UNIVERSITY OF NOVA GORICA
GRADUATE SCHOOL

SEMANTICS WITHIN: THE REPRESENTATION OF MEANING THROUGH CONCEPTUAL SPACES

DISSERTATION

Gregor Strle

Mentor: prof. dr. Jelica Šumič Riha

Nova Gorica, 2012

Author's Declaration

I hereby declare that I am the sole author of this thesis and that no part of this thesis has been published or submitted for publication.

I declare that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices.

Gregor Strle

Acknowledgements

This thesis would not materialize without the following people and institutions.

I thank my mentor prof. Jelica Šumič Riha for her overall support, help and guidance over the course of my doctoral research. As any good mentor would, she helped me keep focus on the bigger agenda everytime I got stuck in the details.

I thank my supervisor prof. Peter Gärdenfors for his generous support, valuable discussions and feedback while I was working as a guest researcher at LUCS (Lund University Cognitive Science Department). I look back at my time in Lund as one of the most inspiring, with great colleagues, challenging discussions and seminars.

I thank assist. prof. Matija Marolt for valuable discussions and help with the technical aspects of implementing *SpaceWalk* in Matlab.

I thank Swedish Institute for guest scholarship grants received in 2007/08 and 2008/09.

I thank SRC SASA for a grant "for prospective researchers to work in scientific research and higher education institutions abroad", and for co-financing my doctoral studies.

Finally, and most importantly, I thank my girls Liv and Susanne for their support and putting up with me over this roller-coaster journey.

Table of Contents

List of figures

List of tables

List of abbrevations

ACT-R – (Adaptive Control of Thought—Rational)

AI – Artificial Intelligence

ANN – Artificial Neural Network

DST – Dynamical Systems Theory

LDA – Latent Dirichlet allocation

LSA – Latent Semantic Analysis

MDS – Multidimensional Scaling

NLP – Natural Language Processing

PCA – Principal Component Analysis

pLSA – probabilistic Latent Semantic Analysis

PSS – Physical Symbol System

PSSH – Physical Symbol System Hypothesis

RAAM – Recursive Auto-Associative Memory

SOM – Self Organizing Map

SVD – Singular Value Decomposition

TASA – Touchstone Applied Science Associates

TOEFL – Test of English as a Foreign Language

U-Matrix – Unified Distance Matrix

Abstract

The aim of the thesis is to propose an alternative to the existing traditional approaches of modeling semantic representations. The practical outcome of the thesis is a computer prototype for modeling lexical semantics, based on the theory of *conceptual spaces* and various methods for natural language processing.

Traditional symbolic and connectionist approaches, it is argued, offer no credible explanation of meaning and semantics and attack the problem on two different, and to a large extent, incommensurable levels. Classical symbolic approach is rule based, using top-down processing and manipulation of discrete symbolic structures to generate appropriate representations, whereas connectionism uses a bottom-up functioning of a neural network to generate distributed representations. In their critique of connectionist approach, Fodor and Pylyshyn (1988) claimed that connectionism cannot naturally account for the compositionality of language and thought, designating it as merely implementational strategy simulating the functionality of a symbol system. Abstract thought and problem solving, they argued, are highly structural everyday activities that cannot be successfully explained by connectionism. Unlike symbolic representations, neural networks simply do not have structural or methodological means to account for more abstract and hierarchical representations, and to use these same representations for further reasoning – the network does not operate upon the representations in the sense of 'being detached from', as is the case in symbolic approach, but within representational structures.

Classical symbolic approach, on the other hand, has its own set of problems. Reserving the domain of abstract thought and problem solving, symbolic approach has no answer to the challenges brought up by lower-level cognition, such as perception or bodily experience. In a classical computational system, to solve a specific problem, the decisions need to be hand-coded into the system as rules, in a top-down manner. Such system cannot represent the emergent properties of the environment, nor the bottom-up influences of lower-level cognition on higher-level cognition (van Gelder 1990). A further, more pressing problem for symbolic approach is its psychologically inadequate theory of semantics, based on the realist view of the world. By this view, concepts are discrete symbols that correspond to entities and categories in the world. Our conceptual symbol system is innate and

made meaningful via its capacity to correspond correctly to these entities and categories in the world. Our representation is representation of external reality, a mirror of logical relations independent of individual's belief, knowledge, perception, modes of understanding, or any other aspect of individual's cognition. The success of our interacting with the world depends on our ability to successfully represent this external reality. Thought then, becomes a manipulation of abstract symbols, which get their meanings via correspondence with objectively defined entities and categories. The essential features forming such categories are abstract, amodal, arbitrary elements that take on their meaning by a principle of compositionality. From a standpoint of cognitive psychology, such approach has serious logical and empirical problems. For example, experimental research has shown that categories do not conform to the rules of logic and ontological view of the world as one based on defining features. In most cases, the structure of a category is "radial"– that is, the category has some central or prototypical members with other members more or less related to these central members. Learning the meanings of words is not analogous to processing abstract symbol structures. Meaning is not defined by a set of necessary and sufficient conditions, nor is it a part of static, ontologically defined view of the world, rather, *meaning is a conceptual entity*, affected by individual's beliefs, background knowledge and context. The deterministic structure of categories and concepts, it seems, could be appropriate only in matters of mathematics and logic.

Gärdenfors' theory of *conceptual spaces* (2000) is proposed as a solution to modeling semantic representations, both as a mathematical framework for building computer applications, as well as a plausible semantic theory. The main argument goes as follows: *meanings are conceptual structures*. Since the semantic relations are inherently conceptual, they should be modeled on a conceptual level by employing *conceptual spaces*. Furthermore, the conceptual level should be taken as a mediating level between traditional symbolic and connectionist representations in order to mitigate well-known issues of both accounts.

The practical outcome of the thesis is a computer prototype, based on the theory of *conceptual spaces* coupled with various *statistical* and *probabilistic* methods for natural language processing. These are proposed as alternatives to the traditional symbolic and connectionist models. Probabilistic approach, particularly, brings fresh

air into the traditional accounts of language and cognition. It is conceptually closer to the symbolic approach (by utilizing top-down processing), but overcomes many of its vices. For one, it allows for hybridity and coupling of different representational architectures. It utilizes associative, approximating data structures and thus allows the 'environment' to influence the representational structure of the system. Furthermore, the notion of probability represents a set of top-down constraints which, taken as *inductive biases* (e.g., as the constraints on learning and memory), can account for effects in human similarity judgments (see Griffiths et al. 2008, 2010, Clark (in press)). Coupled with conceptual spaces, the proposed model offers a more flexible framework for creating and exploring semantic representations, and the effects inductive biases have on individual's representation of meaning and semantics.

The role of conceptual spaces in modeling meaning and semantics of natural languages is significant. What we get, in machine-readable form, are not only conceptual representations of objects, concepts, properties and similarity relations, but the framework that exploits the underlying quality dimensions and projects them onto conceptual space according to the mode of graded categorization. By connecting various levels of analyticity, e.g. by coupling conceptual space with the top-down and bottom-up approaches for natural language processing, such a system becomes truly hybrid.

Keywords: cognitive modeling, conceptual spaces, computer model, meaning, probability, representations, semantics

There are no shades of grey in this. Truth is, after all, a binary function.
(Doug Edwards, the long-time online marketing guy at Google)

1 Introduction

One of the main issues in cognitive science is how meaning is being represented. Most cognitive theories support the constructive approach to cognition and argue that meaning and semantics should be modeled by employing some kind of representational structure. Beyond this general idea, the opinions quickly diverge on many of the essential aspects, starting with the notion of representation, the nature of representational relation, the degree of representational complexity (e.g. levels of description) needed for modeling particular cognitive phenomenon, the significance of certain cognitive functions for explaining main features of human mind, etc. All of which over the years resulted in the two, apparently competing and to some extent incommensurable, traditional paradigms: a) *symbolic approach*, also termed *classical computationalism*, which defines cognition as computation over abstract symbolic structures, or b) *connectionist approach*, or *connectionism*, which models dynamic and emergent properties of cognition (such as perception or motor control) using artifical neural networks (ANN) and argues these should be grounded in the environment.

An alternative to the traditional representational theories are various *situated* (Clancey 1997) and *embodied* approaches to cognition (e.g., Dynamical Systems Theory or DST; Beer 1995a, Thelen and Smith 1994). In similar spirit as connectionism, the embodied approach claims cognition strongly depends on the interaction with the external world, but unlike connectionism emphasizes the claim that in many cases our cognizing is direct and unmediated, with no real use for (internal) representations (see e.g., Brooks 1991, Varela *et al.* 1991, Wheeler 1994, Thelen and Smith 1994, van Gelder 1995). Proponents of embodied approach further argue that the traditional notion of representation is only a theoretical construct with

the aim to illuminate (whether describe, explain or model) some cognitive phenomena, not a feature of human mind[1].

While embodied approach certainly poses some challenges to the more traditional symbolic and connectionist views, most of its examples refer to lower-level cognition (e.g. perception, motor control), where interaction with the environment is largely unmediated. Arguably, in such cases we might not really need to employ representations – in case of *reactive systems* (Brooks 1991, Beer 1995a) for example, a robot builds its internal model exclusively by directly interacting with the environment. Higher-level cognitive processes, on the other hand, are per se highly representational and abstract (e.g. language comprehension, planning, problem solving etc.), and hence generally evade the scope of the embodied approach (e.g. see Svensson and Ziemke 2005).

Overall, the discussion has been fruitful (e.g., Brooks 1991, Beer 1995a/b, 2003, Bickhard and Terveen 1995, Bickhard 1998, 2000, Clark 1997, 1998, Bechtel 1998, 2001, van Gelder 1995, 1998, Chemero 2000a, Dretske 1988, Grush 1997, 2004, Millikan 1984, 1996, and Ramsey 2007). According to Chemero, the embodied cognition approach "has changed the tenor of recent writings on representation: the debate has changed, in part at least, from being about how to determine the content of representations to a debate about what it is to be a representation in the first place" (2000b, p. 11). The embodied approach does signal a general warning that cognition reflects bodily experience and that explanation of cognitive phenomena cannot be isolated from our interactions with the world, nor should they necessarily involve representations. As this thesis is about representing meaning and semantics, both in the domain of abstract thought, the embodied approach is not further discussed. Moreover, in following chapters I argue that meaning and semantics of natural languages are highly representational, but cannot be properly explained by either of the traditional accounts. Gärdenfors' theory of *conceptual spaces* (2000) is proposed as a solution to modeling semantic representations, both as a mathematical framework for building computer applications, as well as a necessary theoretical input to the theory of cognitive semantics. The construction of computer prototype

[1] A more radical form of embodied approach, commonly termed as *anti-representationalism*, strongly supports non-representational alternatives to modeling cognition (e.g., Brooks 1990, 1991, Beer and Gallagher 1992, Wheeler 1994).

for modeling lexical semantics, based on the theory of *conceptual spaces* and various methods for natural language processing, is presented in the final part of the thesis.

2 Motivation, goals and methods

In the spirit of cognitive science, this is an attempt to review and implement some (predominately computational) ideas regarding cognition, and meaning and semantics in particular – a smörgåsbord of research covering areas of linguistics, philosophy, cognitive psychology and artificial intelligence (AI). A large part of the thesis is a critical analysis of traditional symbolic and connectionist approaches to modeling representations, and corresponding semantic theories, with the aim to illuminate the fundaments and set up the argument for a more plausible semantic theory. Main methodological underpinnings of the thesis, both from constructive and explanatory view, are the theory of *cognitive semantics* (Lakoff 1987, Langacker 1986, 1987, Lakoff and Johnson 1980), the theory of *conceptual spaces* (Gärdenfors 2000, 2011) and the *prototype* theory (Rosch *et al.* 1976, Rosch 1978a/b). The practical goal of the thesis is the construction of *SpaceWalk*: a computer model for representing semantics of natural languages based on the theory of *conceptual spaces*.

The proposed theory and methodology behind *SpaceWalk* is part of *cognitive semantics*. I reject both traditional symbolic and connectionist approaches in favor of Gärdenfors' theory of *conceptual spaces*, anchoring meaning in the conceptual realm of individual language user. An underlying argument is that both traditional theories are unsuitable for modeling semantics of natural languages. More importantly, classical computationalism and connectionism approach the modeling of cognition on two different, non-complementary levels. Whereas the former operates on symbolic level and aims to address higher-level cognitive processes (such as abstract thought and reasoning), the latter operates on subsymbolic (and subconceptual) level, focusing on lower-level emergent cognitive processes (such as perception and motor control). As a consequence, these architectures are incompatible and cannot be directly mapped onto each other or result in some hybrid form that could offer a more unifying cognitive theory.

My main argument goes as follows: *meanings are conceptual structures*. Since the semantic relations are inherently conceptual, they should be modeled on conceptual level by employing the theory of *conceptual spaces*. I argue that traditional realist view of semantics, supported by philosophical heritage of propositional logic, where meanings are represented as abstract symbolic structures governed by truth conditional semantics and syntactic rules, is generally flawed. Instead, and this will be emphasized throughout the thesis, *meaning*s, *concepts* and *categories* are highly imbued. Meaning is not something static and rule-governed, but largely dependent on context and conceptual and categorical knowledge. Furthermore, how we act and reason, or do anything else for that matter, is constrained by our environment: social, cultural and physical. Therefore, our perception of the world is as much constrained and influenced by our beliefs as by the 'environmental context' we live in. Any plausible cognitive theory should be able to address these issues. As we shall see, both classical computationalism and connectionism come at a cost.

It is often said that the purpose of modeling in cognitive science is both constructive and explanatory. On the constructive side I argue, that conceptual level – as a mediating level between symbolic and connectionist representations – should be employed in order to mitigate well-known issues of both traditional accounts. Practical alternatives to traditional accounts are proposed in the later parts of the thesis, where I discuss various *statistical* and *probabilistic* approaches to natural language processing and their implementation in *SpaceWalk*. Especially the latter, probabilistic models, have recently received a growing attention from cognitive science community (see e.g., Chater 2006, Chater and Brown, 2008, Chater *et al.* 2010, Griffiths *et al.* 2010, Clark 2012), as they promise to solve some of the issues unsuccessfully addressed by more traditional, statistical approaches. Hence, I argue that probabilistic approach, coupled with the theory of *conceptual spaces*, provides a superior functional and explanatory model of semantics, compared to more traditional methods. It is a novel and innovative approach to semantics, with the end result (*SpaceWalk*) immediately applicable to a wide area of systems, as well as areas of research in cognitive semantics, machine learning, knowledge representation and semantic web.

3 Structure

The thesis is divided in 4 parts. I start off by defining the general notion of representation (Part I) and divide representational genera into three forms: symbolic (language-like or logical), iconic (image schemas) and distributed (neural networks). I discuss the nature of individual representational relation as well as its content and structure. Part II is a detailed analysis of the two traditional approaches to modeling cognition, symbolic approach (or classical computationalism) and connectionism. My aim is to examine the underlying theory and discuss the validity of main arguments brought forward by proponents of each approach. The focus is on general hypotheses, rather than on more specific and peculiar instances[2]. In similar spirit, the strengths and weaknesses of both approaches in relation to language will be discussed.

