Language and Cognition*, https://doi.org/10.1017/langcog.2014.30. (Cited on page 95).

Ittzés, Nóra, ed. 2011. *A magyar nyelv nagyszótára III-IV*. Akadémiai Kiadó. (Cited on page 226).

Jackendoff, Ray S. 1972. *Semantic Interpretation in Generative Grammar*. MIT Press. (Cited on pages 13, 40, 72, 73).

———. 1983. *Semantics and Cognition*. MIT Press. (Cited on pages 13, 40).

———. 1990. *Semantic Structures*. MIT Press. (Cited on pages 12, 40–44, 72).

Jenatton, Rodolphe, Nicolas Le Roux, Antoine Bordes, and Guillaume Obozinski. 2012. "A Latent Factor Model for Highly Multi-relational Data." In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2*, 3167–3175. NIPS'12. Lake Tahoe, Nevada, USA: Curran Associates Inc. http://dl.acm.org/citation.cfm?id=2999325.2999488. (Cited on pages 162, 163).

Jurafsky, Dan. 2014. *Charles J. Fillmore*. https://doi.org/10.1162/COLI_a_00201. (Cited on page 51).

Kalivoda, Ágnes. 2021. "Igekötős szerkezetek a magyarban." PhD diss., Pázmány Péter Katolikus Egyetem, Bölcsészet- és Társadalomtudományi Kar, Nyelvtudományi Doktori Iskola. (Cited on page 176).

Kamp, Hans, Josef van Genabith, and Uwe Reyle. 2011. "Discourse representation theory." In *Handbook of philosophical logic*, 125–394. Springer. https://doi.org/10.1007/978-94-007-0485-5_3. (Cited on page 64).

Kartsaklis, Dimitri, and Mehrnoosh Sadrzadeh. 2014. "A Study of Entanglement in a Categorical Framework of Natural Language." In *The 11th workshop on Quantum Physics and Logic.* ArXiv:1412.8102. June. https://doi.org/10.4204/EPTCS.172.17. (Cited on pages xii, 166, 167, 169).

Katz, Jerrold J. 1987. "Common Sense in Semantics." In *New Directions in Semantics,* edited by E. Lepore, 157–234. Academic Press. (Cited on page 36).

Katz, Jerrold J., and Jerry A. Fodor. 1963. "The structure of a semantic theory." *Language* 39:170–210. https://doi.org/10.2307/411200. (Cited on pages 12, 30, 33, 34).

Kilgarriff, Adam, Pavel Rychlý, Pavel Smrž, and David Tugwell. 2004. "Sketch engine." In *Proceedings of Euralex,* edited by Geoffrey Williams and Sandra Vessier, 105–116. Lorient, Université de Bretagne-Sud, Faculté des lettres et des sciences humaines, July. (Cited on pages 164, 237).

Kipper, Karin, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. "A large-scale classification of English verbs." *Language Resources and Evaluation* 42 (1): 21–40. https://doi.org/10.1007/s10579-007-9048-2. (Cited on page 53).

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. "Moses: Open source toolkit for statistical machine translation." In *Proceedings of the 45th annual meeting of the ACL,* 177–180. Association for Computational Linguistics. (Cited on pages 205, 214).

Kolda, Tamara G, and Brett W Bader. 2009. "Tensor decompositions and applications." *SIAM review* 51 (3): 455–500. https://doi.org/10.1137/07070111X. (Cited on pages 160, 165, 166).

Koller, Alexander, Stephan Oepen, and Weiwei Sun. 2019. "Graph-based meaning representations: Design and processing." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts,* 6–11. https://doi.org/10.18653/v1/P19-4002. (Cited on pages 15, 66).

Komlósy, András. 1982. "Deep structure cases reinterpreted." In *Hungarian General Linguistics,* edited by Ferenc Kiefer, 351–385. John Benjamins. Amsterdam–Philadelphia. https://doi.org/https://doi.org/10.1075/llsee.4.11kom. (Cited on page 152).

Kornai, András. 2008. *Mathematical Linguistics.* Advanced Information and Knowledge Processing. Springer. ISBN: 9781846289859. https://doi.org/10.1007/978-1-84628-986-6. (Cited on pages 51, 72).

Kornai, András. 2010a. "The algebra of lexical semantics." In *Proceedings of the 11th Mathematics of Language Workshop,* edited by Christian Ebert, Gerhard Jäger, and Jens Michaelis, 174–199. LNAI 6149. Springer. https://doi.org/10.5555/1886644.1886658. (Cited on pages 12, 69).

———. 2010b. "The treatment of ordinary quantification in English proper." *Hungarian Review of Philosophy* 54 (4): 150–162. (Cited on page 75).

———. 2012. "Eliminating ditransitives." In *Revised and Selected Papers from the 15th and 16th Formal Grammar Conferences,* edited by Ph. de Groote and M-J Nederhof, 243–261. LNCS 7395. Springer. https://doi.org/10.1007/978-3-642-32024-8_16. (Cited on pages 10, 69, 73, 78).

———. 2019. *Semantics.* Springer Verlag. ISBN: 978-3-319-65644-1. https://doi.org/10.1007/978-3-319-65645-8. (Cited on pages 33, 67, 69, 72, 74, 77, 87, 90, 94).

———. 2023. *Vector semantics.* Springer Verlag. https://doi.org/10.1007/978-981-19-5607-2. (Cited on pages 23, 69–71, 75, 77, 91, 157).

Kornai, András, Judit Ács, Márton Makrai, Dávid Márk Nemeskey, Katalin Pajkossy, and Gábor Recski. 2015. "Competence in lexical semantics." In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics,* 165–175. Denver, Colorado: Association for Computational Linguistics. https://doi.org/10.18653/v1/S15-1019. (Cited on pages 69, 75–77, 88).

Kornai, András, Péter Halácsy, Viktor Nagy, Csaba Oravecz, Viktor Trón, and Dániel Varga. 2006. "Web-based frequency dictionaries for medium density languages." In *Proc. 2nd Web as Corpus Workshop (EACL 2006 WS01),* edited by A. Kilgariff and M. Baroni, 1–8. https://doi.org/10.3115/1628297.1628298. (Cited on page 77).

Kornai, András, and Márton Makrai. 2013. "A 4lang fogalmi szótár." In *IX. Magyar Számitógépes Nyelvészeti Konferencia,* edited by Attila Tanács and Veronika Vincze, 62–70. (Cited on pages viii, 52, 70, 75, 147, 157).

Kossaifi, Jean, Yannis Panagakis, Animashree Anandkumar, and Maja Pantic. 2016. "Tensorly: Tensor learning in python." ArXiv preprint arXiv:1610.09555, *Journal of Machine Learning Research (JMLR)* 20:1–6. (Cited on page 167).