Part III focuses on theories of meaning and semantics. Traces of philosophy of logic, which prevailed in traditional symbolic approach to cognition, are clearly evident in the classical realist semantics. I argue against realist semantics, proposing cognitive semantics as psychologically more plausible solution. After arguing that realist semantics has little explanatory value when considering the notion of meaning and semantics in natural languages, and discovering that the image-schematic formalisms proposed by cognitive semantics are opaque and under-defined, I argue for the theory of *conceptual spaces* as the most appropriate approach to modeling semantic structures. While sharing main tenets of cognitive semantics, conceptual spaces, unlike image-schemas, offer a precisely defined mathematical framework upon which to build and exploit these semantic structures.

Part IV discusses the construction of computer prototype for modeling lexical semantics, called *SpaceWalk*. I start with two main alternatives to the existing symbolic and connectionist models of language: *probabilistic* and *similarity-space* approaches to natural language processing. The former is a top-down approach and therefore roughly corresponds to symbolic view on modeling cognition, whereas the latter uses bottom-up processing similar to connectionist modeling. Both approaches are compared and tested based on their structural and computational characteristics,

[2]For example, when discussing classical computationalism, I won't review the variety of symbolic architectures, but instead look for fundamental ideas that are characteristic and essential to symbolic approach.

as well as on the theoretical assumptions that they bring to the discussion of semantics. I conclude with the empirical evaluation of *SpaceWalk*, along with mentioned methods for natural language processing.

Before we start I'd like to emphasize three things. First, due to the complexity of the field, none of the theories presented here should be taken as universal or self-sufficient. Each addresses problems on a different level and characteristically focuses on only a small subset of cognition. Hence, the need for an appropriate hybrid architecture involving multiple representational architectures is emphasized throughout the text. Second, I argue that cognitively realistic account of semantics is possible by implementing conceptual spaces as a mediating level between symbolic and subconceptual representations, and that such solution is also step towards building a hybrid system. Third, methods and models discussed in the thesis are taken as explanatory tools, i.e. instruments used to explore, simulate or explain particular aspects of cognition, not mechanisms of human mind. Thus, while conceptual spaces make a beautiful analogy, we do not really employ Voronoi tessellations in our thought.

PART I: REPRESENTATION

Section 1: The notion of representation

1 Introduction

The origin of the word *representation* comes from Latin *repraesentatio(n-),* from *repraesentare*: to "bring before, exhibit". According to the Oxford English Dictionary, *Representation* (*n.*) is

> a. The action or process of presenting to the mind or imagination;
>
> b. *Philos*. An image, concept, or thought in the mind, esp. as representing an object or state of affairs in the world; *spec*. a mental image or idea regarded as an object of direct knowledge and as the means by which knowledge of objects in the world may indirectly be acquired ... Also: the formation or possession of images, concepts, or thoughts in the mind, esp. as representing, or as a means of acquiring knowledge of, objects or states of affairs in the world. (OED 2011)

In cognitive science, the general notion of representation is twofold: on one hand *representation* refers to an entity that is used to represent some thing. On the other, the aim of representation is to *represent* – to *denote* the relation between representation and what it represents: a *meaning relation* of some sorts (Cummins 1989). As Cummins (1989) points out, there are two ways of understanding the notion of representation: as *representations* (vehicles carrying information) or as a *representation* (without s; a *relation* of sort between representation and what is being represented). The former presents the problem of determining which representational structures or states are used by cognitive systems to represent. The latter presents the problem of defining the relation between representations and what they are representations of. For Cummins (2002), the former is a scientific problem, the latter a philosophical (metaphysical) one. Here, I deal with both, but the emphasis is on the explanatory and functional aspects of modeling representations. The focus is mostly on computational approaches, different levels of representational complexity, and problems of modeling semantics of natural languages. I start by reviewing the general notion of *representation*, as defined by Haugeland (1991).

2 The notion of *representation*

One of the most discussed and influential reflections on the notion of *representation*, is Haugeland's article Representational Genera (1991). There, Haugeland identifies three "canonical accounts" of representational genera, *logical* or *language-like*, *distributed* and *image-like* representations:

> A genus of representation is a general kind, within which there can be more specific kinds, importantly different from one another, yet generically alike. The level of generality intended can be indicated by example. Natural languages, logical calculi, and computer programming languages, as well as numerous more specialized notations, are all interestingly different species; but they are generically alike in being broadly *language-like* or *logical* in character. By contrast, pictures, though equally representational, are not linguistic at all, even in this broad sense; rather, they, along with maps, scale models, analog computers, and at least some graphs, charts, and diagrams, are species in another genus of broadly *image-like* or *iconic* representations. So the level of generality intended for representational genera is that of logical versus iconic representations, thus broadly construed. (Haugeland 1991, p. 171)

In literature, these respectively relate to *classical symbolic approach* (computationalism), *neural networks* (connectionism; employing distributed representations) and *iconic* approach (image-schemas, mental images). Whereas the first two have been dominant paradigms over the years, the latter has served more as an underlying hypothesis in different cognitive theories (cf., Johnson-Laird 1980, Kosslyn 1981, Lakoff 1987, Langacker 1987).

The main question then is: *what is the distinctive 'essence' of each genus*?

To answer this question we first need to define essential features upon which we would be able to discriminate between representational genera. Haugeland argues that traditional approaches are based on wrong assumptions. Most standard characterizations of the notion of representation focus on "standing in for", that is, on relational structure of representations – a distinctive relation between representing tokens (representational structure) and their represented contents or *designation*. However, the essential differentia, so Haugeland, should be sought in what is represented, i.e. in the contents, not in the nature of relation itself.

2.1 Distingushing the genera by relation

The distinct characteristic of *logical* representations is their generative *compositional semantics*. Tokens such as sentences, rules, formulae etc. are complex structures – they have a recursively specifiable structure and determinate atomic constituents allowing the semantic significance of the whole to be determined by its syntax and semantic significance of its constituents. As we shall see, the contributions of possible structures and their constituents are fixed arbitrarily, but the significance of the compound object is not at all arbitrary given a particular set of atomic elements.

Iconic or *image-schematic* representations, on the other hand, represent their contents through some form of *isomorphism*, from very obvious, carrying strong *resemblance* or *similarity* to things in pictures and scale models for example, to purely mathematical or abstract, such as graphs, wiring diagrams and analog computers. There are many forms of isomorphism and the ones relevant to particular representation are those determined by the scheme to which they belong. A monochromatic picture token might represent its object as monochromatic or color, depending to which scheme it belongs to (e.g., color or grayscale). Thus, the selection of an isomorphic structure is initially arbitrary or conventional; but once fixed, the contents of particular iconic tokens are not arbitrary.

The distinctive feature of *distributed* representations is *superposition*. Each element in a network is encoded across all the token elements and the different contents are superimposed on one another – hence, distributed. Each element of the representational token in some way represents each portion of the represented contents. This requires the prior specification of representational tokens and the identification of various content portions, of what and how is being encoded within particular scheme. Once the details of particular representational scheme have been fixed, what particular token represents is not arbitrary.

2.2 Representing vs. recording

We can differentiate representational genera based on distinctive forms of representing relations: *logical* representations are characterized by *compositional semantics*, *iconic* representations by *structural isomorphism* and *distributed* representations by spread-out *superposition*. Intuitively, these distinctive relations

should be mutually exclusive (and thus sufficient) to distinguish between individual representational genera. But, as Haugeland (1991) argues, that is not the case. These distinctive relations should be understood more in terms of 'recording devices' or processes, than of respective representational genera. The distinction between something being representation or mere recording device is evident in Haugeland's condition for sameness of representational genus:

> ... if the representations of one scheme can be witlessly transformed into equivalent representations from another scheme by a general procedure, then those schemes are species of the same genus. (ibid., p. 181)

The point Haugeland makes is that cross-generic 'translation' (or transformation) requires "wits"; otherwise it is not a representation but a mere recording of primary scheme. This problem is clearly evident in the case of photographed inscription. A photograph of a written description on a piece of paper preserves the original representation (or more precisely, the ability of the original to represent). But it is clearly not an image of what inscribed text was about, i.e. there is nothing in the image (read 'image-like' or 'iconic') that conveys the intended meaning of the original – the representing remains essentially logical in character. Thus, the logical representation has in no sense been translated into iconic one, but merely 'recorded' onto iconic medium.

What then is the difference between recording and representing? The confusion arises from the fact, to stay with the example of the photographed inscription, that the photo both represents and records the text, but the text itself is a representation of something else – it does not follow that the photo is representing what is being represented in the text inscription, it merely records it; this issue is often overlooked. Recording is a trivial mechanical production process, like copying: it is reversible ('playback'), it can reproduce copies from the original, and it is partial in the sense only certain aspects of the original are being recorded. The aim of the recording is to preserve the "schematic type", i.e. the "type-identity" of the representation being

recorded (ibid., p. 180). There's no cognitive load, both recording and copying are completely witless, oblivious to content and ignorant to the world[3].

By contrast, representing is not *witless*. Of course, to continue with the case of photographed inscription example, there is an iconic representation of inscription on a piece of paper, but the intended meaning of the original representation is not conveyed, only recorded – the original representation is still logical in character.

The question is can representational genera be differentiated based on the nature of representational relation alone? All traditional accounts seem to focus primarily on the representational relation. For Haugeland, such criterion is not discriminatory, but rather insufficiently exclusive. The essential differentia characteristic of iconic representations, for example, is to be abstractness and isomorphism. Yet, argues Haugeland, both could be found in many other tokens.

As a counter-example, consider recursively generated maps or floor plans. What gives them semblance of being logical or language-like? An architectural drawing created with computer software is stored within system as a set of bytes and formal specifications (line-drawing commands) defined by programming language. The process of drawing is just a "witless process" of recovering the image. Thus, recursively generated maps are not logical representations, but icons recovered from logical recordings (in which the image is being stored).

In similar spirit, take for example Wittgenstein's picture theory of meaning (1922), where sentences represent worldly facts by "picturing" their logical structure. Wittgenstein argued that language is not sufficient for expressing its own logical structure – constituent parts of a proposition can correspond to some aspect of the world, but correspondence itself can only be shown (Stern 1995). Following this idea, for some logical representational scheme the picture theory could provide plausible semantics. As in previous example, the character of the scheme remains logical, but recorded iconically. A special kind of recording though: "... since there have been no actual prior sentences to record iconically, ... the recordings are of 'virtual sentences' – something like the *facts* pictured" (Haugeland 1991, p. 182).

[3] *Witlessness* is not level bound – there is no level at which a witless process could count as intelligent or sensible (for example, a certain process can be seen as intelligent on cognitive level but mechanical on lower level)

This emphasizes two things about representational relation. First, isomorphism cannot be an essential discriminating feature of iconic representations because it can apply equally well to other cases that are not necessarily iconic, the relation between sentences and facts, for example. Moreover, isomorphism can be found everywhere: chess transcripts and the game's moves are one example, music and notation, etc. Second, isomorphism might just be the way the contents are being recorded, not represented. Thus, representational relation cannot be a sufficient criterion for discriminating among genera.

2.3 Skeletal vs. fleshed-out contents

Haugeland argues that the essence of representational genera should be searched for in the nature of represented content, instead of relation. But what are the contents of representational genera? What is being represented?

The problem we need to attack first is how to differentiate the contents of representational genus from the real-life, "fleshed-out" contents. In real environment, only part of the stimulus or information (whether perceptual or conceptual) is usually represented, for the rest, our cognitive capacities, such as categorization and memory, help us 'fill-in'. These capacities are context dependent and dynamic (much experimental psychology research confirms this, e.g., Rosch 1978, Karmiloff-Smith 1992, Barsalou 1999, 2008). To get to the skeletal contents then, we need to derive to-be-represented contents from background contextual information and qualitatively differentiate these contents among genus. This formula should give us the *skeletal* contents of individual genus, stripped off of the effects of real-life environment, that is, of other representations and background information. By analyzing this substructure, one would derive the essential characteristics of particular representational genus. In essence, *skeletal* contents is seen as a foundation for *fleshed-out* contents of real-life environment – a kind of primal substructure detached from background information and any characteristics of alternative representational schemes. Take for example the skeletal contents of language as defended by symbolic paradigm: language is composite, with atomic sentences and content as atomic facts (in the spirit of formal Fregean semantics), where the meaning of an expression is a function of the meanings of its parts and of how they are syntactically combined.

2.4 Discussion

Even if we accept the premises given above some open questions remain? How can this *skeletal* content carry sufficient explanatory power? It is an abstraction, or more precisely, a substratum of certain (but always partial) information derived from analysis of cognitive processes in real-life situations, or their simulations, being further constrained by the nature of particular representational genus. This provokes related questions: Can different representational schemes combine or complement? If yes, as Haugeland argues, how do they combine? As we shall see, this is not just a technical issue. The main question is how can competing theories complement each other? And furthermore, to what extent are potential hybrid systems psychologically plausible? In attempt to construct a computer model based on psychologically plausible theory of semantics, these issues will re-emerge throughout the thesis.

As representations have a functional role of 'standing in for', they depend crucially on the general background knowledge and context. What is often overlooked when discussing representational approaches to cognition is that by back-engineering a particular cognitive process into some composite representational structure, we lose large amounts of contextual information that, in reality, in the flux of our everyday life experiences, is essential to our being and functioning in the world. The problem is contextual information and background knowledge cannot be simply reduced to a set of simple representational elements. This, of course, is the deficiency of any formal attempt of modeling cognition: modeling any significant aspect of cognition (e.g. language comprehension) inevitably produces a very partial, 'chunky' image. And, while any credible representational theory should be able to tell us something about the represented world, it should also indicate the shortcomings of individual approach, particularly as representation is taken to be explanatory primitive.

Haugeland acknowledges these problems, drawing from artificial intelligence research[4]. Using language as an example he argues, that the whole picture of the

[4] A frequent situation in using certain representational formalisms is that we tend to ascribe them more functionality and explanatory power that they really (can) carry. Good old fashioned AI or GOFAI (Haugeland 2000, p. 301) is a good example: a system using representations for navigating and reacting to the environment is often compared to human-like intelligence and motor skills. Such anthropomorphic explanations ignore a plethora of cognitive issues in everyday situations where there are many factors at play. Here, Haugeland agrees with Searle (1980, p. 288) that such system understands nothing – all these representations are external to the system and hardwired by the designer.

situation, fleshed-out contents of the situation, can be generated from the combination of representational structures:

> In other words, the full content of a discourse, in terms of which it is workable at all, is simultaneously a function of two determining factors: the skeletal content of those linguistic tokens themselves, plus whatever else the relevant sensible speakers of that language can count on one another to grasp in that context. That the latter is essential in practice does not show that the former is impossible in theory, or indeed inessential. (ibid., p. 188)

By this view, the contextual information becomes a part of some internal representational structure (a mental image, for example) of individual language user. This notion has a long history in philosophy of mind. The only difference is that, traditionally, thought "is the locus of all contentfulness" (where the contents are somehow conferred on linguistic tokens), whereas for Haugeland, this locus is in the symbiosis between (logical representation of) skeletal linguistic content and internal, mental representation of background knowledge. How exactly should such symbiosis (between different representational structures) work, underlies large part of this thesis.

It is important to note that the character of representational relation is an important discriminatory factor and, while these relations (whether *logical*, *distributed* or *image-like*) by themselves might not be sufficient criteria for discriminating among genera, they are still an important part of modeling cognitive processes. The character of representational relation inevitably constrains the contents, levels of abstraction and consequently the nature of cognitive process it aims to describe. For example, higher cognitive processing such as language use characteristically involves logical and conceptual representations, which cannot be sufficiently represented on the lower-level, dealing with perception, sensory-motor tasks etc. – the latter are, characteristically, subsymbolic and subconceptual. Moreover, there is rarely only one kind of representational structure involved.

The above discussion opened some general theoretical and functional aspects of modeling representations. Following chapters present two traditional paradigms, the classical symbolic approach and the connectionist approach, in more detail. But most of Haugeland's intuitions remain potent, especially in the context of formalism,

power and scope of individual representational approach, and the possibility of combining them into a hybrid system.

PART II: TWO PARADIGMS

Section 2: Symbolic paradigm

3 Conceptual foundations: Homo Ex Machina

There are many different flavors to symbolic approach of modeling cognition, but the common underlying theoretical foundation characteristic to all is the idea of cognition as computation. Computational theory of mind has a long and firm tradition in cognitive science, taking from philosophy of logic, mathematics and classical artificial intelligence (AI). With the metaphor of 'mind is a computer', cognitive systems are being modeled as formal symbol manipulation systems: the mind is a symbol processor and mental states are related to computational states. This underlying hypothesis is best represented by the *Turing machine*.

3.1 Can machines think?

Alan Turing's seminal paper *Computing Machinery and Intelligence* (Turing, 1950) started the debate whether human intelligence could be modeled by a digital computer. While the back-drop to this inquiry is a question *"Can machines think?"*, Turing finds this question difficult to define and instead proposes to solve it by 'imitation game', now called a Turing test. The suppletory question to "*Can machines think?*" becomes "*Can computer pass the Turing test?*". Many variations of the Turing test exist (for an overview see Akman and Blackburn 2000, Moor 2000, Rapaport 2005), but in essence, their purpose is to test the computer's ability of human-like performance in natural language conversation. The participant (human or computer) passes the test if it convinces the interrogator that it is human[5].

> The new form of the problem can be described in terms of a game which we call the 'imitation game'. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart front the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he

[5] The rules being "an average interrogator will not have more than 70 percent chance of making the right identification after five minutes of questioning" (Turing 1950, p. 442).

> says either "X is A and Y is B" or "X is B and Y is A." The interrogator is allowed to put questions to A and B thus:
>
> C: Will X please tell me the length of his or her hair?
>
> Now suppose X is actually A, then A must answer. It is A's object in the game to try and cause C to make the wrong identification. His answer might therefore be:
>
> "My hair is shingled, and the longest strands are about nine inches long."
>
> In order that tones of voice may not help the interrogator the answers should be written, or better still, typewritten. The ideal arrangement is to have a teleprinter communicating between the two rooms. Alternatively the question and answers can be repeated by an intermediary. The object of the game for the third player (B) is to help the interrogator. The best strategy for her is probably to give truthful answers. She can add such things as "I am the woman, don't listen to him!" to her answers, but it will avail nothing as the man can make similar remarks.
>
> We now ask the question, "What will happen when a machine takes the part of A in this game?" Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, "Can machines think?" (Turing 1950, p. 433-4).

Turing test is an empirical test of a machine's ability to exhibit intelligent behavior – a simple but powerful examination underlying some of AI and philosophical hypotheses about machines imitating human intelligence[6]. Propositional character of language takes the main stage, and "the question and answer method seems to be suitable for introducing almost any one of the fields of human endeavor that we wish to include." (Turing 1950, p. 435).

But, what kind of machine could perform such a test? And further, what kind of device could perform any computation whatsoever? What follows, is an explanation of computation as a mechanical procedure.

[6] The question whether passing a Turing test is a sufficient demonstration of cognition has met a strong criticism from a number of scholars (for example Dreyfuss (1972, 1992), Harnad (1990), but most notably Searle (1980)). Since universal Turing machine is the fundamental ingredient of computational approach, its critique is targeting computational approach in general.

3.2 The Turing machine

Turing (1936, 1938) designed a very simple device (now called the Turing machine) with a finite read/write control head that operates on an unbounded tape and can do four things: move the tape in both directions, read (a symbol on the tape), write/overwrite (a symbol on the tape), and change its 'internal' state. The input is given in binary form on the machine's tape (divided into squares with or without symbols) and the output consists of the contents of the tape when the machine halts (stops operation). The idea is to decompose an object's behavior into finite, easily manageable 'chunks' or states. At any given point the machine is in one of its states. The possible operations are represented by instructions, table of rules or machine's 'machine table'. The machine table can be thought of as the machine's 'program': it tells machine what to do. The upshot of operating on formal symbolic encodings is that we can further encode the operations (machine table) and the contents of the tape of any Turing machine into strings and feed them to another Turing machine. We get a universal Turing machine that can simulate any other machine (Figure 1).

> It is possible to invent a single machine which can be used to compute any computable sequence. If this machine I is supplied with a tape on the beginning of which is written the S.D ["standard description" of a machine table] of some computing machine M, then I will compute the same sequence as M. (Turing 1936, p. 241-2)

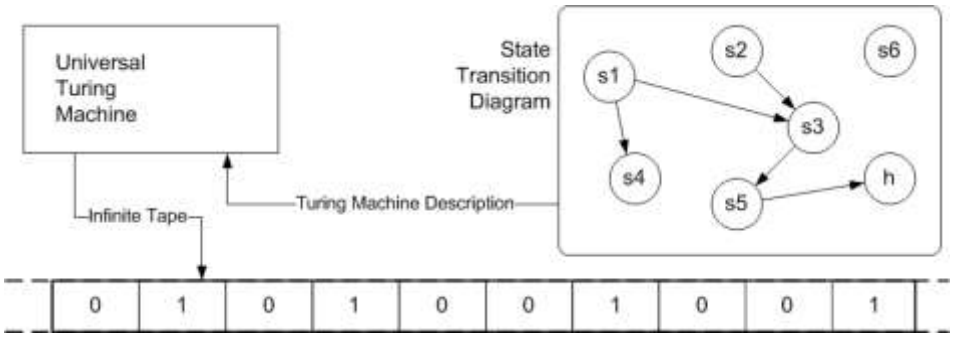


Figure 1: Universal Turing machine

The Turing machine is not a physical machine, but an abstract (and universal) theoretical specification of a possible machine, a mathematical tool equivalent to the

digital computer[7]. The essence of universal Turing machine is in its simplicity[8] and power: it serves as a standard against which to compare computational systems.

Turing's goal was to define general properties of algorithms and computations characteristic of any computer and to define the limits of computation and the capacities of physical computing machines, later presented in the Church-Turing thesis[9]. But the underlying theoretical foundation was purely mathematical. It started with the discovery of paradoxes in Cantor's set theory (and later Russell's Paradox; for an overview see (Anellis and Drucker 1991)) and Hilbert's attempt of redefining mathematics as a study of formal systems. Hilbert aimed to avoid paradoxes by axiomatization of classical arithmetic, stripping off the traditional contents of mathematics into purely formal system, in order to construct a complete formal theory of classical arithmetic. The main problem became finding a definite finitary formal procedure that could be used to unequivocally decide the provability of any claim in formalized mathematics (Cleland 1995, 2004). This decision problem later became known as Hilbert's *Entscheidungsproblem*.

Thus, it is important to note that, initially, Turing built his machines to solve *Entscheidungsproblem* (see his early paper (Turing: 1936)), and Hilbert's hypothesis is strongly reflected in Turing's account of computation: computation is processing of symbols as formal meaningless structures, such processing must be fixed in advance, and there can only be a finite number of steps in any computation. Only later has Turing analysis been extended to computational capacities of physical machines, outlined in the Church-Turing thesis. According to the Church-Turing thesis (Church 1936, Turing 1936, Kleene 1967) all possible number theoretic functions which can be computable, can be computable by universal Turing machine. Later, the thesis has been extended (with no restriction to the number theoretic functions) to define the limit of computation in general (see Minsky 1967, p. 132-

[7] Modern computer is much like Turing machine, except that computers have finite memory while Turing machine has infinite memory. Turing machine operates with a movable read/write head on an unbounded storage tape; if we restrict the head to move in only one direction and operate on finite tape we get a finite automata or a finite-state machine FSM (a modern computer could be thought of as a large network of finite-state machines).

[8] Numerous attempts to define smallest possible universal Turing machine have been made (see for example Shannon 1956, Minsky 1956, 1962, Rogozhin 1996)

[9] For a great philosophical view on Turing machine, see Crane (2003).

138): anything computable can be computed on a Turing machine[10]. This shift, together with psychological (and anthropomorphic) interpretation of machine's operations, set the foundations of AI.

3.3 Can thought be mechanically explained?

For beneath Turing's "Can machines think?" there is another important question lurking around: "Can thought be mechanically explained?" For Turing, both questions are intimately connected. In "trying to imitate an adult human mind" with computational processes of the machine, Turing's approach becomes purely anthropomorphic. The justification for computational explanation of mind

> ... lies in the fact that the human memory is necessarily limited. ... We may compare a man in the process of computing a real number to a machine which is only capable of a finite number of conditions.... The machine is supplied with a 'tape' … running through it, and divided into sections … each capable of bearing a 'symbol'. At any moment there is just one square ... which is 'in the machine'. We may call this square the 'scanned symbol'. The 'scanned symbol' is the only one of which the machine is, so to speak, 'directly aware'. (Turing 1936, p.231)

The anthropomorphic character of computation is evident from the following passages.

> Computing is normally done by writing certain symbols on paper. We may suppose this paper is divided into squares, like a child's arithmetic book ...
>
> The behaviour of the computer at any moment is determined by the symbols which he is observing, and his "state of mind" at that moment ...
>
> Let us imagine the operations performed by the computer to be split up into "simple operations" which are so elementary that it is not easy to imagine them further divided. Every such operation consists of some change in the physical [sic] system consisting of the computer and his tape. We know the state of the system if we know the sequence of symbols on the tape, which of these are observed by the computer (possibly with a special order), and the state of mind of the computer. We may suppose that in a simple operation not more than one symbol is altered. Any other changes can be set up into simple changes of this kind... .

[10] Not all agree with such loose interpretation of Church-Turing thesis (see Copeland 1998 and 2002, Cleland 1993). The precise formulation of the Church-Turing thesis does not account for what can be calculated by a machine. Rather, it states that whatever can be computed by a mathematician working in accordance with 'mechanical' methods (that is, given a finite number of instructions, and being unaided by machinery)using only paper and pencil, can also be computed by a Turing machine (Turing 1948, p.9).

> The operation actually performed is determined ... by the state of mind of the computer and the observed symbols. In particular, they determine the state of mind of the computer after the operation is carried out.
>
> We may now construct a machine to do the work of this computer... .
>
> We suppose ... that the computation is carried out on a tape; but we avoid introducing the "state of mind" by considering a more physical and definite counterpart of it. It is always possible for the computer to break off from his work, to go away and forget all about it, and later to come back and go on with it. If he does this he must leave a note of instructions (written in some standard form) explaining how the work is to be continued. This note is the counterpart of the "state of mind". We will suppose that the computer works in such a desultory manner that he never does more than one step at a sitting. The note of instructions must enable him to carry out one step and write the next note. Thus the state of progress of the computation at any stage is completely determined by the note of instructions and the symbols on the tape. (Turing 1936, p. 250-4)

This is a general praxis of advocating computationalism, with the quasi-cognitive terms being ascribed to the operations of the machine: the machine has a 'state of mind' (it is in a certain state), it 'observes' the environment (the symbols) and 'behaves' accordingly (to the rules of operation), it can 'forget' (erase symbol, change state), 'go on with it' etc.

By such view, human calculation is purely mechanical and devoid of any cognitive content: as Dennett (1986) pointed out, the Turing machines "presuppose no intelligence" (p. 83). Thought processes are broken down into a series of smaller, easily definable and mechanically 'traceable' steps: calculation depends on our brains following a set of simple mechanical rules and sub-rules, which are such that they can also be followed by a machine. In the process of computing, the 'machine' could be at any point replaced by the 'human machine' and vice versa. Intelligence-like-behavior emerges from the overall complexity of the system: machine's ability to simulate the creative aspect of human problem-solving (see Turing 1947, p. 103–4). Thus, the early AI became "[t]he science of making machines do things that would require intelligence if done by men" (Minsky 1968, p. V).

4 Machine as a psychological paradigm

4.1 Physical Symbol Systems Hypothesis

The science of making intelligent machines started with the reformulation of Turing's idea of mechanical symbol manipulation systems. In 1972, and later in (Newell and Simon 1976, Newell 1980), Newell and Simon laid the foundations with the definition of Physical Symbol Systems (PSS) and Physical Symbol Systems Hypothesis (PSSH), which marked the start of classical AI.

> A physical symbol system consists of a set of entities, called symbols, which are physical patterns that can occur as components of another type of entity called an expression (or symbol structure). Thus, a symbol structure is composed of a number of instances (or tokens) of symbols related in some physical way (such as one token being next to another)... Besides these structures, the system also contains a collection of processes that operate on expressions to produce other expressions: processes of creation, modification, reproduction and destruction. A physical symbol system is a machine that produces through time an evolving collection of symbol structures. Such a system exists in a world of objects wider than just these symbolic expressions themselves. ...
>
> *The Physical Symbol Systems Hypothesis*. A physical symbol system has the necessary and sufficient means for general intelligent action. (Newell and Simon 1976, p. 116)

Newell and Simon were not interested in philosophical issues, rather, their mission is empirical: to apply and test computational models in domain of cognitive science, and anchor computational theory of mind as a prevailing approach within AI and cognitive psychology. PSSH is an empirical hypothesis with aim to define a universal class of systems capable of intelligent behavior[11]. As they point out, "[n]ot only are psychological experiments required to test the veridicality of the simulation models as explanations of the human behavior, but out of the experiments come new ideas for the design and construction of physical symbol systems" (Newell and Simon 1976, p. 120).

For Newell and Simon "the symbolic behavior of man arises because he has the characteristics of a physical symbol system" (ibid., p. 119). For PSSH, human

[11] This agenda is similar to Turing's definition of Universal Turing Machines. In his later, more detailed account of PSS (Newell, 1980), Newell explicitly defines PSS as universal, relative to physical limitations.

thinking succumbs to the rules of the formal logic and machine's syntactic processing of symbols – "that intelligence resides in physical symbol systems [becomes] ... computer sciences' most basic law of qualitative structure" (ibid., p. 125). The main research paradigm became human problem solving, decision making, routine action, inductive behavior, long-term memory research etc. – anything that could be successfully modeled with the syntactic representational structure of PSS.