Kovács, Ádám, Judit Ács, András Kornai, and Gábor Recski. 2020. "Better Together: Modern methods plus traditional thinking in NP alignment." In *Proc. LREC 2020,* 3635–3639. https://www.aclweb.org/anthology/2020.lrec-1.448/. (Cited on page 90).

Kovács, Ádám, Kinga Gémes, Eszter Iklódi, and Gábor Recski. 2022. *POTATO: exPlainable infOrmation exTrAcTion framewOrk.* https://doi.org/10.1145/3511808.3557196. arXiv: 2201.13230 [cs.CL]. (Cited on pages 69, 71, 90, 156).

Kovács, Ádám, Kinga Gémes, András Kornai, and Gábor Recski. 2022. "Explainable lexical entailment with semantic graphs." *Natural Language Engineering,* https://doi.org/https://www.doi.org/10.1017/S1351324922000092. (Cited on page 60).

Kovács, Ádám, and Gábor Recski. 2018. "Knowledge base population using natural language inference." In *Proceedings of the Automation and Applied Computer Science Workshop 2018 : AACS'18,* edited by Dmitriy Dunaev and István Vajk, 19–24. Budapest University of Technology / Economics. (Cited on page 156).

Kovaleva, Olga, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. "Revealing the dark secrets of BERT." *arXiv preprint arXiv:1908.08593,* https://doi.org/10.18653/v1/D19-1445. (Cited on page 129).

Krizhevsky, A., and G. Sutskever I.and Hinton. 2012. "ImageNet classification with deep convolutional neural networks." In *NIPS'2012.* (Cited on pages 124, 179).

Lahat, Dana, Tülay Adali, and Christian Jutten. 2015. "Multimodal data fusion: an overview of methods, challenges, and prospects." *Proceedings of the IEEE* 103 (9): 1449–1477. https://doi.org/10.1109/JPROC.2015.2460697. (Cited on page 166).

Landauer, T.K., P.W. Foltz, and D. Laham. 1998. "Introduction to Latent Semantic Analysis." *Discourse Processes* 25:259–284. https://doi.org/10.1080/01638539809545028. (Cited on pages 95, 96).

Landauer, Thomas K, and Susan T Dumais. 1997. "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." *Psychological review* 104 (2): 211. https://doi.org/10.1037/0033-295X.104.2.211. (Cited on pages 160, 237).

Langacker, Ronald. 1987. *Foundations of Cognitive Grammar.* Vol. 1. Stanford University Press. (Cited on page 12).

Lazaridou, Angeliki, Georgiana Dinu, and Marco Baroni. 2015. "Hubness and Pollution: Delving into Cross-Space Mapping for Zero-Shot Learning." In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers),* 270–280. Long, Oral. Beijing, China: Association for Computational Linguistics. https://doi.org/10.3115/v1/P15-1027. (Cited on page 228).

Le, Q.V., and T. Mikolov. 2014. "Distributed Representations of Sentences and Documents." In *ICML*. (Cited on page 109).

Lee, Lillian. 1999. "Distributional similarity models: Clustering vs. nearest neighbors." In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, 33–40. https://doi.org/10.3115/1034678.1034694. (Cited on page 103).

Lenat, Douglas B., and R.V. Guha. 1990. *Building Large Knowledge-Based Systems*. Addison-Wesley. (Cited on pages 31, 35, 75).

Levelt, Willem J. M., Ardi Roelofs, and Antje S. Meyer. 1999. "A theory of lexical access in speech production." *Behavioral and brain sciences* 22:1–75. https://doi.org/10.1017/S0140525X99001776. (Cited on pages 20, 74, 235).

Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press. (Cited on pages 14, 44, 159, 177).

Levy, Omer, and Yoav Goldberg. 2014a. "Dependency-Based Word Embeddings." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 302–308. Baltimore, Maryland: Association for Computational Linguistics, June. https://doi.org/10.3115/v1/P14-2050. (Cited on page 160).

———. 2014b. "Linguistic Regularities in Sparse and Explicit Word Representations." In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, 171–180. Ann Arbor, Michigan: Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-1618. (Cited on pages 109, 111, 179).

———. 2014c. "Neural Word Embedding as Implicit Matrix Factorization." In *Advances in Neural Information Processing Systems 27*, edited by Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, 2177–2185. (Cited on pages 94, 109, 110, 160).

Levy, Omer, Yoav Goldberg, and Ido Dagan. 2015. "Improving Distributional Similarity with Lessons Learned from Word Embeddings." *Transactions of the Association for Computational Linguistics* 3:211–225. https://doi.org/10.1162/tacl_a_00134. (Cited on pages 109, 111–113).

Levy, Omer, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. "Do Supervised Distributional Methods Really Learn Lexical Inference Relations?" In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 970–976. Denver, Colorado: Association for Computational Linguistics. https://doi.org/10.3115/v1/N15-1098. (Cited on pages 109, 122, 164, 200).

Li, Jiwei, and Dan Jurafsky. 2015. "Do Multi-Sense Embeddings Improve Natural Language Understanding?" In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing,* 1722–1732. Lisbon, Portugal: Association for Computational Linguistics, September. https://doi.org/10.18653/v1/D15-1200. (Cited on pages 124, 227).

Linzen, Tal. 2016. "Issues in evaluating semantic spaces using word analogies." In *RepEval.* https://doi.org/10.18653/v1/W16-2503. (Cited on page 122).

Liu, Hugo, and Push Singh. 2004. "ConceptNet—a practical commonsense reasoning tool-kit." *BT technology journal* 22 (4): 211–226. https://doi.org/10.1023/B:BTTJ.0000047600.45421.6d. (Cited on page 55).

Ljubešić, Nikola, and Tomaž Erjavec. 2011. "hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene." In *Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings,* edited by Ivan Habernal and Václav Matousek, 395–402. Lecture Notes in Computer Science. Springer. https://doi.org/https://doi.org/10.1007/978-3-642-23538-2_50. (Cited on page 204).

Luong, Minh-Thang, Hieu Pham, and Christopher D Manning. 2015. "Bilingual Word Representations with Monolingual Quality in Mind." In *Proceedings of NAACL-HLT,* 151–159. https://doi.org/10.3115/v1/W15-1521. (Cited on page 139).

Majewska, Olga, Ivan Vulić, Diana McCarthy, Yan Huang, Akira Murakami, Veronika Laippala, and Anna Korhonen. 2018. "Investigating the cross-lingual translatability of VerbNet-style classification." *Language Resources and Evaluation* 52 (3): 771–799. https://doi.org/10.1007/s10579-017-9403-x. (Cited on page 177).