4.2 Some aspects of PSS architecture

4.2.1 The heuristics of human problem solving

Chess became the natural environment for studying the processes the human mind employs when solving problems. In his classical analysis of chess thinking, De Groot (1965) defines the four phases of problem solving that are very similar to machines' processing:

> 1. The First Phase of Orientation, especially orientation to possibilities. What we find here is largely 'looking at' the consequences of moves and general possibilities in a certain direction.
>
> 2. The Phase of Exploration. The subject tries out rather than 'investigates' possibilities for action. He calculates a few moves deep a few sample variations, or what he considers to be the main variation; if these are unsatisfactory he puts the move(s) in question temporarily aside.
>
> 3. The Phase of Investigation. There is a deeper, more serious search for possibilities, strengthenings, etc., that are quantitatively and qualitatively quite sharply defined. The investigation is more directed and much more exhaustive: more variants are calculated and they are calculated more deeply.
>
> 4. The Final Phase of Proof. The subject checks and recapitulates, he strives for proof; the obtained results are made into a subjectively convincing argument. A certain completeness is sought in the calculation of results, be it for the positive or negative part. (p. 267)

In PSS, this 'mechanical' notion of human reasoning is defined as a function of the heuristics of human problem solving and as a function of the complexity of the domain in question (Newell and Simon 1959, 1972, 1976, Simon 1979). The latter defines the problem/search space as a "space of symbol structures in which problem situations, including the initial and goal situations can be represented" (Newell and Simon 1976, p.121) – for any step in problem solving there is a search space. The

idea of the search space is common to all areas of AI, including problem solving, natural language processing, robotics, vision, knowledge representation, and machine learning (for an overview see (Russell and Norvig 1995)). The former, the heuristics of human problem solving is realized by searching through a problem space.

> *Heuristic Search Hypothesis*. The solutions to problems are represented as symbol structures. A physical symbol system exercises its intelligence in problem solving by search – that is, by generating and progressively modifying symbol structures until it produces a solution structure. (Newell and Simon 1976, p. 120)

4.2.2 Chunking

Classical AI then, sees human thinking as a rule-governed mental activity, highly amenable to hierarchical recursive analysis of PSS. Because of the potential for the exponential explosion of operations upon number of possible representations describing complex problem space, the heuristic search through problem space is sequential. The individual symbolic structures of possible solutions the search operates on are represented as 'chunks'. Chunks are 'episodic' records of 'knowledge'[12] collecting the experience of a system at the given time and level of the operation (Newell, 1990). Thus, chunking operates as a knowledge-transfer process, recording what one did in the prior situation and using this knowledge in further operations. This effectively reduces computational load since the system can execute the already familiar tasks without repeating instructions. The role of chunking is to narrow down the search operations on problem space on a set of relevant symbolic structures (for a specific time and the task given), and then the knowledge accumulated in the process can be used to guide the search.

> For general intelligent systems (and humans), life is a sequence of highly diverse tasks and the system has available a correspondingly large body of

[12] In *Unified Theories of Cognition* (Newell 1990), Newell discusses SOAR cognitive architecture and the role of long-term and short-term memory. While short-term memory (working memory) is explained solely by functional requirements of the system's architecture, SOAR's single long-term memory represents both episodic and semantic knowledge (Tulving 1983). New knowledge is acquired in long-term memory through an experience-based-learning mechanism called *chunking* (Laird *et al.* 1987, Rosenbloom and Newell 1986). Chunking is episodic, while semantic knowledge is abstract. The transformation from episodic to semantic knowledge is illustrated by an example of block-stacking problems (Newell 1990), which shows how episodic knowledge can be abstracted from conditions of learning situation to form semantic knowledge. As with PSS and PSSH, critics doubt both biological and psychological plausibility of SOAR's architecture.

> knowledge. ... intelligence is the ability to use the knowledge the system has in the service of the system's goals (Newell 1992, p. 428).

4.2.3 Designation and interpretation

The accumulation of the abovementioned computational ideas resulted in the attempt to define a unified cognitive theory. In Unified Theories of Cognition (Newell 1990), Newell proposed a detailed computational theory of human cognition, using computational cognitive architecture SOAR (Laird, Newell and Rosenbloom 1987) as a primary example. SOAR represented a pure symbolic approach to cognition, shaped primarily by three functional requirements: (a) exhibiting flexible, goal-driven behavior, (b) learning continuously from experience, and (c) exhibiting real-time cognition (Lewis 1996). The overall aim of SOAR was to give a unified theory of cognition by explaining a wide range of cognitive tasks (e.g. problem solving, concept acquisition, long-term memory etc.). Its architecture is basically a PSS with multiple levels of abstraction to model different cognitive tasks, from simple input-output and control-operational levels, to higher, knowledge level structures (cf. Newell 1990, Marr 1982, Pylyshyn 1984). These levels of abstraction were taken to be analogous, or at least try to respectfully resemble the structure of human cognition. Though the approaches to constructing symbolic architectures vary, such as *production systems* (proposed by Newell), *formal logics* (McCarthy 1968), *frames* by Minsky (1975), *semantic networks* by Quillian (1968), *scipts* (Schank and Abelson 1977) etc., they all share the underlying PSSH. In effect, they all have to do with internal manipulation of expressions to make the two basic functions of PSS as powerful as possible.

Since symbols are abstract, meaningless (no information is encoded) and arbitrary (any symbol can designate any entity), their connection to the external environment or to the other symbol structures is via their mode of operation: via *designation* and *interpretation*. These are two most important functional capacities of PSS (Newell and Simon 1976, p.116):

> *Designation*. An expression designates an object if, given the expression, the system can either affect the object itself or behave in ways dependent on the object. In either case, access to the object via the expression has been obtained, which is the essence of designation.

> *Interpretation*. The system can interpret an expression if the expression designates a process and if, given the expression, the system can carry out the process.

The notion of *interpretation* indicates the ability of the system to run from a description: that system's own data can be interpreted and that system can create expressions for its own behavior and then produce that behavior (Newell 1980, p. 29). On the other hand, the concept of designation is the most fundamental concept of PSS, one "which gives symbols their symbolic character ... or ... meaning" (Newell 1980, p. 26). The power of *designation* as standing-in-for some entity inside (some other symbol structure in system's memory) or outside of the system depends entirely on the nature of the process to which it is coupled[13]. Here, Newell (1980) offers a very formal definition:

> *Designation*: An entity X designates an entity Y relative to a process P, if, when P takes X as input, its behavior depends on Y. (p. 26)

According to computational approach, representation is simply another term for a structure that *designates*. These structures are "systematically built by combining atomic constituents into molecular assemblies, which (in complex cases) make up whole data structures in turn" (Clark, 1989, p.19). Thus, the formal symbolic representational structures of the PSS are compositional – they may be composed and interpreted – with "combinatorial syntax and semantics" (Fodor and Pylyshyn 1988). As such, they are amenable to the rules of formal logic. The raw materials of thought became symbol structures and syntactic operations, the fundamental constraining element from 'intangible' and 'ineffable', and the "progress was first made by walking away from all that seemed relevant to meaning and human symbols [carrying information]. We could call this the stage of formal symbol manipulation." (Newell and Simon 1976, p. 117).

The notion of *designation*, i.e. the way the representational token, symbol, pattern or structure 'stands in for' the object in the world, is an essential ingredient of representational theories of cognition. And, general differentiation between these

[13] For Newell, "designation is at the heart of universality. For one machine to behave as an arbitrary other machine, it must have symbols that designate that other." (Newell 1980, p. 27). Recall the notion of Universal Turing Machine.

theories largely depends on the nature of *designation* they employ – the type[14] of coupling between some artificial or biological system and the environment (cf. different views of Clark and Toribio 1994, Clark 1997, 1998, Brooks 1991, Beer 1995a, Bechtel 1998, 2001, van Gelder 1995, Haugeland 1991, 2000, Dretske 1988, Grush 1997, 2004, Millikan 1984, 1996, Ramsey 2007, Chemero 2009). Overall, the notion of *designation* defines whether something is a representation, the nature of representational relation and the choice of representational system. Whether Newell's (and hence symbolists) definition of *designation* poses a sufficient constraint, is still a hotly debated topic.

Here, and throughout the thesis, the notion of designation is understood as a semantic relation. And, as we shall see, it is the very character of symbol systems – the disembodied abstractness of computational approach – that will become the main target of our criticism.

4.3 Criticism

There are many problems with symbolic approach to human cognition. The discrete and disembodied nature of PSS is widely open to the skepticism about its biological and psychological plausibility. As has been argued by many (most notably by Searle 1980, Dreyfus 1972, 1992, Harnad 1990, Winograd and Flores 1987), any theory of human cognition has to deal with questions of how our cognition is grounded in the physical world (Harnad 1990) and how this grounding is represented. Not only lower-level cognition (such as perception), but most of high-level cognitive processes (for example language comprehension) depend on the context and interaction with the environment. As already noted, our knowledge is not only part of an abstract formal symbol structure, devoid of any subjective content. It is by our interaction with the environment (biological, social, cultural etc.) where, for example, language acquisition and reasoning, two fundamental and distinctively human characteristics, evolve (see Deacon 1997). None of these questions can be successfully explained by symbolic approach alone. As critics argue (e.g.,

[14] Concepts like genuine, cognitive, intelligent, meaningful, grounded, situated, embodied etc. are all strongly linked to (interpretation of) the way some representational system designates – 'stands in for' – some entity in the environment. In some theories of cognitive science, the mode of designation, or more precisely, the ability or disability of the system to decouple from the environment and still functionally carry and use representational operations, is the mark between simple (having no intelligence) or intelligent system (Clark 1998).

Smolensky 1988, Clark 1989, Brooks 1991, Churchland 1995, Beer 1995a/b, van Gelder 1995, and Wheeler 2005, among others), most of the flaws of symbolic approach are due to the underlying philosophical assumptions given by the computational approach to cognition. For example, symbol system is *restrictive* in that inconsistency is not allowed: all conditions have to be precisely specified in advance. Further, disembodied and abstract syntactic structures do not reflect the environment: symbols designate distant memory locations within PSS and any relationship within the internal structure must be explicitly quantified. This leads to the *frame problem* (McCarthy and Hayes 1969, Dennett 1987) – a situation where logical inferences upon such explicitly quantified structure lead to combinatorial and hence computational explosion[15]. Moreover, symbolic approach is unable to deal with partial, incomplete or approximate information. In reality, cognition consists of complex cooperation of dynamic and interactive processes based on a large amount of noisy and inconsistent data. Unified theories of cognition should be able to explain these phenomena. Following chapters review alternatives to symbolic approach.

[15] Different solutions to the computational (logical) aspect of frame problem have been proposed (e.g., McDermott (1987)), but philosophical/epistemological issues remain (see, Dennett 1987, Dreyfus 1992, Wheeler 2005, Wheeler 2008). The epistemological problem goes to the core of the classical symbolic approach to cognition and its underlying semantic theory (see Chapters 6 and 7). The question is: How can ongoing, context-sensitive information be captured with a set-propositional semantics of classical AI? Related to the computational aspect: How can such system drill out the information relevant to the current state?

Section 3: Connectionism

5 Introduction

Connectionism aims to explain some of the modeling issues that have not been successfully addressed by symbolic approach. It presents the alternative to the static, discrete view on cognition by taking into account dynamic changes in the environment and local context. As Chown and Kaplan (1992) waggishly remark: "the difference in approach could be characterized by saying that the classical approach fits the environment into the representation whereas the adaptive approach fits the representation to the environment" (p. 443).

Symbolic approach treats cognition as an abstract cognitive process: human thought is represented by finite structures composed of discrete and meaningless atomic symbols and arranged in accord with a finite number of syntactic relations, devoid of temporary context. As consequence, these highly structured symbolic representations (operated upon by recursive analysis) are intimately related to the storage and performance issues, since such systems typically have to employ large amounts of information to successfully solve individual tasks (Gregory 1969, Gibson 1979, Clark 1989, Rumelhart, McClelland *et al.* 1986, Fodor and Pylyshyn 1988). Heuristic search on problem spaces does release some of the tension on computational load, but objectively, the knowledge still has to be ready and fully accessible to the system. For symbolic approach then, the accrual of knowledge mantra – the more knowledge the better (see Newell 1990) – is necessary for the overall functioning of the system, but the problem of how these unlimited amounts of information could be successfully processed by humans is not addressed[16].

Modeling human information processing definitely involves static and discrete symbolic states and accrual of declarative knowledge. But it is almost an intuitive

[16] Newell's argues that "we must be able to learn from the environment, not occasionally but continuously, and not about a few specifics, but everything and every way under the sun" (Newell 1990, p. 19). Even though he later admits that humans can't deal with unlimited amounts of knowledge, this issue remains largely untouched. In his pursue for the unified theories of cognition, Newell admits: "The final risk is the rising tide of connectionism, which is showing signs of sweeping over all of cognitive science at the moment. The excitement is palpable - we are all hot on the trail of whether neuroscience and the cognitive world can finally be brought together. That is indeed an exciting prospect. But my message relates to symbolic architectures and all the other good things that connectionism sees as the conceptual frame to overthrow." (Newell 1990, p. 38)

notion that human beings cannot handle unbound amounts of information and knowledge in arbitrarily deeply embedded structures. Thus, Gregory (1969), Gibson (1979) and Clark (1989), among others, argue for a more reasonable and cognitively plausible strategy. In a flux of perceptual information that surrounds us, a system with a limited capacity (which human brain is) must treat information *differentially*. As Clark (1989) points out, in many situations we can only afford to know as much as we need to function. And Gregory, in "How so little information controls so much behavior" (1969), further argues, that we are able to function with remarkably little information since our cognitive processes are assisted by generality and simplifications of stimuli from our everyday experiences. Symbolic approach cannot successfully explain these properties.

Hence, connectionism sets to explain where symbolic approach fails: the *emergent*, *dynamic* and *distributed* properties of cognition.

> The idea of emergence in cognitive science is the contrasting idea that there are more basic or elementary processes that are really the fundamental ones, and that physical symbol systems of the kind Newell described are sometimes useful approximate characterizations which, however, have difficulties in capturing in full the context-sensitive, flexible, graded, and adaptive nature of human cognitive abilities. ...
>
> When it comes to intelligence, the real stuff consists of human success in everyday acts of perception, comprehension, inductive inference, and real-time behavior—areas where machines still fall short after nearly 60 years of effort in artificial intelligence...
>
> I do not think anyone who emphasizes the importance of emergent processes would deny that planful, explicitly goal-directed thought plays a role in the greatest human intellectual achievements. However, such modes of thought themselves might be viewed as emergent consequences of a lifetime of thought-structuring practice supported by culture and education... (McClelland 2010, p. 752-3)

5.1 Connectionist representations

Compared to symbolic approach, connectionism gives a fundamentally different way of viewing representations and processing of human cognition. The performance issues, coupled with the dynamic properties, partial information and the ability to treat information differentially, are an integral part of connectionist theory from the beginning. Thus, connectionism advocates two main ideas:

> (1) the idea that processing in a multilayered processing system is continuous, so that information accumulates gradually over time and is propagated as it builds up, and
>
> (2) the idea that this kind of continuous processing may be interactive, so that influences can be bidirectional, flowing both from higher to lower levels and from lower levels to higher levels. These ideas are well captured in the connectionist framework. They are generally not captured well in highly symbolic processing frameworks, in which the objects manipulated are discrete tokens that stand in an all-or-none fashion for some mental object. (McClelland 1988, p. 115)

According to MacLennan (1994), this proves that connectionist systems satisfy a different set of pragmatic invariants (such as flexible, robust, adaptive, and responsive), that are in some contexts more important than those of discrete symbol system[17]. Connectionist representations are continuous (meaning small errors in processing have small effects), adaptive in a way they can gradually change behavior, responsive to the environment and can use partial information and inferences in real time processing. These characteristics are essential for modeling lower-level cognitive processes, the microstructure of cognition[18].