Makrai, Márton. 2013. "Fogalmak fontossága a definíciós gráf vizsgálatával [Importance of concepts based on the analysis of the definition graph." In *VII. Alkalmazott Nyelvészeti Doktoranduszkonferencia,* edited by Tamás Váradi. MTA Nyelvtudományi Intézet Budapest. ISBN: 978-963-9074-59-0. http://www.nytud.hu/alknyelvdok13/proceedings13/ANyD7-Makrai-Marton.pdf. (Cited on pages 15, 71, 78, 79, 86, 147).

Makrai, Márton. 2014a. "Causality in vectors space language models." In *Spring Wind,* 6:192–200. Association of Hungarian PhD / DLA Students (DOSZ). (Cited on page 195).

———. 2014b. "Deep cases in the `4lang` concept lexicon." In *X. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2014),* edited by Attila Tanács, Viktor Varga, and Veronika Vincze, 50–57 (in Hungarian), 387 (English abstract). ISBN: 978-963-306-246-3. (Cited on pages 78, 79, 106, 147).

———. 2015. "Comparison of distributed language models on medium-resourced languages." In *XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015),* edited by Attila Tanács, Viktor Varga, and Veronika Vincze, 22–33. Szegedi Tudományegyetem Informatikai Tanszékcsoport. ISBN: 978-963-306-359-0. (Cited on pages 198, 199, 210, 211, 216).

———. 2016. "Filtering Wiktionary triangles by linear mapping between distributed models." In *LREC.* (Cited on pages 214, 234).

———. 2022. "Three-order normalized PMI and other lessons in tensor analysis of verbal selectional preferences." In *XVIII. Magyar Számítógépes Nyelvészeti Konferencia,* edited by Gábor Berend, Gábor Gosztolya, and Veronika Vincze, 105–120. Szegedi Tudományegyetem TTIK, Informatikai Intézet. ISBN: 978-963-306-848-9. (Cited on pages 161, 174).

Makrai, Márton, and Veronika Lipp. 2018. "Do multi-sense word embeddings learn more senses?" In *K + K = 120 Workshop Dedicated to László Kálmán and András Kornai on the occasion of their 60th birthdays.* (Cited on pages 118, 224, 234).

Makrai, Márton, Dávid Márk Nemeskey, and András Kornai. 2013. "Applicative structure in vector space models." In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality,* 59–63. Sofia, Bulgaria: ACL, August. http://www.aclweb.org/anthology/W13-3207. (Cited on pages 157, 180, 190).

Makrai, Márton, Ákos Máté Tündik, Balázs Indig, and György Szaszák. 2022. "Towards abstractive summarization in Hungarian." In *In XVIII. Magyar Számítógépes Nyelvészeti Konferencia.* (Cited on page viii).

Manin, Dmitrii Y. 2008. "Zipf's Law and Avoidance of Excessive Synonymy." *Cognitive Science* 32 (7): 1075–1098. https://doi.org/https://doi.org/10.1080/03640210802020003. (Cited on page 161).

Marcus, Mitchell, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. "Building a Large Annotated Corpus of English: The Penn Treebank." *Computational Linguistics* 19:313–330. https://doi.org/10.21236/ADA273556. (Cited on page 200).

Marneffe, Marie-Catherine de, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. "Universal Dependencies." *Computational Linguistics* 47, no. 2 (July): 255–308. ISSN: 0891-2017. https://doi.org/10.1162/coli_a_00402. eprint: https://direct.mit.edu/coli/article-pdf/47/2/255/1938138/coli_a_00402.pdf. (Cited on page 156).

McCann, Bryan, James Bradbury, Caiming Xiong, and Richard Socher. 2017. "Learned in translation: Contextualized word vectors." In *Advances in Neural Information Processing Systems,* 6294–6305. (Cited on page 125).

McGill, William. 1954. "Multivariate information transmission." *Transactions of the IRE Professional Group on Information Theory* 4 (4): 93–111. https://doi.org/10.1109/TIT.1954.1057469. (Cited on page 162).

McInnes, Leland, John Healy, and Steve Astels. 2017. "hdbscan: Hierarchical density based clustering." *The Journal of Open Source Software* 2, no. 11 (March). https://doi.org/10.21105/joss.00205. (Cited on pages xii, 175).

McInnes, Leland, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. "UMAP: Uniform Manifold Approximation and Projection." *The Journal of Open Source Software* 3 (29): 861. https://doi.org/10.21105/joss.00861. (Cited on pages xiv, 175).

Meyers, Adam, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. "The NomBank Project: An Interim Report." In *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL 2004,* 24–31. Boston, Massachusetts, USA: Association for Computational Linguistics, May. https://aclanthology.org/W04-2705. (Cited on page 60).

Miháltz, Márton, Csaba Hatvani, Judit Kuti, György Szarvas, János Csirik, Gábor Prószéky, and Tamás Váradi. 2008. "Methods and results of the Hungarian WordNet project." In *Proceedings of the Fourth Global WordNet Conference (GWC-2008).* Citeseer. (Cited on page 52).

Mikolov, Tomáš. 2010. *Recurrent neural network based language model.* Presentation at Google. Brno University of Technology. https://doi.org/10.21437/Interspeech.2010-343. (Cited on page 107).

Mikolov, Tomas, Kai Chen, G.s. Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." In *1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings,* edited by Y. Bengio and Y. LeCun. May. arXiv: 1301.3781 [cs.CL]. http://arxiv.org/abs/1301.3781. (Cited on pages 109, 111, 112, 162, 179, 183, 199, 200, 216, 229).

Mikolov, Tomas, Quoc V Le, and Ilya Sutskever. 2013. "Exploiting similarities among languages for machine translation." ArXiv preprint arXiv:1309.4168. (Cited on pages 109, 199, 203, 205, 206, 212, 213, 215, 226, 227, 234).

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. "Distributed Representations of Words and Phrases and their Compositionality." In *Advances in Neural Information Processing Systems 26,* edited by C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, 3111–3119. Curran Associates, Inc. https://bit.ly/39HikH8. (Cited on pages 107–109, 160, 179, 206, 226).

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig. 2013. "Linguistic Regularities in Continuous Space Word Representations." In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013),* 746–751. Atlanta, Georgia: Association for Computational Linguistics. (Cited on pages 93, 107, 109, 112, 123, 179, 199, 200, 204, 229).

Miller, George A. 1995. "WordNet: a lexical database for English." *Communications of the ACM* 38 (11): 39–41. https://doi.org/10.1145/219717.219748. (Cited on pages 52, 71, 191, 195).

Mnih, Andriy, and Geoffrey Hinton. 2007. "Three new graphical models for statistical language modelling." In *Proceedings of the 24th international conference on Machine learning,* 641–648. ACM. https://doi.org/10.1145/1273496.1273577. (Cited on page 162).

Mnih, Andriy, and Geoffrey E Hinton. 2008. "A scalable hierarchical distributed language model." *Advances in neural information processing systems* 21:1081–1088. (Cited on page 109).

———. 2009. "A scalable hierarchical distributed language model." *Advances in neural information processing systems* 21:1081–1088. (Cited on pages xii, 107, 180, 192, 195).

Mnih, Andriy, and Koray Kavukcuoglu. 2013. "Learning word embeddings efficiently with noise-contrastive estimation." In *Advances in Neural Information Processing Systems 26 (NIPS 2013).* (Cited on page 208).

Moens, Marc, and Mark Steedman. 1988. "Temporal Ontology and Temporal Reference." *Computational Linguistics* 14 (2): 15–28. https://www.aclweb.org/anthology/J88-2003. (Cited on page 64).

Mohammad, Saif, Bonnie Dorr, and Graeme Hirst. 2008. "Computing word-pair antonymy." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* 982–991. Association for Computational Linguistics. https://doi.org/10.3115/1613715.1613843. (Cited on page 116).

Montague, Richard. 1970. "English as a formal language." In *Formal Philosophy,* edited by R. Thomason, 1974:188–221. Yale University Press. (Cited on page 72).

Moravcsik, J. M. 1975. "Aitia as Generative Factor in Aristotle's Philosophy." *Dialogue* 14:622–36. https://doi.org/10.1017/S00122173 0002655X. (Cited on page 48).

Morin, Frederic, and Yoshua Bengio. 2005. "Hierarchical Probabilistic Neural Network Language Model." In *Aistats,* 5:246–252. Citeseer. (Cited on page 107).

Mrkšić, Nikola, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. "Counter-fitting Word Vectors to Linguistic Constraints." In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 142–148. San Diego, California: Association for Computational Linguistics, June. https://doi.org/10.18653/v1/N16-1018. (Cited on page 116).

Mrkšić, Nikola, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. "Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints." *Transactions of the Association for Computational Linguistics* 5:309–324. https://doi.org/10.1162/tacl_a_00063. (Cited on page 117).

Nakov, Preslav, Antonia Popova, and Plamen Mateev. 2001. "Weight functions impact on LSA performance." *EuroConference RANLP,* 187–193. (Cited on page 95).

Neelakantan, Arvind, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. "Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 1059–1069. Doha, Qatar: Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1113. (Cited on page 226).

Nemeskey, Dávid, Gábor Recski, Márton Makrai, Attila Zséder, and András Kornai. 2013. "Spreading activation in language understanding." In *Proceedings of the 9th International Conference on Computer Science and Information Technologies (CSIT 2013),* 140–143. Yerevan, Armenia: Springer. https://hlt.bme.hu/media/pdf/nemeskey_2013.pdf. (Cited on pages 10, 17, 69, 87, 90, 180).

Nemeskey, Dávid, Gábor Recski, and Attila Zséder. 2012. "Miből lesz a robot MÁV-pénztáros? [The makings of a robotic ticket-clerk." In *IX. Magyar Számítógépes Nyelvészeti Konferencia [Ninth Conference on Hungarian Computational Linguistics.* (Cited on page 90).

Nemeskey, Dávid Márk. 2017. "`emLam` – a Hungarian Language Modeling baseline." In *XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017),* 91–102. Szeged. arXiv: 1701.07880 [cs.CL]. (Cited on pages 118, 119, 211).

———. 2020. "Natural Language Processing Methods for Language Modeling." PhD diss., Eötvös Loránd University. (Cited on pages 119, 204, 211).

Ng, Andrew Y., Michael I. Jordan, and Yair Weiss. 2001. "On Spectral Clustering: Analysis and an algorithm." In *Advances in neural information processing systems,* 849–856. MIT Press. (Cited on page 193).

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, et al. 2016. "Universal Dependencies v1: A Multilingual Treebank Collection." In *Proc. LREC 2016,* 1659–1666. May. (Cited on pages 148, 167).

Niwa, Yoshiki, and Yoshihiko Nitta. 1994. "Cooccurrence vectors from corpora vs. distance vectors from dictionaries." In *Proceedings of the 15th conference on Computational linguistics-Volume 1,* 304–309. Association for Computational Linguistics. https://doi.org/10.3115/991886.991938. (Cited on page 94).

Hendrix, G. G. 1975. *Partitioned Networks for the Mathematical Modeling of Natural Language Semantics.* Technical report. Department of Computer Science, University of Texas at Austin. (Cited on page 73).

Novák, Attila. 2014. "A New Form of Humor – Mapping Constraint-Based Computational Morphologies to a Finite-State Representation." In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), Paris: European Language Resources Association (ELRA).* (Cited on page 119).

Novák, Attila, and Borbála Novák. 2018a. "Cross-lingual generation and evaluation of a wide-coverage lexical semantic resource." In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. (Cited on pages 91, 118).

———. 2018b. "Magyar szóbeágyazási modellek kézi kiértékelése." In *Magyar Számítógépes Nyelvészeti Konferencia,* 14:67–77. ISBN: 978-963-306-578-5. http://acta.bibl.u-szeged.hu/58555/. (Cited on page 206).

Novák, Attila, Péter Rebrus, and Zsófia Ludányi. 2017. "Az `emMorph` morfológiai elemző annotációs formalizmusa." In *XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017)*. Szeged. https://doi.org/10.30716/RSZ/2017/1/1. (Cited on page 119).

Novák, Attila, Borbála Siklósi, and Csaba Oravecz. 2016. "A New Integrated Open-source Morphological Analyzer for Hungarian." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016),* edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, et al. Portorož, Slovenia: European Language Resources Association (ELRA), May. ISBN: 978-2-9517408-9-1. (Cited on page 119).

Och, Franz Josef, and Hermann Ney. 2003. "A Systematic Comparison of Various Statistical Alignment Models." *Computational Linguistics* 29 (1): 19–51. https://doi.org/10.1162/089120103321337421. (Cited on pages 205, 214).

Oepen, Stephan, and Jan Tore Lønning. 2006. "Discriminant-based MRS banking." In *LREC,* 1250–1255. (Cited on page 68).

Ogden, C.K. 1944. *Basic English: A General Introduction with Rules and Grammar*. Psyche miniatures: General Series. Kegan Paul, Trench, Trubner. http://books.google.hu/books?id=1-EtAAAAYAAJ. (Cited on page 76).

Olshausen, Bruno A, and David J Field. 1997. "Sparse coding with an overcomplete basis set: A strategy employed by V1?" *Vision research* 37 (23): 3311–3325. https://doi.org/10.1016/S0042-6989(97)00169-7. (Cited on page 181).

Oravecz, Csaba, Tamás Váradi, and Bálint Sass. 2014. "The Hungarian Gigaword Corpus." In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. Reykjavik, Iceland: European Language Resources Association (ELRA). http://www.aclweb.org/anthology/L14-1536. (Cited on pages xii, 204, 208, 216, 229).