> We view macrotheories as approximations to the underlying microstructure which the distributed model ... attempts to capture. As approximations they are often useful, but in some situations it will turn out that an examination of the microstructure may bring much deeper insight. (Rumelhart and McClelland 1986, p. 125).
>
> Subsymbolic models accurately describe the microstructure of cognition, while symbolic models provide an approximate description of the macrostructure. (Smolensky 1988, p. 11)

From theoretical perspective (e.g., Smolensky 1988, Harnad 1990), one of the fundamental advantages of connectionist representations is the ability to explain how the grounding of symbolic processes could be realized in the subsymbolic substrate. Smolensky sees connectionism as a prerequisite and prerogative to symbolic

[17] The difficulty for PSS framework to model above ideas has been demonstrated by Anderson (1983). Anderson, inspired by Newell's SOAR, set out to prove that interactive activation model of visual word perception (McClelland and Rumelhart 1981) could be simulated with the production system formalism by using his ACT-R (Adaptive Control of Thought—Rational) cognitive architecture. Extensive modifications of ACT-R, and the fact that a large part of ACT-R architecture remained unutilized, proved connectionist approach more appropriate (McClelland 1988, p. 116).

[18] Because of their neural network structure, some argue connectionist representations are 'brain-like' and tend to simulate the processing in the brain (e.g., Churchland 1989, Clark 1989, Bechtel and Abrahamsen 1991, Pulvermüller 1999). But this might be an oversimplification of the neurological facts (see Clark 1998).

modeling of higher cognitive processes. Similarly, for MacLennan (1994, p. 121), "the fact that people can handle discrete symbols more flexibly than conventional computers ... [shows, that] ... human symbolic cognition is implemented in terms of continuous subsymbolic processes, and ... can partake of the flexibility of these processes when that is advantageous". But, as we shall is in later discussion on hybrid systems, the coupling between symbolic and subsymbolic architecture doesn't quite live up to the expectations.

5.2 Connectionist architecture

Connectionist networks, neural networks or parallel distributed processing (PDP) models are different names for describing connectionist architecture. In these models, cognitive processes are being modeled through the interactions of large numbers of simple processing units. The tasks involved operate upon relatively automatic processes based on prior experience: perception (perceiving the world of objects and events and interpreting it for the purpose of organized behavior), memory (for example, retrieving contextually appropriate information from memory), intuitive semantics and language (perceiving and understanding language, natural language processing), categorization, reading, and, in general, intuitive or implicit reasoning (McClelland 1988, 1999). About the essential elements of connectionist architecture:

> Like all cognitive models, connectionist models must propose some building blocks and some organization of these building blocks. In connectionist models, the primitives are *units* and *connections*. Units are simple processing devices which take on activation values based on a weighted sum of their inputs from the environment and from other units. Connections provide the medium whereby the units interact with each other; they are weighted, and the *weights* may be positive or negative, so that a particular input will tend to excite or inhibit the unit that receives it, depending on the sign of the weight...
>
> Any particular connectionist model will make assumptions about the number of units, their pattern of connectivity to other units, and their interactions with the environment. These assumptions define the architecture of a connectionist model. The set of units and their connections is typically called a *network*. (McClelland 1988, p. 108)

Besides their focus on modeling natural cognitive tasks, there are two distinctive characteristics of connectionist networks that differentiate them from the classical symbolic models. First, inferences or solutions to a problem are discovered in a network of processing units without the explicit application of a predefined algorithm (McClelland 1999, p. 137). The units are "...truly active, in the sense that they give rise to further processing activity directly, without any need for a central processor or a production-matching-and-application mechanism that examines them and takes action on the basis, of the results of this examination" (McClelland 1988, p. 109). Another distinctive characteristic of connectionist representations are patterns of activation. Patterns of activation in the connectionist networks are in some way similar to symbolic representations, if we take the latter as patterns of 0s and 1s. But, as McClleland (1988) points out, there is a difference: connectionist representations are in general graded, "in the sense that each unit's activation need not be one of two binary values... typically each unit may take on a continuous activation value between some maximum and minimum" (ibid., p. 109). Thus, all connectionist representational structures are emergent and tightly coupled with the processing of the input data from the environment.

5.1.1 Supervised and unsupervised learning

Many different connectionist architectures exist, from simple to very complex[19]. These architectures differ in the organization and number of units, layers and interconnections among them. For example, a *feed forward* network has multiple layers and restrictions on connectivity among layers; the processing of units is directed forward through a series of layers. On the other hand, *fully recurrent* networks have usually no restrictions on connectivity, whereas *simple recurrent* networks introduce some restrictions to allow for certain dependencies among successive inputs, for example temporal dependencies for learning complex temporal tasks (Elman 1990a). While these architectures differ in structure and learning procedures, what they all have in common are general constraints, such as *internal*

[19] There is a vast amount of literature on connectionist architectures (see the classic (Rumelhart *et al.* 1986) for a systematic study of connectionist modeling of a wide range of cognitive phenomena). Any detailed analysis of connectionist architectures is out of scope of this thesis. My intention here is to use some well-known examples of connectionist modeling and examine some of their general properties, for two reasons: a) to illustrate the basic principles behind connectionist approach to language, and b) to evaluate the significance of these principles within symbolic vs. connectionist discussion. The overall aim, of course, is to set the stage for the third theoretically and functionally necessary ingredient: conceptual spaces.

coverage, *input and output*, *values of connections*, *weighing* and *processing* (Figure 2 diagrams a general architecture of neural network). *Internal coverage* restricts the extent to which individual units represent stimuli from environment or particular conceptual objects (letters, words, concepts etc.). *Input* and *output* restrictions define the role and interaction of units in network: some units may receive no input from the environment or send no output outside the net, and some of the interconnections among units in the network may be deleted. Furthermore, there may be restrictions on the *values* (positive or negative) of some of the connections and the *strength* of certain group of units. Learning in a connectionist network depends on network's architecture and the representation of activation patterns. It involves modification, i.e. *weighing* and *processing* of weights on connections in a network, in a way that influences the pattern of unit activations produced in response to a given input. In *supervised* learning, changing of the weights to achieve desired result can be influenced externally by explicit feedback based on the behavior of the network, as is the case in error-correction learning. In the *unsupervised* learning, only input provided to the network along with internal biases is being used.

The most general connectionist network is a three layer feed-forward network (Figure 2) where all units are interconnected and process from input to output through hidden layers, without any restrictions on connectivity or external influences (for thorough analysis of different variations, see Hinton and Anderson 1981, McClelland and Rumelhart 1981, McClelland and Rumelhart 1985, Rumelhart *et al.* 1986).

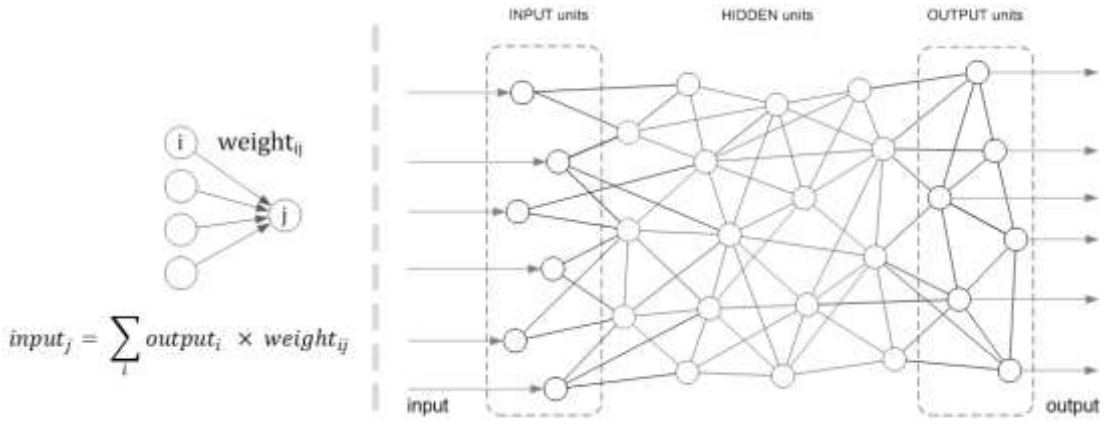


Figure 2: A diagram of general feed-forward connectionist architecture (adapted after (Plaut 2003, p. 148)). All units in the network are interconnected and process forward from input through a hidden layer to output. The activity of each unit is a non-linear function of the summed weighted input from other units.

Some neural networks use *localist* representations where individual conceptual object (whether letter, word or particular visual feature) is represented by a single unit (e.g., Dell 1986, McClelland and Rumelhart 1981). Other, more complex networks operate upon *distributed* representations where individual conceptual object is distributed over a pattern of activations from a number of simple processing units (e.g. Hinton, McClelland and Rumelhart 1986). The Interactive Activation model of word perception (McClelland and Rumelhart 1981) is an example of *localist* network: each unit stands for an individual conceptual object and there are three layers of units: letter-feature units, letter units, and word units (Figures 3 and 4). However, some further constraints apply: "units within the same rectangle stand for incompatible alternative hypotheses about an input pattern and are mutually inhibitory. Bidirectional excitatory connections between levels are indicated for one word and its constituents" (McClelland 1988, p. 109).

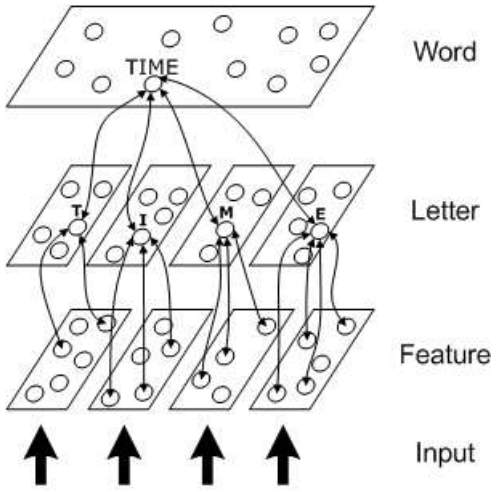


Figure 3: A three layered Interactive Activation model of word perception (McClelland and Rumelhart 1981). Only some of the units are activated and each letter in the middle layer is a generalization of particular pattern in input layer).

In the Interactive Activation model the functionality of the network is based on interactive processing of weighted activations between layers: units in each layer

receive excitatory connections from consistent units at other layers and inhibitory connections from inconsistent alternatives within the same layer. Such simple architecture can explain some cognitive aspects of language use, for example a number of context effects in perception, including the word superiority effect where the perception of a letter is enhanced when it occurs in the context of a word compared with when it occurs in isolation or in a random letter string (for review, see Plaut 2003).

Take as an example a simulation of a word superiority effect in Figure 4. Here, the reader is processing the letter T in the beginning of a word. All the letters in Figure 4 apply only to the first letter of the word. The bottom layer represents visual feature detectors, where similar features, i.e. those that match the features of an uppercase T, are active (see two nodes on the left), whereas the three nodes on the right are not active because they don't match. Nodes in the lower visual feature detector level are connected nodes in the letter detector level. All connections in the network are either excitatory (represented with an arrow at the end of the connection) or inhibitory (represented with a circle at the end of the connection). The activation spreads throughout the network. For example, the note representing the letter T is sending excitatory activation to all the words that start with T and inhibitory activation to all the other words. As word nodes gain in activation, they will send inhibitory activation to all other words, excitatory activation back to letter nodes from letters in the word, and inhibitory activation to all other letter nodes. Letters in positions other than the first are needed in order to figure out which of the words that start with T is being read.

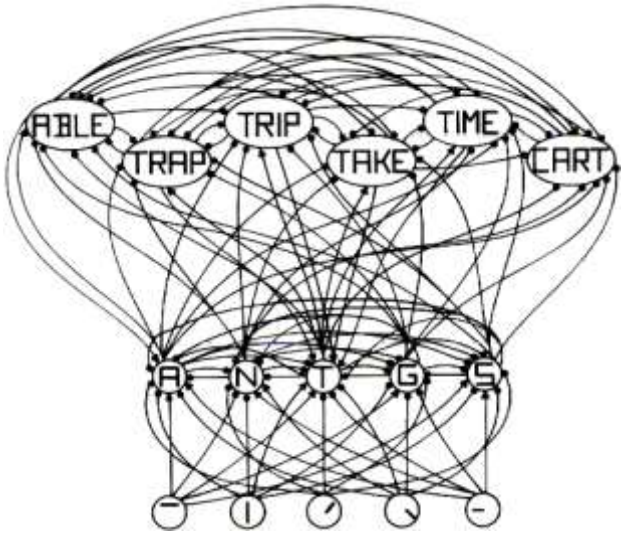


Figure 4: Interactive Activation model: a simulation of word superiority effect (McClelland and Rumelhart 1981)

The advantage of the connectionist approach is in the ability to learn and model the dynamics involved in cognitive behavior. Learning occurs through the evolution of patterns of activation over time. In the three-layer networks, for example, the propagation of activation among the units is directed via weighted connections in the hidden layer. The hidden layer is playing a representational role because the processing of the units in the hidden layer responds to the input of the network (via weighted connections), which results in partitioning of the activation space of the hidden units. Weights represent the enduring 'knowledge' of the network, whether it uses supervised or unsupervised learning, and determine how the network will react to incoming stimuli. This is an essential characteristic of connectionist representation.

5.3 Connectionist models of language

Even though examples shown above are cases of unsupervised learning, a large part of psychological connectionist modeling uses *supervised* learning models. According to Plaut (2003), unsupervised learning may be successful at modeling simple cognitive tasks (for example simple perceptual tasks) where "the similarities among representations provided by the environment may be sufficient to guide the behavior" (p. 145). Supervised learning, on the other hand, is more effective at modeling complex transformations involved in many forms of cognitive processing, and thus

contributes to the understanding of learning, generalization, and the flexibility and productivity of cognition. Since language comprehension exhibits many of these features, it quickly became the main area of connectionist research.

A typical example of connectionist supervised learning procedure is a *back-propagation*[20] (Rumelhart, Hinton and Williams 1986). Back-propagation is a type of error-correction algorithm with the aim to "[c]hange each weight in a way that reduces the discrepancy between the correct response for a given input and the one actually generated by the system" (Plaut 2003, p. 145). It does this by manipulating internal representations over hidden units, based on calculated changes in each unit's activation and by modifying the unit's incoming weights accordingly.

An early connectionist model from Rumelhart and McClelland (1986) used back-propagation to generate the past tense forms of both regular and irregular English verbs from their stems. Their model learned various categories of verbs and a direct association between the phonology of all types of verb stems and their past-tense forms using a single neural network, thereby obviating the need for rule-based syntactically structured representations (typically used in symbolic approach). This is no trivial task, since different categories have different type and token frequencies and the network has to infer phonemic forms from both regular forms (i.e. by adding to the verb stem in one of the three regular ways, either /ed/ (add - added), /d/ (play - played) or /t/ (walk walked)) and irregular past tense forms (arbitrary mappings (is - was, go - went), identity mappings (hit - hit), or vowel change mappings (run - ran, ring - rang)).