Osgood, Charles E., William S. May, and Murray S. Miron. 1975. *Cross Cultural Universals of Affective Meaning.* University of Illinois Press. (Cited on page 93).

Ostler, Nicholas. 1979. *Case-Linking: a Theory of Case and Verb Diathesis Applied to Classical Sanskrit.* MIT: PhD thesis. (Cited on page 12).

Palmer, Martha, Daniel Gildea, and Nianwen Xue. 2010. "Semantic role labeling." *Synthesis Lectures on Human Language Technologies* 3 (1): 1–103. https://doi.org/10.1007/978-3-031-02135-0_1. (Cited on pages 34, 40, 52).

Panchenko, A, E Ruppert, S Faralli, S.P Ponzetto, and C Biemann. 2018. "Building a Web-Scale Dependency-Parsed Corpus from Common Crawl." In *Proceedings of LREC 2018.* ELRA. (Cited on page 167).

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12:2825–2830. (Cited on pages 176, 185).

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. "Glove: Global Vectors for Word Representation." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 1532–1543. Doha, Qatar: Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162. (Cited on pages 110, 111, 160, 162, 229).

Peters, Matthew, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. "Deep Contextualized Word Representations." In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers),* 2227–2237. New Orleans, Louisiana: Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1202. (Cited on pages 125, 138, 179).

Polajnar, Tamara, Laura Rimell, and Stephen Clark. 2014. "Using Sentence Plausibility to Learn the Semantics of Transitive Verbs." In *NIPS Learning Semantics Workshop.* In arXiv, some minor errata fixed. (Cited on page 160).

Pustejovsky, James. 1995. *The Generative Lexicon.* MIT Press. (Cited on pages 8, 14, 46).

Pusztai, Ferenc, ed. 2003. *Magyar értelmező kéziszótár.* Akadémiai Kiadó. (Cited on page 226).

Putnam, H. 1976. "Two Dogmas Revisited." *Printed in his (1983) Realism and Reason, Philosophical Papers* 3. https://doi.org/10.1017/CBO9780511625275. (Cited on page 87).

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 101–108. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.14. (Cited on page 60).

Qiu, Xipeng, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. "Pre-trained models for natural language processing: A survey." *arXiv preprint arXiv:2003.08271*, https://doi.org/10.1007/s11431-020-1647-3. (Cited on page 125).

Quillian, M. Ross. 1968. "Word concepts: A theory and simulation of some basic semantic capabilities." *Behavioral Science* 12:410–430. https://doi.org/10.1002/bs.3830120511. (Cited on page 16).

———. 1969. "The teachable language comprehender." *Communications of the ACM* 12:459–476. https://doi.org/10.1145/363196.363214. (Cited on pages xiv, 9, 15, 30, 69, 70).

Quine, W.V. 1969. "Natural kinds." In *In Ontological Relativity and other essays*. Columbia University Press. https://doi.org/10.1007/978-94-017-1466-2_2. (Cited on page 32).

Quine, Willard van Orman. 1951. "Two dogmas of empiricism." *The Philosophical Review* 60:20–43. https://doi.org/10.2307/2181906. (Cited on page 86).

Rabanser, Stephan, Oleksandr Shchur, and Stephan Günnemann. 2017. "Introduction to Tensor Decompositions and their Applications in Machine Learning." ArXiv:1711.10781 [stat.ML], November. arXiv: http://arxiv.org/abs/1711.10781v1 [stat.ML, cs.LG]. http://arxiv.org/abs/1711.10781v1. (Cited on pages 165, 166).

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. "Improving language understanding by generative pre-training." Https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf. (Cited on pages xii, 126).

Radovanović, M, A Nanopoulos, and M Ivanović. 2010. "Hubs in space: Popular nearest neighbors in high-dimensional data." *Journal of Machine Learning Research* 11:2487–2531. https://doi.org/10.1145/1553374.1553485. (Cited on page 228).

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." *Journal of Machine Learning Research* 21 (140): 1–67. http://jmlr.org/papers/v21/20-074.html. (Cited on page 134).

Ramachandran, Prajit, Peter J Liu, and Quoc V Le. 2017. "Unsupervised pretraining for sequence to sequence learning." In *EMNLP*. https://doi.org/10.18653/v1/D17-1039. (Cited on page 125).

Rebrus, Péter, András Kornai, and Dániel Varga. 2012. "Egy általános célú morfológiai annotáció." *Általános Nyelvészeti Tanulmányok* 24. (Cited on page 119).

Recski, Gábor. 2016a. "Building Concept Graphs from Monolingual Dictionary Entries." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Portorož, Slovenia: European Language Resources Association (ELRA). ISBN: 978-2-9517408-9-1. (Cited on pages 11, 90, 156).

———. 2016b. "Computational methods in semantics." PhD diss., Eötvös Loránd University, Budapest. https://doi.org/10.15476/ELTE.2016.126. (Cited on pages 10, 14, 74, 91, 156, 174).

———. 2018. "Building concept definitions from explanatory dictionaries." *International Journal of Lexicography* 31 (3): 274–311. https://doi.org/10.1093/ijl/ecx007. (Cited on pages 11, 71, 90, 156).

Recski, Gábor, and Judit Ács. 2015. "MathLingBudapest: Concept Networks for Semantic Similarity." In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 138–142. Denver, Colorado: Association for Computational Linguistics. https://doi.org/10.18653/v1/S15-2025. (Cited on pages 90, 156).

Recski, Gábor, Gábor Borbély, and Attila Bolevácz. 2016. "Building definition graphs using monolingual dictionaries of Hungarian." In *XI. Magyar Számitógépes Nyelvészeti Konferencia [11th Hungarian Conference on Computational Linguistics*, edited by Attila Tanács, Viktor Varga, and Veronika Vincze. (Cited on pages 11, 90, 156).

Recski, Gábor, Eszter Iklódi, Katalin Pajkossy, and András Kornai. 2016. "Measuring Semantic Similarity of Words Using Concept Networks." In *Proceedings of the 1st Workshop on Representation Learning for NLP*, 193–200. Berlin, Germany: Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-1622. (Cited on pages 69, 70, 90, 91, 156).

Řehůřek, Radim, and Petr Sojka. 2010. "Software Framework for Topic Modelling with Large Corpora" [in English]. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks,* 45–50. Valletta, Malta: ELRA, May. http://is.muni.cz/publication/884893/en. (Cited on page 204).

Reisinger, Joseph, and Raymond J Mooney. 2010. "Multi-prototype vector-space models of word meaning." In *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics,* 109–117. Association for Computational Linguistics. (Cited on page 226).

Al-Rfou', Rami, Bryan Perozzi, and Steven Skiena. 2013. "Polyglot: Distributed Word Representations for Multilingual NLP." In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning,* 183–192. Sofia, Bulgaria: Association for Computational Linguistics, August. http://www.aclweb.org/anthology/W13-3520. (Cited on page 195).

Rogers, Anna, Aleksandr Drozd, and Bofang Li. 2017. "The (too many) problems of analogical reasoning with word vectors." In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017),* 135–148. https://doi.org/10.18653/v1/S17-1017. (Cited on pages 122, 123, 200).

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky. 2020. "A primer in bertology: What we know about how bert works." *arXiv preprint arXiv:2002.12327,* https://doi.org/10.18653/v1/S17-1017. (Cited on pages 126, 133, 136, 177).

Rubinstein, Ron, Michael Zibulevsky, and Michael Elad. 2008. "Efficient implementation of the K-SVD algorithm and the Batch-OMP method." *Department of Computer Science, Technion, Israel, Tech. Rep,* (cited on page 181).

Ruder, Sebastian. 2018. *NLP's ImageNet moment has arrived.* https://ruder.io/nlp-imagenet/. (Cited on page 125).

Ruhl, C. 1989. *On monosemy: a study in lingusitic semantics.* State University of New York Press. (Cited on page 71).

Rumelhart, D., and J. McClelland. 1986. "On learning the past tenses of English verbs." In *Parallel distributed processing: Explorations in the microstructure of cognition,* edited by D. Rumelhart and J. McClelland. Cambridge MA: Branford. https://doi.org/10.7551/mitpress/5236.001.0001. (Cited on page 103).

Ruppenhofer, Josef, Michael Ellsworth, Miriam R.L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2006. *FrameNet II: Extended Theory and Practice.* Distributed with the FrameNet data. Berkeley, California: International Computer Science Institute. (Cited on page 53).

Russell, Stuart, and Peter Norvig. 2002. "Artificial intelligence: a modern approach," (cited on page 48).

Rychlý, Pavel. 2008. "A Lexicographer-Friendly Association Score." In *Proceedings of Recent Advances in Slavonic Natural Language Processing,* 6–9. (Cited on page 162).

Sahlgren, M. 2006. "The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces." PhD diss., Department of Linguistics, Stockholm University. https://doi.org/10.1145/361219.361220. (Cited on pages 93, 96–98, 112).

Salton, Gerard, Anita Wong, and Chung-Shu Yang. 1975. "A vector space model for automatic indexing." *Communications of the ACM* 18 (11): 613–620. https://doi.org/10.1145/361219.361220. (Cited on page 94).

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." arXiv: 1910.01108 [cs.CL]. (Cited on page 102).

Saralegi, Xabier, Iker Manterola, and Iñaki San Vicente. 2011. "Analyzing methods for improving precision of pivot based bilingual dictionaries." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* 846–856. Association for Computational Linguistics. (Cited on page 215).

Sass, Bálint. 2011. "Igei szerkezetek gyakorisági szótára." PhD Thesis, Péter Pázmány Catholic University. (Cited on page viii).

———. 2015. "28 millió szintaktikailag elemzett mondat és 500000 igei szerkezet [28 million syntactically analyzed sentences and 500 000 verb constructions in Hungarian]." In *XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015),* edited by Tanács Attila, Varga Viktor, and Vincze Veronika, 303–308. Szegedi Tudományegyetem Informatikai Tanszékcsoport. ISBN: 978-963-306-359-0. (Cited on pages 159, 176).

Schank, Roger C. 1972. "Conceptual dependency: A theory of natural language understanding." *Cognitive Psychology* 3 (4): 552–631. https://doi.org/10.1016/0010-0285(72)90022-9. (Cited on pages 10, 21, 73, 76).

———. 1973. *The Fourteen Primitive Actions and Their Inferences.* Stanford AI Lab Memo 183. (Cited on pages 21, 22, 195).

———. 1975. *Conceptual Information Processing.* North-Holland. (Cited on pages 26, 70).

Schnabel, Tobias, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. "Evaluation methods for unsupervised word embeddings." In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing,* 298–307. https://doi.org/10.18653/v1/D15-1036. (Cited on page 124).

Schuster, Sebastian, and Christopher D. Manning. 2016. "Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16),* 2371–2378. Portorož, Slovenia: European Language Resources Association (ELRA). https://www.aclweb.org/anthology/L16-1376. (Cited on page 60).

Schütze, Hinrich. 1993. "Word Space." In *Advances in Neural Information Processing Systems 5,* edited by SJ Hanson, JD Cowan, and CL Giles, 895–902. Morgan Kaufmann. (Cited on page 98).

———. 1998. "Automatic Word Sense Discrimination." *Computational Linguistics Special-Issue-on-Word Sense Disambiguation* 24 (1). http://www.aclweb.org/anthology/J98-1004. (Cited on page 224).

Schwenk, Holger, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2019. *CCMatrix: Mining Billions of High-Quality Parallel Sentences on the WEB.* arXiv: https://arxiv.org/abs/1911.04944. https://arxiv.org/abs/1911.04944. (Cited on page 219).

Sellars, Roy Wood. 1917. *The essentials of logic.* Houghton Mifflin. (Cited on page 30).

Sen, M.U., and H. Erdogan. 2014. "Learning word representations for Turkish." In *Signal Processing and Communications Applications Conference (SIU), 2014 22nd,* 1742–1745. April. https://doi.org/10.1109/SIU.2014.6830586. (Cited on page 199).

Shannon, Claude E., and Warren W. Weaver. 1949. *The Mathematical Theory of Communication.* Urbana: University of Illinois Press. (Cited on page 162).

Sharan, Vatsal, and Gregory Valiant. 2017. "Orthogonalized ALS: A Theoretically Principled Tensor Decomposition Algorithm for Practical Use." In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017,* 3095–3104. August. http://proceedings.mlr.press/v70/sharan17a.html. (Cited on pages 160, 162, 163, 166, 176).

Sidiropoulos, Nicholas D., Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E. Papalexakis, and Christos Faloutsos. 2017. "Tensor Decomposition for Signal Processing and Machine Learning." *IEEE Transactions on signal processing* (Piscataway, NJ, USA) 65, no. 13 (July): 3551–3582. ISSN: 1053-587X. https://doi.org/10.1109/TSP.2017.2690524. (Cited on page 165).

Siklósi, Borbála, and Attila Novák. 2016. "Beágyazási modellek alkalmazása lexikai kategorizációs feladatokra." In *Proc. MSZNY 2016,* edited by Attila Tanács, Viktor Varga, and Veronika Vincze, 3–14. Szegedi Tudományegyetem. (Cited on page 118).

Sinclair, John M. 1987. *Looking up: an account of the COBUILD project in lexical computing.* Collins ELT. https://doi.org/10.1016/0022-5193(74)90110-6. (Cited on pages xi, 225).

Smith, J. Maynard. 1974. "Theory of games and the evolution of animal conflicts." *Journal of Theoretical Biology* 47:209–221. https://doi.org/10.1016/0022-5193(74)90110-6. (Cited on pages 19, 20).