Many aspects of Rumelhart and McClelland's approach were strongly criticized, most notably by (Pinker and Prince 1988; Lachter and Bever 1988), and in more general terms by Fodor and Pylyshyn (1988). One general criticism of the original model, put forward by Pinker and Prince (1988), is the lack of any explicit linguistic rules. The consequence, they claim, is that the model does not capture the fine details of the data and learns certain rules that are not characteristic of any human

[20] The back-propagation algorithm iteratively: 1) computes activations forward from input units to output units, usually via one or more hidden layers; 2) computes a measure of performance error over the output units, 3) propagates this error backward through the network to determine the partial derivative of the error with respect to each weight in the network; and finally 4) changes the weights based on these derivatives so as to reduce the error (Rumelhart, Hinton and Williams 1986).

language[21]. Another criticism concerned the training corpus, arguing that it was artificially structured with unnatural proportion of regular to irregular verbs.

In subsequent work (Plunkett and Marchman 1991, 1993; Cottrell and Plunkett 1995), many of the specific limitations of the model have been addressed. Modifying the original approach of Rumelhart and McClelland (1986), Plunkett and Marchman (1991) trained the network by activating the phonemic representation of a verb on the input layer and generating an output by successive application of the back-propagation algorithm, constantly changing the weights in the network to generate a smaller error when the same input pattern is presented in the future. Apart from learning to generate the proper past tense forms of English words, the back-propagation based learning demonstrated some important similarities to the way children learn the English past tense forms. In comparison with the original approach, Plunkett and Marchman trained the network on a large number of irregular words first, gradually adding larger set of regular and irregular words to the training. At this point, the model showed some overgeneralization. Like children, the network first learned the correct form for the irregulars in its corpus, but subsequently overgeneralized the regular form and applied it to some irregular forms (thus producing, using an English example, 'dreamed' instead of 'dreamt') before finally learning the proper form of all verbs (Markman 1999; Plunkett and Marchman 1991). Thus, Plunkett and Marchman argued their model can explain correct uses of both regular forms and overgeneralizations of irregulars, and does this in a psychologically credible way.

Despite the success of Plunkett and Marchman's model, it is important to note that simulating learning English past tense forms is only one small part among a wide variety of tasks related to modeling language. Moreover, especially in modeling higher level cognitive processes, there are obvious deficiencies of connectionist approach. A main problem is the lack of hierarchy and systematicity of connectionist architecture. In more specific terms, Fodor and Pylyshyn (1988) argued that connectionist approach cannot account for the compositional nature of language, which is essential to any credible theory on language and cognition in general. Thus, if Fodor and Pylyshyn's criticism holds, connectionism has problems with both,

[21] This opened a heated debate (cf. Markman 1999; Haselager and Rappard 1998; Haselager, Bongers and Van Rooij 2003; van Gelder 1990; Bechtel and Abrahamsen, 1991)

theoretical and functional validity. These issues will be discussed later in Section 5. First, let us shortly review some of the attempts to overcome structural and functional deficiencies of individual approach by building a hybrid system.

Section 4: Hybrid systems

6 Hybrid systems

In general, both camps acknowledged some of the shortcomings of modeling language and cognition and sought solution in hybrid alternatives.

> While connectionist models have had considerable success in many areas of cognition, their full promise for addressing higher level aspects of cognition, such as reasoning and problem solving, remains to be fully realized (McClelland 1999, p. 139).

> ... the motivation for the design and construction of hybrid models, both within cognitive science and more practical applications, is that those models, by inheriting the virtues of the component technologies, can thereby also avoid the often cited vices. Thus, a hybrid model might avoid the symbolic vice of "brittleness" by employing a micro-feature based distributed representation which engenders graceful degradation. Similarly, the semantic opacity of a distributed connectionist system might be ameliorated by employing a symbolic component. (Cooper and Franks 1994, p. 5)

The idea behind most hybrid systems is that connectionist systems can provide the underlying architecture for high level symbolic processing. A large number of different proposals has emerged over the years (e.g., Rumelhart and McClelland 1986, Plunkett and Marchman 1991, Hinton 1981, 1988, Touretzky 1986, Pollack 1990), which, according to Franks and Cooper (1995), could be defined within three general categories of hybrid models: implementational connectionism, implementational computationalism and real hybridness.

6.1 Top-down: implementational connectionism

First, consider a physical combination of both architectures, or what Pinker and Prince (1988) called "implementational connectionism". By this view, neural networks could complement symbolic rule processing by creating distributed representations of elementary information upon which the functions of symbolic model could operate. In such cases, the network would present an intermediate level tightly coupled to the hardware of the system, and compute input to output according to the rules given by the symbol system (for some of the proposals see, e.g. Hinton 1981, 1988, Hinton, McClelland and Rumelhart 1986, Touretzky 1986, Hinton and

Touretzky 1985). From the technical, engineering perspective of building such a system, the focus is on "the way in which entities in the domain are mapped into the hardware changes during the course of the inference" (Hinton 1988, p. 50). In most cases, such systems are primarily symbol systems that use neural functions where needed (in processing perceptual tasks, for example). Thus, the contribution is not equal: the mediation between both architectures is typically directed from top-down symbolic operations. From theoretical perspective, even though both (physically distinct) types of architectures are involved, there is still no cognitively plausible explanation on how representations on different levels should interact.

6.2 Bottom-up: implementational computationalism

An alternative approach is a fully developed connectionist network that can simulate hierarchy, i.e. compositional structure and syntactic transformations of symbolic approach. This is a prevalent approach in constructing hybrid systems and deserves some more attention[22]. Some of the notable examples include connectionist implementation of syntactic transformations (Touretzky 1986, Pollack 1990), a connectionist implementation of production systems (Touretzky and Hinton 1988), or implementation of structure and part-whole hierarchies (Hinton 1989, Elman 1989, Smolensky 1990). For example, Hinton's arguments for hardware modularity emphasize the need for putting compositional structure into the system:

> One of the best and commonest ways of fighting complexity is to introduce a modular, hierarchical structure in which different modules are only loosely coupled ... Self-supervised back-propagation ... was originally designed to allow efficient bottom-up learning in domains where there is hierarchical modular structure ... Given a sufficiently large ensemble of input vectors and an "innate knowledge" of the architecture of the generator, it should be possible to recover the underlying structure by using self-supervised back-propagation to learn compact codes for the low-level variables of each leaf module . It is possible to learn codes for all the lowest-level modules in parallel. Once this has been done, the network can learn codes at the next level up the hierarchy. The time taken to learn the whole hierarchical structure (given parallel hardware) is just proportional to the depth of the tree ... it is helpful to allow top-down influences from more abstract representations to less abstract ones, and a working simulation. (Hinton 1989, p. 228)

[22] Proponents of the classical approach argue that much of the criticism directed towards connectionist approaches to language also holds for this kind of hybrid systems. See Chapter 5 for a discussion.

6.2.1 Recursive Auto-Associative Memory model (RAAM)

In general, hierarchical structure is the key problem for connectionism. One example of implementing compositional structure into connectionist network is a Recursive Auto-Associative Memory (RAAM) developed by Pollack (1990). RAAM is a three-layer feed-forward network which can transform syntactically structured sentences (or any information in symbolic tree structures) into distributed representations. It does this by employing back-propagation learning algorithm on input representations and giving back distributed representations as an output.

Moreover, RAAM can translate these representations recursively (in both directions) between symbolic trees and numeric vectors. It generates the translation using two basic components of the architecture: the *compressor* and the *reconstructor*. The role of the *compressor* is to encode symbolic tree structures in a bottom-up fashion, from leaves up to the root. Using Pollack's example (1990, p.84), the hypothetical tree structure of ((A, B) (C, D)) can be encoded in three steps: first A and B are compressed into a pattern R1, following by C and D being compressed into pattern R2, and finally, both patterns (R1 and R2) compressed into pattern R3 (whole), thus resulting in a hierarchically distributed representation of symbolic tree structure (see Figure 5).

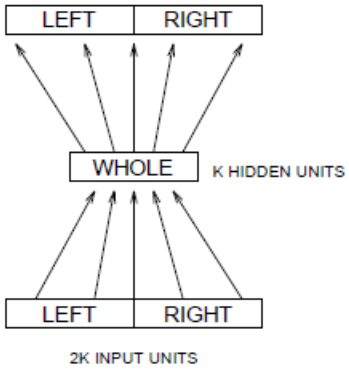


Figure 5: A single feed-forward network composed of both *compressor* (input units) and *reconstructor* (output units).

The *compressor* acts as the encoder or compositionality generator for RAAM's distributed representations. The role of the *reconstructor* is just the opposite: it

reconstructs the constituent structure from the distributed representation. In a top-down fashion, the reconstructor recursively decodes the distributed representation into an original symbolic tree structure: first, by decoding R3 into R1 and R2, and consequently, from R1 into A and B and from R2 into C and D. The implications of RAAM are significant for the theoretical and functional aspects of modeling language and cognition, and for the general connectionist vs. symbolic discussion (see Section 5).

RAAM uses a purely connectionist approach. Delineation between symbolic (read abstract) and connectionist (read physical) properties of the system is lost, since the functions of both architectures conflate onto a modular neural network. Consequently, the 'knowledge' is stored implicitly in hierarchical layers of weighed distributes representations, based on the back-propagation of some low-level variables of the input information. Such network is able to reveal graded patterns of generalization from lower to higher variables, but the higher, abstract level of analysis (typical for symbol systems), offering an interpretation of this 'knowledge', is missing. Unlike typical symbolic representations, which are explicit and directly amenable to rule-like operations, the connectionist, and in this case simulated symbolic representations are implicit and not directly affordable to the system – they need compressor and reconstructor in order to translate. At the end, the underlying psychological theory is purely connectionist and succumbs to the typical problems of connectionist networks. True, compared to basic networks some of the compositionality (of symbolic model) is preserved, but (as we will argue later in the thesis) psychologically valid theory of how these structures could be semantically interpreted is missing.

6.3 Real 'hibridity'?

The third option is a 'real' hybrid system where both levels complement each other in a bottom-up and top-down fashion[23]. Here, "the system's behavior is generated by both connectionist and symbolic functions and by theoretically significant causal relations between them" (Boden 2006, p. 976). And for Franks and Cooper (1995), the real hybrid system "... that may be cognitively plausible and explanatorily acceptable is of the behaviourally hybrid type" (p. 61).

[23] This idea reflects philosophical discussion in Minsky's The Society of Mind (1985).

On higher level, certain symbolic functions (e.g., based on semantic and/or syntactic properties) could influence the learning process in the neural network (e.g., with the aim of learning regular and irregular forms of English past tense) and in return, the 'knowledge' of the system would be ameliorated by the information given through the output of network's error-correction learning algorithm. For example, a back-propagation model of McClelland and Rumelhart (1986) showed a genuine empirical validity in simulating learning English past tense forms without predefined linguistic rules, but such network would probably be even more stable (and psychologically valid), when complemented by compositional and rule based structure of symbol system (recall Pinker and Prince's (1988) argument).

Ideally, such hybrid system should overcome many of the problems that individual approach faces when modeling cognitive behavior individually. In reality, this is not the case. Among proposed solutions there is currently no common unified theory nor is there a hybrid model that could successfully explain basic properties of language, let alone other areas of cognition. Arguably, one reason might be the fact that the theoretical positions of both camps have, for the most part, remained static. Apart from the necessary modifications of their models to account for the novel scientific discoveries, the underlying theories themselves (whether connectionist, symbolic or hybrid) have not changed. Thus, for the most part of previous century we have been witnessing gradual improvement of the architectures, but this development has hardly been reflected in the core theoretical positions.

6.4 Discussion: A need for intermediate level

Many of the shortcomings of individual models from 1990's were due to the constraints of technological development. But the main reason lays in the variety and complexity of human cognition and therefore inability to develop general, psychologically plausible theories (or, in Newell's terms, "unified theories of cognition") and computational models. Moreover, the ongoing research (especially in neurosciences; e.g., Cabeza and Nyberg 2000, Martin *et al.* 1996, Martin and Simmons 2008) has started to shift some of traditional cognitive explanations. Taken together, with the advance of modern technologies, more detailed scientific introspections into human cognition are being possible now, then ever before.

6.4.1 Problems with hybrid account

The general problem with existing hybrid models is the theoretical and functional incompatibility of competing architectures: direct coupling between levels is difficult because the direct translation from connectionist onto symbolic representations, and vice versa, is not possible, or approximate at most. On one side we have distributed representations revealing graded patterns of generalization from the environment, on the other, hierarchical symbolic structures governed by rules and compositional semantics. From both functional and theoretical perspective, there is a large abyss to cross.

Some of these issues echo from the studies of (Cooper and Franks 1993; Franks and Cooper 1995). They identify a number of dimensions where connectionist and symbolic implementations in hybrid models conflict. Since the exact mappings between symbolic and connectionist representations are not possible, they can only approximate the functioning and representational structure of the other model. The issue becomes what is 'lost' when mapping the two models. Or, as Franks and Cooper (1995, p. 67) put it, "how much mismatch between components is permissible?" That is, to what extent does connectionist model, for example, approximate the symbolic structure. One way of addressing this question is to try and differentiate operational and instructional aspects of respective architectures from physical or behavioral aspects of cognitive content. According to Haugeland (1991, 2000), the operational aspects, such as skeletonisation (i.e. a process of getting representational essentials from the context-dependent information in the environment) and implementation, cannot give exact mappings. For hybrid systems to be psychologically plausible, the physical and behavioral aspects should be preserved as much as possible. Moreover, the evaluation of hybridity should follow commitments of particular cognitive theory, not only functional (operational and implementational) aspects of the system in question[24].

6.4.2 Incompatabile representational mechanisms

A related issue concerns the incommensurability of symbolic and connectionist representations. We have already touched upon the notion of connectionist vs.

[24] Fodor and Pylyshyn (1988) use similar argument to claim connectionism can only explain implementational aspects of modeling cognition, not cognitive.

symbolic representation: the first being continuous, distributed and implicit (i.e. hidden in the system), and the latter being explicit, discrete and affordable to the system. The difference between implicit vs. explicit means the connectionist system can only employ representations to the extent they are an implicit part of the processing, but cannot systematically operate upon/over them, this latter characteristic is natural for symbolic models. Therefore, the connectionist architecture and algorithms are fundamentally different from the symbolic.

> There is no clear method for ensuring the preservation of a symbolic algorithm in a dynamical equation that does not itself constitute an algorithm; nor one for ensuring the preservation of a symbolic semantics (or function) in a connectionist account of the discipline of real world entities ... There is no possibility of an exact mapping between the connectionist and the symbolic. (Franks and Cooper 1995, p. 69)

The other significant (and tightly related) problem is that both approaches operate on different and to a large extent disjunctive representational levels. The issue becomes, how do we "reconcile the static ontologies of standard knowledge representation with a continuously changing world described using the ontologies of physics" (Hallam 1995, p. v)? How do we translate (map) from the dynamic distributed architecture into the static compositional one, and vice versa? How do we interpret, for example, patterns of continuous interactions into discrete symbolic states?