Smolensky, Paul. 1990. "Tensor product variable binding and the representation of symbolic structures in connectionist systems." *Artificial intelligence* 46 (1): 159–216. https://doi.org/10.1016/0004-3702(90)90007-M. (Cited on pages 104, 105).

Socher, R., M. Ganjoo, H. Sridhar, O. Bastani, C. D. Manning, and A. Y. Ng. 2013. "Zero-shot learning through cross-modal transfer." In *International Conference on Learning Representations (ICLR 2013).* (Cited on page 116).

Somers, Harold L. 1987. *Valency and case in computational linguistics.* Edinburgh University Press. (Cited on pages 148, 151).

Sowa, JF. 1976. "Conceptual Graphs for a Data Base Interface." *Journal of Research and Development,* https://doi.org/10.1147/rd.204.0336. (Cited on pages 11, 23).

Sowa, John F. 1992. "Conceptual graphs as a universal knowledge representation." *Computers & Mathematics with Applications* 23 (2): 75–93. https://doi.org/10.1016/0898-1221(92)90137-7. (Cited on pages 23, 24).

Speer, Robert, Joshua Chin, and Catherine Havasi. 2017. "ConceptNet 5.5: An Open Multilingual Graph of General Knowledge." In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17),* 4444–4451. https://doi.org/10.1609/aaai.v31i1.11164. (Cited on page 55).

Speer, Robert, and Catherine Havasi. 2012. "Representing General Relational Knowledge in ConceptNet 5." In *LREC,* 3679–3686. (Cited on pages 55, 56).

Sra, Suvrit. 2018. "Directional statistics in machine learning: a brief review." *Applied Directional Statistics: Modern Methods and Case Studies* 225:1–12. (Cited on page 164).

Subramanian, Anant, Danish Pruthi, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Eduard Hovy. 2018. "SPINE: SParse Interpretable Neural Embeddings." *AAAI,* https://doi.org/10.1609/aaai.v32i1.11935. (Cited on page 181).

Sun, Lin, Anna Korhonen, Thierry Poibeau, and Cédric Messiant. 2010. "Investigating the cross-linguistic potential of VerbNet: style classification." In *Proceedings of the 23rd International Conference on Computational Linguistics,* 1056–1064. Association for Computational Linguistics. (Cited on page 177).

Suzuki, Ikumi, Kazuo Hara, Masashi Shimbo, Marco Saerens, and Kenji Fukumizu. 2013. "Centering Similarity Measures to Reduce Hubs." In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing,* 613–623. Seattle, Washington, USA: Association for Computational Linguistics. http://www.aclweb.org/anthology/D13-1058. (Cited on page 228).

Swadesh, Morris. 1950. "Salish internal relationships." *International Journal of American Linguistics,* 157–167. https://doi.org/10.1086/464084. (Cited on page 76).

Szécsényi, Tibor. 2019. "Argumentumszerkezet-variánsok korpusz alapú meghatározása [Corpus-based identification of Hungarian argument structure variants]." In *XV. Magyar Számítógépes Nyelvészeti Konferencia,* edited by Gábor Berend, Gábor Gosztolya, and Veronika Vincze, 315–331. Szegedi Tudományegyetem TTIK, Informatikai Intézet, January. (Cited on page 176).

Talmy, L. 1988. "Force dynamics in language and cognition." *Cognitive science* 12 (1): 49–100. https://doi.org/10.1207/s15516709cog1201_2. (Cited on pages 12, 13, 37).

Tanaka, Kumiko, and Kyoji Umemura. 1994. "Construction of a bilingual dictionary intermediated by a third language." In *Proceedings of the 15th conference on Computational linguistics-Volume 1,* 297–303. Association for Computational Linguistics. https://doi.org/10.3115/991886.991937. (Cited on pages 213–215).

Tang, Gongbo, Rico Sennrich, and Joakim Nivre. 2019. "Encoders Help You Disambiguate Word Senses in Neural Machine Translation." In *EMNLP.* https://doi.org/10.18653/v1/D19-1149. (Cited on pages 138, 139).

Tesniére, Lucien. 1959. *Élements de syntaxe structurale.* Paris: Klincksieck. (Cited on page 16).

Thorndike, Edward L. 1921. *The teacher's word book.* New York Teachers College, Columbia University. (Cited on page 76).

Tiedemann, Jörg. 2012. "Parallel Data, Tools and Interfaces in OPUS." In *LREC,* edited by Nicoletta Calzolari. Istanbul, Turkey: European Language Resources Association (ELRA), May. ISBN: 978-2-9517408-7-7. http://www.lrec-conf.org/proceedings/lrec2012/summaries/463.html. (Cited on pages 205, 214, 216).

Tomašev N., Mladenić D. 2013. "Hub Co-occurrence Modeling for Robust High-Dimensional *k*NN Classification." In *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2013. Lecture Notes in Computer Science,* edited by Blockeel H., Kersting K., Nijssen S., and Železný F., 8189:643–659. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-40991-2_41. (Cited on page 228).

Tsvetkov, Yulia, Manaal Faruqui, and Chris Dyer. 2016. "Correlation-based intrinsic evaluation of word vector representations." *arXiv preprint arXiv:1606.06710,* https://doi.org/10.18653/v1/W16-2520. (Cited on page 95).

Tucker, Ledyard R. 1966. "Some mathematical notes on three-mode factor analysis." *Psychometrika* 31 (3): 279–311. https://doi.org/10.1007/BF02289464. (Cited on page 166).

Turian, Joseph, Lev-Arie Ratinov, and Yoshua Bengio. 2010. "Word Representations: A Simple and General Method for Semi-Supervised Learning." In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics,* 384–394. Uppsala, Sweden: Association for Computational Linguistics. (Cited on pages 124, 195).

Turney, Peter D. 2006. "Similarity of semantic relations." *Computational Linguistics* 32:379–416. https://doi.org/10.1162/coli.2006.32.3.379. (Cited on pages 123, 199).

Turney, Peter D., and Patrick Pantel. 2010. "From Frequency to Meaning: Vector Space Models of Semantics." *Journal of Artificial Intelligence Research* 37:141–188. https://doi.org/10.1613/jair.2934. (Cited on pages 93, 94, 99, 101, 103, 160, 164).

Van de Cruys, Tim. 2009. "A Non-negative Tensor Factorization Model for Selectional Preference Induction." In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics,* 83–90. Athens, Greece: Association for Computational Linguistics, March. https://doi.org/10.3115/1705415.1705426. (Cited on pages 159, 160, 162, 164).

———. 2011. "Two Multivariate Generalizations of Pointwise Mutual Information." In *Proceedings of the Workshop on Distributional Semantics and Compositionality,* 16–20. Portland, Oregon, USA: Association for Computational Linguistics, June. https://www.aclweb.org/anthology/W11-1303. (Cited on pages 162–164).

Van de Cruys, Tim, Thierry Poibeau, and Anna Korhonen. 2013. "A Tensor-based Factorization Model of Semantic Compositionality." In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 1142–1151. Atlanta, Georgia: Association for Computational Linguistics, June. https://www.aclweb.org/anthology/N13-1134. (Cited on pages 160, 162–164).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention is All you Need." In *Advances in Neural Information Processing Systems 30,* edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 5998–6008. Curran Associates, Inc. arXiv: 1706.03762 [cs.CL]. http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf. (Cited on pages 126, 139).

Vendler, Zeno. 1967. *Linguistics and Philosophy.* Ithaca, NY: Cornell University Press. https://doi.org/10.7591/9781501743726. (Cited on pages 12, 47).

Villada Moirón, M. B. 2005. "Data-driven identification of fixed expressions and their modifiability." PhD diss., University of Groningen. (Cited on page 164).

Virpioja, Sami, Peter Smit, Stig-Arne Grönroos, Mikko Kurimo, et al. 2013. "Morfessor 2.0: Python implementation and extensions for Morfessor Baseline," (cited on page 120).

Vossen, P., W. Meijs, and M. den Broeder. 1989. "Meaning and structure in dictionary definitions." In *Computational lexicography for natural language processing,* 171–192. (Cited on page 50).

Vulić, Ivan, Nikola Mrkšić, and Anna Korhonen. 2017. "Cross-lingual induction and transfer of verb classes based on word vector space specialisation." *arXiv preprint arXiv:1707.06945* (Copenhagen, Denmark) (September): 2546–2558. https://doi.org/10.18653/v1/D17-1270. (Cited on page 177).

Wang, Bin, and C.-C. Jay Kuo. 2020. *SBERT-WK: A Sentence Embedding Method by Dissecting BERT-based Word Models.* https://doi.org/10.1109/TASLP.2020.3008390. arXiv: 2002.06652 [cs.CL]. (Cited on pages 142, 143).

Watanabe, Satosi. 1960. "Information theoretical analysis of multivariate correlation." *IBM Journal of research and development* 4 (1): 66–82. https://doi.org/10.1147/rd.41.0066. (Cited on page 162).

Weeds, Julie, and David Weir. 2003. "A general framework for distributional similarity." In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing,* 81–88. https://doi.org/10.3115/1119355.1119366. (Cited on pages 3, 180, 239).

Weinreich, U. 1964. "Webster's Third: A Critique of its Semantics." *International Journal of American Linguistics* 30:405–409. https://doi.org/10.1086/464799. (Cited on page 46).

Whitney, William Dwight. 1885. "The roots of the Sanskrit language." *Transactions of the American Philological Association (1869–1896)* 16:5–29. https://doi.org/10.2307/2935779. (Cited on page 77).

Wierzbicka, Anna. 1972. *Semantic Primitives.* Frankfurt: Athenäum. (Cited on pages 13, 35, 70).

———. 1985. *Lexicography and conceptual analysis.* Ann Arbor: Karoma. (Cited on pages 12, 76).

Wilks, Y., D. Fass, C. Guo, J. McDonald, T. Plate, and B. Slator. 1989. "A tractable machine dictionary as a resource for computational semantics," 193–228. (Cited on page 51).

Wilks, Yorick. 1977. "What sort of taxonomy of causation do we need for language understanding?" *Cognitive Science* 1 (3): 235–264. https://doi.org/10.1016/S0364-0213(77)80019-0. (Cited on page 26).

Winograd, T. 1972. "Understanding natural language." *Cognitive Psychology* 3 (1): 1–191. ISSN: 0010-0285. https://doi.org/10.1016/0010-0285(72)90002-3. (Cited on page 16).

Woods, William A. 1975. "What's in a link: Foundations for semantic networks." *Representation and Understanding: Studies in Cognitive Science,* 35–82. https://doi.org/10.1016/B978-0-12-108550-6.50007-0. (Cited on pages 10, 11, 22, 73).

Xing, Chao, Chao Liu, Dong Wang, and Yiye Lin. 2015. "Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation." In *NAACL,* 1005–1010. https://doi.org/10.3115/v1/N15-1104. (Cited on pages 216, 228, 230).

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. "XLNet: Generalized Autoregressive Pretraining for Language Understanding." In *Advances in neural information processing systems,* 5754–5764. arXiv: 1906.08237 [cs.CL]. https://papers.nips.cc/paper/8812-xlnet-generalized-autoregressive-pretraining-for-language-understanding.pdf. (Cited on page xiv).

Yogatama, Dani, Manaal Faruqui, Chris Dyer, and Noah A. Smith. 2015. "Learning Word Representations with Hierarchical Sparse Coding." In *ICML*. Previous version in NIPS Deep Learning and Representation Learning Workshop 2014. (Cited on page 182).

Youn, Hyejin, Logan Sutton, Eric Smith, Cristopher Moore, Jon F. Wilkins, Ian Maddieson, William Croft, and Tanmoy Bhattacharya. 2016. "On the universal structure of human lexical semantics." *PNAS* 113 (7): 1766–1771. https://doi.org/10.1073/pnas.1520752113. (Cited on pages 37, 225).

Zellers, Rowan, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. "HellaSwag: Can a Machine Really Finish Your Sentence?" *arXiv preprint arXiv:1905.07830,* https://doi.org/10.18653/v1/P19-1472. (Cited on page 28).

Zgusta, Ladislav. 1971. *Manual of lexicography.* Prague: Academia. https://doi.org/10.1515/9783111349183. (Cited on page 225).

Zhao, Peng, Gulherme Rocha, and Bin Yu. 2009. "The composite and absolute penalties for grouped and hierarchical variable selection." *The Annals of Statistics* 37(6A):3468–3497. https://doi.org/10.1214/07-AOS584. (Cited on page 182).

Zhuang, Yimeng, Jinghui Xie, Yinhe Zheng, and Xuan Zhu. 2018. "Quantifying Context Overlap for Training Word Embeddings." In *EMNLP*. https://doi.org/10.18653/v1/D18-1057. (Cited on page 164).

Zséder, Attila, Gábor Recski, Dániel Varga, and András Kornai. 2012. "Rapid creation of large-scale corpora and frequency dictionaries." In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12),* 1462–1465. Istanbul, Turkey: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/783_Paper.pdf. (Cited on pages 204, 206).

Zsibrita, János, Veronika Vincze, and Richárd Farkas. 2013. "magyarlanc: A Tool for Morphological and Dependency Parsing of Hungarian." In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2013),* 763–771. Hissar, Bulgaria: INCOMA Ltd. Shoumen. (Cited on page 120).