Thus, the crux of the matter is lack of intermediate semantic theory that could translate one system's representational architecture into another. For example, the traditional realist view of the world is strongly reflected in the symbolic approach. Traditional realist view argues for combinatorial syntax and semantics, and the use of discrete states, formal ontologies and context-independent semantic relations, to describe and represent things in the world. As such, it fails to explain some of the essential cognitive phenomena, such as learning on partially accessible information, social and cultural context, graded structure of concepts and categories etc., and therefore lacks psychologically validity (Chapters 10 and 11 present these issues in more detail).

Connectionism, on the other hand, focuses on the low-level dynamic and 'fine-grained' aspects of cognitive behavior, with distributed representations generally too unstable to afford any durable form of semantics. And, as noted earlier, connectionist

representations are not affordable to the system. Moreover, as Pinker and Prince (1988) and others (especially Fodor 1975, Pylyshyn 1985, Fodor and Pylyshyn 1988) have argued, some of the detail, i.e. some of the properties and relationships among conceptual objects, is lost or hidden in the representation of the network.

The rest of the thesis presents a quest for a semantic theory that could alleviate some of traditional issues about modeling representations, especially in regard to meaning of natural language. We begin with the analysis of three basic properties, namely *systematicity*, *productivity* and *compositionality* of language, and solutions offered by symbolic and connectionist approach. The following chapters act as an introduction to the discussion of semantics in Part III.

Section 5: Clash of two paradigms

7 Systematicity, productivity and compositionality of language and thought

7.1 A classicist's critique of connectionism

Number of criticisms have been directed towards connectionist approach to cognition, most notably from Fodor and Pylyshyn (1988), Pinker and Prince (1988) and Fodor and McLaughlin (1990). While systems like RAAM showed the capability of connectionist architectures to employ structural representations, for Fodor and Pylyshyn the structural problems of connectionist models still persist, even when in hybrid configuration. For one, defenders of classical symbolic approach have criticized connectionism for its inability to represent *concatenative* compositionality: a *concatenative constituent* structure of language and thought. For Fodor (1975) and Fodor and Pylyshyn (1988), representational system must be compositional, meaning that representations have a combinatorial syntax and semantics as a result of their *concatenative* compositionality. In practice, this results in the part/whole relationship between simple and complex representations. The concatenative constituent structure of complex representations is built up from the simple elements, by some mode of combination, to match the structure of the information represented; for example, the representation of "John loves Mary" is a concatenation of tokens "John", "loves" and "Mary". The mode of combination defines the relations between elementary representations that "explicitly represent the relations between parts of the information (e.g. that in "John loves Mary" it is Mary being loved, not John)" (Haselager 1998, p. 168).

7.1.1 Concatenative compositionality

For classical approach, the concatenation of tokens is necessary since it preserves the internal syntactical structure of symbolic representations. The symbolic tokens must be explicit and preserved in the representations, as constituents of the composite expression, for possible extraction and decomposition. Syntactic structure plays a crucial role in the classical approach to modeling cognition: it purports to explain

how a physical system (such as computer or brain) can exhibit semantically coherent behavior. Symbolic representations need to mediate between semantic and physical states of the system and do this in a systematically determined fashion via syntactic structure. In this way, the "semantic properties of representations come to be reflected in their syntactic (and hence physical) properties. ... The result is a purely physical system that can respect semantic criteria in its behavior by virtue of the causal role in the system of the syntactic structure of its representations; it is a purely "syntactic engine" generating meaningful activity" (ibid., p. 366). That this "syntactic engine" employs explicit physical tokens of the original constituents and thus satisfies the "semantical criteria of coherence" of thought, is essential for classical approach to cognition.

> If, in principle, syntactic relations can be made to parallel semantic relations, and if, in principle, you can have a mechanism whose operations on formulas are sensitive to their syntax, then it may be possible to construct a syntactically driven machine whose state transitions satisfy semantical criteria of coherence. Such a machine would be just what's required for a mechanical model of the semantical coherence of thought: Correspondingly, the idea that the brain is such a machine is the foundational hypothesis of Classical cognitive science. (Fodor and Pylyshyn 1988, p.30)

The "syntactic engine" plays a double role. It enforces structure sensitivity of representations and at the same time ensures that similarly structured representations also bear semantic similarities:

> ... on one hand, the meaning of the representation as a whole is fixed by the meanings assigned to the constituent tokens, and on the other, the presence of these tokens simply is the formal syntactic structure which drives the causal processes in the system (van Gelder 1990, p. 367).

For the causal processes in the system to occur, the syntactic structure must be enforced not only at the functional level, but on the implementational level as well.

> ... the symbol structures in a Classical model are assumed to correspond to real physical structures in the brain and the combinatorial structure of a representation is supposed to have a counterpart in structural relations among physical properties of the brain. For example, the relation "part of", which holds between a relatively simple symbol and a more complex one, is assumed to correspond to some physical relation among brain states ... This bears emphasis because the classical theory is committed not only to there being a system of physically instantiated symbols, but also to the claim that the physical properties onto which the structure of the symbols is mapped are

> the very properties that cause the system to behave as it does." (Fodor and Pylyshyn 1988, pp. 13-14)

Fodor and Pylyshyn (1988) conclude their argument by pointing out the inherent structure sensitivity of syntactic-semantic relation in classical approach:

> It is perhaps obvious by now that all the arguments we've been reviewing-the argument from systematicity, the argument from compositionality, and the argument from inferential coherence-are really much the same: If you hold the kind of theory that acknowledges structured representations, it must perforce acknowledge representations with similar or identical structure ... So, if your theory also acknowledges mental processes that are structure sensitive, then it will predict that similarly structured representations will generally play similar roles in thought. (p. 48)

The internal syntactic structure of symbol system, it is argued, ensures the representation of structural similarities which can be capitalized by mental processes: similarly structured representations can be treated in similar ways (recall 'John loves Mary' example). This shows a deep theoretical commitment of the classical approach to modeling cognition via internal syntactic structures, not only in terms of functional and implementational properties of some physical system, but of language and thought as well. Hence, mental representations of the Language of Thought are strictly concatenative and employ

> a combinatorial syntax and semantics, in which (a) there is a distinction between structurally atomic and structurally molecular representations; (b) structurally molecular representations have syntactical constituents that are themselves either structurally molecular or structurally atomic; and (c) the semantic content of a (molecular) representation is a function of the semantic contents of its syntactic parts, together with its constituent structure. (ibid., p. 12)

7.1.2 Productivity and systematicity

Based on the classical definition of compositionality[25], Fodor and Pylyshyn (1988) argue for two essential characteristics that distinguish classical cognitive theories from connectionist: *productivity* and *systematicity* of language and thought.

[25] A more loose interpretation of compositionality as *functional compositionality* is proposed by (Chalmers 1990a, 1993 and van Gelder 1990). Differences between *concatenative* and *functional compositionality* will be discussed later.

Productivity refers to the ability to produce an infinite number of novel propositions, sentences or thoughts, with finite means. This is only possible when representational system has a combinatorial syntax and semantics (Fodor 1975; Fodor and Pylyshyn 1988). The theoretical arguments for productivity are closely connected to Chomsky's generative grammar (Chomsky 1965, 1968): the knowledge underlying linguistic competence is generative, as it allows us to generate and understand an unbounded number of sentences. Similarly, Fodor and Pylyshyn (1988, p. 21) note "there are indefinitely many propositions which the system can encode. However, this unbounded expressive power must presumably be achieved by finite means". For Chomsky (1965, 1968), as for Fodor (1975), many aspects of the formal structure of language are innate, a part of human biology. In classical computational interpretation then, the language faculty is an internal computational-representational system.

> It seems clear that we must regard linguistic competence – knowledge of a language – as an abstract system underlying behaviour, a system constituted by rules that interact to determine the form and intrinsic meaning of a potentially infinite number of sentences. Such a system – a generative grammar – ... defines a language ... as a recursively generated system, where the laws of generation are fixed and invariant, but the scope and the specific manner in which they are applied remain entirely unspecified. (Chomsky 1968, p. 62)

Systematicity, in terms of linguistic processing, refers to the intrinsic connection between our ability to understand and produce certain linguistic forms and expand on others. In terms of encoding, any language that can encode certain sentence will automatically be able to encode a variety of related sentences. Thus, understanding or producing certain sentences is intrinsically related to the ability to understanding and producing certain others (Fodor and Pylyshyn 1988). For example, we cannot understand the sentence "John loves the girl" without also being able to understand "the girl loves John", or any other utterance similar to the form of "X loves Y". This follows from compositionality principle:

> ... a straightforward (and quite traditional) argument from the systematicity of language capacity to the conclusion that sentences must have syntactic and semantic structure: If you assume that sentences are constructed out of words and phrases, and that many different sequences of words can be phrases of the same type, the very fact that one formula is a sentence of the language will often imply that other formulas must be too: in effect, systematicity follows

> from the postulation of constituent structure. (Fodor and Pylyshyn 1988, p. 25; italics added)

In the same manner, thought also shows systematicity: any cognitive system that can think one of these sentences will also be able to think the other. By classical view, a form of structural relations (e.g. subject – predicate – object in the "John loves the girl" and "The girl loves John" example) is the same in both thoughts, only certain atoms have changed place; thus understanding the first thought by way of systematicity of structural relations implies understanding the second thought as well. Moreover, these kinds of thoughts are systematically related not only on a structural level, but also "from a semantic point of view" (pp. 31 -42). Herein lays the difference between the "systematicity of cognitive representation" and the "compositionality of representations".

Productivity and *systematicity* are essential properties supporting classicist's position on language and cognition: the need for underlying abstract structures that can be freely composed, instantiated with novel items, and interpreted on the basis of their structure (Lewis 1999). Moreover, both express compositionality and follow combinatorial syntax and semantics, and structure sensitivity of process.

> ... the main argument stands: systematicity depends on compositionality, so to the extent that a natural language is systematic it must be compositional too. ... The standard argument for compositionality is that it is required to explain how finitely representable language can contain infinitely many nonsynonymous expressions. (Fodor and Pylyshyn 1988, p. 29)

7.1.3 Fodor and Pylyshyn's further arguments

Fodor and Pylyshyn (1988) point out they have no problem accepting connectionist (or any kind of hybrid system that is essentially connectionist) implementation of symbolic features, such as compositionality for example. What they argue for is that the sole implementation or simulation of these characteristically symbolic features does not make the simulating system systematic and compositional per se. Hence, merely providing counterexamples of simulated compositionality (e.g., in connectionist architectures such as RAAM) is not sufficient: compositionality has to be inherent in the system. Fodor and Pylyshyn argue that connectionist systems have no logical syntax and consequently no "mechanism to enforce the requirement that

logically homogeneous inferences should be executed by correspondingly homogeneous computational processes", no combinatorial structure, and "nothing to prevent minds that are arbitrarily unsystematic" (ibid., p. 33). Thus, for connectionism to really count as an independent cognitive theory, it would have to show that systematicity necessarily follows from (and is essential to) the connectionist architecture. According to Fodor and Pylyshyn and proponents of classical view, it does not. Although connectionist architecture can employ structured representations and show some level of compositionality, those structures neither support semantic evaluation nor exhibit the properties of classical symbolic constituents (for one, they lack logical syntax). Consequently, connectionism cannot express psychological generalizations that classical theories capture (McCauley 1998). Moreover, even if neural networks can address some of the cognitive states and processes, they do so at an analytic level that is subconceptual and therefore, non-cognitive. In other words:

> It's not enough just to stipulate systematicity; one is also required to specify a mechanism that is able to enforce the stipulation. To put it another way, it's not enough for a Connectionist to agree that all minds are systematic; he must also explain how nature contrives to produce only systematic minds. Presumably there would have to be some sort of mechanism, over and above the ones that Connectionism per se posits, the functioning of which insures the systematicity of biologically instantiated networks; a mechanism such that, in virtue of its operation, every network that has an aRb node also has a bRa node… and so forth.
>
> There are, however, no proposals for such a mechanism. Or, rather, there is just one: The only mechanism that is known to be able to produce pervasive systematicity is Classical architecture. And, as we have seen, Classical architecture is not compatible with Connectionism since it requires internally structured representations. (Fodor and Pylyshyn 1988, p. 35)

To reiterate, according to classical interpretation, the language faculty is an internal computational-representational system. Such analogy assumes that symbolic architecture (embracing the notion of logical syntax and consequently compositionality and systematicity) is prerogative for modeling abstract cognitive processes, whereas physical explanations, characteristic of connectionist modeling, contribute nothing to cognitive explanations; their role is purely implementational. Hence, "a theory of the relations among representational states is ipso facto a theory

at the level of cognition, not at the level of implementation" (Fodor and Pylyshyn 1988, p. 48).

For Fodor and Pylyshyn then, there is a clear distinction between the architecture (which, to become a credible cognitive theory, should accept the symbolic approach) and its implementation (which can be realized in various ways, symbolic, connectionist or hybrid).

> ... the implementation, and all properties associated with the particular realization of the algorithm that the theorist happens to use in a particular case, is irrelevant to the psychological theory; only the algorithm and the representations on which it operates are intended as a psychological hypothesis. ...
>
> Given this principled distinction between a model and its implementation, a theorist who is impressed by the virtues of Connectionism has the option of proposing PDP's as theories of implementation. But then, far from providing a revolutionary new basis for cognitive science, these models are in principle neutral about the nature of cognitive processes. In fact, they might be viewed as advancing the goals of Classical information processing psychology by attempting to explain how the brain (or perhaps some idealized brain-like network) might realize the types of processes that conventional cognitive science has hypothesized. (Fodor and Pylyshyn 1988, p. 47)

From classicist's perspective, human language and thought are compositional and governed by rules as opposed to simple activation patterns and associations (Pinker and Prince 1988). Fodor and Pylyshyn further argue that these characteristics are necessary for any adequate theory of cognition. If connectionism cannot properly explain compositionality, then it is not an adequate cognitive theory, but, in best case, a mere implementation of symbolic approach.

7.2 Connectionist's reply: functional compositionality

In *On the Proper Treatment of Connectionism* (1988), Smolensky attempted to rebut Fodor and Pylyshyn's criticism of connectionism as exclusively implementational, and argued for functional and theoretical implications that connectionism brings to modeling cognition, noting how connectionist systems readily accommodate the context sensitivity of representations for which considerable psychological evidence exists. Hence, for Smolensky (1988, 1989) the concatenative compositionality is

rigid, as the content in such architecture is generally context-invariant. As further noted by McCauley (1998), there are many cases where "the representational stability depends not on symbolic form but on a "family resemblance" among those vectors that, in different contexts, carry out some functional, subsymbolic role" (p. 621). For example, the processing of propositional attitudes (characteristically a symbolic operation) is typically supported and initiated by the "intuitive processor" (Smolensky 1988, p.3), which does not involve symbolic manipulation. That is, the processing of propositional attitudes should incorporate representations of agent's behaviour and interaction with the environment when expressing certain belief ascriptions, not merely context-free symbolic structures. As we shall see, symbolic approach alone is unable to express such 'contextual' semantics.

Whether something counts as structural, functional or explanatory aspect of modeling of cognition, or a mere implementation of the architecture (as Fodor and Pylyshyn mark connectionist approach), strongly depends both on the level of analysis and theoretical commitment. For example, classical symbolic approach promotes high-level cognition that exhibits compositionality and is amenable to rules, hence ignoring other areas of cognition where conscious rule processing and predefined structural relations are not initially applicable (e.g., perception, intuition and, practically, all of skilled performance). Thus, what might look like implementational detail from a higher level symbolic perspective (for example, different learning algorithms giving different accounts of conceptually interpretable patterns), might have both functional and theoretical implications from lower level connectionist perspective (see Marr 1982). For example, Rumelhart and McClelland's (1986) back-propagation model showed that learning regular and irregular past tenses of English verbs might not necessary require predefined rules – a discovery, that put a shadow of a doubt on many previously widely accepted theories about language and cognition (e.g. Chomsky 1968, 1980, Pinker and Prince 1988, Fodor and Pylyshyn 1988).

Nevertheless, the classical notion of compositionality and the critique of connectionism posed by Fodor and Pylyshyn have to be taken seriously. What follows, is a comparison of symbolic and connectionist accounts of compositionality

and an attempt to refute Fodor and Pylyshyn's critique of connectionism as a valid cognitive theory.

7.2.1. Functional compositionality

Connectionists argued against concatenative nature of compositionality, most notably Smolensky (1987b, 1988), van Gelder (1990) and Chalmers (1990a, 1993). In symbol systems, the concatenation preserves the constituent structure and sequential relations among tokens in the expression, generating compound representation such as written natural languages, formal languages, mathematics, logic etc. Thus,

> [g]iven the way concatenation is defined, it is obvious that when describing a representation as having a concatenative structure, one is making more than just the grammatical point that it stands in certain abstract constituency relations, and also more than just the quasi-historical point that it happened to have been built up out of a certain set of (recoverable) constituents. One is also saying that it will have an internal formal structure of a certain kind; that is, such that the abstract constituency relations among expression types find direct, concrete instantiation in the physical structure of the corresponding tokens. An appropriate name for this kind of internal structure is syntactic structure. Thus, the syntactic structure of a representation is the kind of formal structure that results when a concatenative mode of combination is used. (van Gelder 1990, p. 360-361)

But as van Gelder points out, the concatenation is only one possible mode of compositionality. Since the notion of compositionality is inherent in representational architecture, there could be other, different kinds of compositionality:

> ...essential to a compositional scheme is the requirement that its expressions stand in certain abstract constituency relations. It is through the mode of combination, which relates primitive tokens to compound expression tokens, that these constituency relations are realized in a particular scheme; and it is because there can be important differences between modes of combination that various styles of compositionality can be distinguished. (van Gelder 1990, p. 359)

Connectionist representations do not employ syntactic structures in a classical sense and do not contain tokens of their constituents. The minimum requirement (by connectionist criteria) for the system to be compositional is:

> ... to have systematic methods for generating tokens of compound expressions, given their constituents, and for decomposing them back into

> those constituents again ... [there is no need for preserving the tokens of these] constituents in the expressions themselves; rather, all that is important is that the expressions exhibit a kind of functional compositionality. ...
>
> Functional compositionality is obtained when there are general, effective, and reliable processes for (a) producing an expression given its constituents, and (b) decomposing the expression back into those constituents. Such processes are *general* if they can be applied, in principle, in constructing and decomposing arbitrarily complex representations ... To be *effective* they must be mechanistically implementable; that is, it must be possible to build a machine that can carry out these processes. ... Finally, for these processes to be *reliable*, they must always generate the same answer for the same inputs. Standardly, of course, concatenative schemes are functionally compositional. (van Gelder 1990, p. 361)

The main difference between *concatenative* and a *functional* compositionality then is in the way tokens are being employed in representations. To some extent, connectionist representations can represent complex structures (as shown in RAAM), but these structures are not concatenative: "this kind of internal structure does not count as syntactic structure, since its parts do not, in general, satisfy the identity criteria for the various constituents" (van Gelder 1990, p. 363). According to connectionists, the implementation of such non-syntactic compositional structure into connectionist architecture is functional. And, as van Gelder (1990) rightly remarks,

> ... the most pertinent and informative contrast between the Classical approach and Connectionism is not, as Fodor and Pylyshyn (1988) have suggested, between a commitment to structured (Classical) as opposed to unstructured (Connectionist) representations; rather, it is between two very different ways of implementing compositional structure. (p. 365)

7.2.2 Local vs. distributed

Moreover, Chalmers (1990a/b, 1993a/b) and Smolensky (1987a, 1990) point out that Fodor and Pylyshyhn's critique of connectionist compositionality is ill founded. According to Chalmers (1990b, 1993a/b), the main flaw of Fodor and Pylyshyhn's critique lays in the general misunderstanding of connectionism and in consequent generalizations about *localist* vs. *distributed* representations. He argues Fodor and Pylyshyn (1988) built their case against connectionist compositionality upon the *localist* representations, which are taken as a typical example of connectionist architecture (see pp. 15-19). In reality, *localist* representations represent only a small

and atypical case of connectionist modeling. Moreover, the difference between the *localist* and *distributed* representations is significant and goes to the core of connectionist theory. To reiterate the debate from previous chapters, the deepest philosophical commitment of connectionist approach is the rejection of the atomic symbol as the bearer of meaning: "atomic tokens simply do not carry enough information with them to be useful in modeling human cognition" (Chalmers 1990b, p. 343). And to a large extent, sans basic associative links, the *localist* representation resembles the traditional notion of isolated atomic symbols: just like in symbolic models, each entity (word, concept, etc.) in the *localist* representation is being represented by a separate node. In typical connectionist architecture, on the other hand, the representation of each individual entity is *distributed* over many nodes, generating complex internal structures and "far more *information* than a single node" (ibid., p. 343). For Chalmers (1990),

> [t]his is the fundamental flaw in F&P's argument: lack of imagination in considering the possible ways in which distributed representations can carry semantics. It is a different variety of distributed semantics that would be carried by a connectionist implementation of a Turing Machine ... And it is a different variety again of distributed semantics that can yield connectionist models of compositionality in important new ways. (p. 343)

The main difference between *localist* and *distributed* representations then is in the "power of distribution" and hence in semantic interpretation of content: whereas *localist* semantics bears the characteristics of traditional symbolic approach, the *distributed* semantics is purely connectionist. Citing connectionist models of Elman (1990a/b), Pollack (1990) and Smolensky (1987a, 1990), Chalmers (1990) remarks:

> It is no accident that three of the most prominent counterexamples to F&P's argument – the models of Elman, Pollack, and Smolensky – all use distributed *representation* in an essential way. Smolensky's tensor-product architecture simply could not work in a localist framework. Its multidimensional tensor representations are by their nature spread over many nodes. Elman's implicit structure which develops in a recurrent network could also not succeed in a localist framework – the many subtle adjustments needed for various syntactic distinctions to develop could not be made. And Pollack's Recursive Auto-Associative Memory has a deep commitment to distribution – if it were one-concept-to-one-node, then its recursive encoding scheme could never get off the ground. (p. 344)

The main question then is whether functional compositionality is sufficient for modeling basic properties of language and thought[26]. If the answer is yes – and connectionists prove that structural similarity can be achieved without concatenative compositionality – then Fodor and Pylyshyn's arguments can be put to rest.

7.2.3 Examples of functional compositionality

Connectionists approached the above question empirically, by building models of systematic cognitive processing and aiming to achieve functional compositionality from non-concatenative representations exclusively (e.g., Pollack 1988, 1990, Smolensky 1987a, 1988, Elman 1990b). Apart from Pollack's RAAM (1988), two other models frequently mentioned in the literature are Hinton's (1988) model of representing hierarchical structures via reduced descriptions and Smolensky's tensor product framework (1987a)[27]. All three connectionist models offer alternative solutions to generating non-concatenative, functional compositionality. Moreover, all three models offer alternative solutions to representing various data structures over strictly limited connectionist resources.

7.2.3.1 Pollack's RAAM and Hinton's reduced description model

Presented in the chapter on hybrid modeling, RAAM (for a detailed analysis, see Pollack 1990 or Chalmers 1990) solves these issues by treating activation patterns in the network as stacks, by using functions of push and pop (for example, branches of binary tree in RAAM (Figure 5) are treated as stacks). Stacks are simple elements, operated upon by push and pop functions. Pushing (compressing) a new element into a stack generates a new pattern, thus expanding the stack, while popping (reconstructing) is a reverse process. Thus, structural relations are being kept or transformed by mode of operation and "[t]his process can be performed recursively, with the result that any given recursively structured sequence can be stored in an appropriately trained network" (van Gelder 1990, p. 369).

[26] This research is relevant for two reasons: to show the diversity of connectionist approaches to modeling language and cognition, and to show how distributed semantics, contra classicist's view, supports a general connectionist claim for functional compositionality.

[27] A throughout analysis and technical details of each approach are out of scope of this thesis and have been tackled in depth elsewhere in the scientific literature (e.g., in the writtings of Smolenksy (1987b, 1988), Elman (1989) and van Gelder (1990)).

An alternative solution to the problem of mapping a part-whole hierarchy into a finite amount of parallel hardware had been proposed by Hinton (1988). Since the neural networks in general are fixed, there are some general constraints: there is only so much information that can be represented at the certain point in time on particular level. Thus, using allegory of a "moveable window scheme", the hierarchical structure is being represented via reduced descriptions on each level. By utilizing structural similarities of the network, the network expands the constituent structure given/gained by reduced descriptions on particular upper level down onto the representations of the lower level.

As van Gelder points out, both RAAM and Hinton's reduced description model, allow for generation and

> ...recovery of all the constituents of that hierarchy; in that sense [both models] can be described as a compositionally structured representation[s]. But ... this is not achieved by having first concatenated those constituents. There is no requirement that constituents figure, in the representation of the whole, in anything like the form they appear in when the constituent has been fully expanded (van Gelder 1990, p. 371).

This argument is further supported by Hinton (1988):

> The crucial property of the moveable window scheme is that the pattern of activity that represents the current whole is totally different from the pattern of activity that represents the very same object when it is viewed as being a constituent of some other whole. (p. 52)

Both Hinton's and Pollack's approach have shown that connectionist representations can support systematic processing. In both cases, there is no need for explicit constituent structure, i.e. the preservation of constituents' tokens. There is no need for explicit tokening of the original expressions – functional compositionality, and hence systematicity, can be achieved by processing exclusively upon implicit structures of connectionist architecture (Pollack 1990; Hinton 1988; Smolensky 1988).

7.2.3.2 Smolensky's tensor product framework

In Smolensky (1987a, 1988), Smolensky describes a *tensor product framework*, another alternative to classical compositionality, and emphasizes the differences between connectionist and symbolic approach in more general terms, mirroring some

of the basic arguments presented above. As with Pollack's and Hinton's models, the emphasis is on the capability of generating structural representations without employing concatenation. Since the representations in neural networks are vectors describing patterns of activity over processing units and the connections between such units, the general problem is "finding a mapping from a set of structured objects (e.g., trees) to a vector space" (Smolensky 1987a, p. 2). Smolensky shows how hierarchical structure and various constituency relations among representations can be generated, preserved and recovered through the mode of combination (complex expression as combinations of simpler parts) operating on these vectorial representations, in this case by the processes of tensor addition and multiplication[28].

Smolensky's definition of *tensor product framework* embodies some essential properties of functional compositionality (and of distributed connectionist architecture in general), showing how functional compositionality offers a genuine alternative to concatenative compositionality of the Language of Thought. The vectorial representations of neural networks are not syntactically structured, i.e. "they do not contain tokens of the primary constituents (i.e., the primitive vectors assigned to the original roles and fillers) in any sense other than that there are processes that can generate those constituents given the compound representation" (van Gelder 1990, p. 373). Unlike atomic symbols, vectorial representations are context-dependent: there is no single canonical position (node) representing individual concept, rather, individual concepts are distributed over patterns, i.e. clusters of vectors, and related and influenced by a kind of similarity or "family resemblance" and weigh distributions in the network. Hence, the modes of composition and decomposition cannot be "precise and uniquely defined", as is the case with symbolic approach, and are subject to various kinds of imperfections (whether context effects, different kinds of ambiguities, interferences etc.; see (Smolensky 1988, p. 14-16)). Moreover, all notable connectionist alternatives to traditional concatenative compositionality (e.g. Pollack 1990, Smolensky 1990, Hinton 1988 and Elman 1990) employ distributed representations.

[28] In tensor product framework, computation is based on the numerical vectors and tensors (e.g., on activation values of vectors).

7.3 Discussion: a need for an unifying semantic theory

The above chapters echo some of the problems for both paradigms. Connectionism, if understood as purely implementational strategy (as Fodor and Pylyshyn would have it), loses much of explanatory power and functionality, especially when modeling higher-level cognition, such as language acquisition and comprehension. Proponents of classical approach rightly argue that productivity, systematicity and compositionality are essential ingredients of language and thought. They further argue concatenative compositionality cannot naturally emerge from connectionist architecture – it requires unlimited or arbitrarily extended resources which are simply not available to the connectionist architecture. Connectionists, on the other hand, argue that functional compositionality can replace concatenation by employing distributed representations of complex structures by mode of superposition. Moreover, Van Gelder (1990) points out that contrary to fundamental computational assumption, cognitive processing involves finite resources that are not arbitrarily extendable[29]:

> Distributing transformations, which take various constituents and superimpose them to achieve a new representation of the whole over the same space, inevitably destroys those tokens in the process (although not necessarily their recoverability), and hence are incompatible with any variety of concatenation. In short, the finite resource restrictions characteristic of Connectionism preclude concatenative styles of compositionality in favor of distributed (and so merely functional) styles. The fact that a broad alternative conception of compositionality is emerging in Connectionist research is thus a fairly direct consequence of one of its basic commitments, a commitment that stands in stark contrast with the Classical assumption that resources are always, at least in principle, arbitrarily extendable (as in, e.g ., the unbounded tape on a Turing Machine). (p. 375)

Thus, to stand as a genuine alternative to the symbolic approach to cognition, connectionism should be able to explain how arbitrarily complex structures could be generated by operating over finite representational resources without employing concatenation. This is a pressing problem for connectionism, since it cannot, apart from relatively successful modeling of some isolated cases (e.g., learning the proper form of English past-tense verbs, or exhibiting functional compositionality), generally account for discrete semantic structures and the abstract nature of human

[29] Also, recall Gregory's (1969) and Clark's (1989) comments about computationalism and the availability of resources in human cognitive processing.

thought. On the other hand, while Fodor and Pylyshyn agree that "understanding both psychological principles and the way that they are neurophysiologically implemented is much better (and, indeed, more empirically secure) than only understanding one or the other" (p. 45), they nevertheless remain fully committed to the computational theory of mind. They happily embrace Turing's idea: the claim that the mind has the architecture of a classical computer is not a metaphor, but a literal empirical hypothesis (Pylyshyn 1984, Fodor and Pylyshyn 1988). And, since Fodor doesn't acknowledge any explanatory power to connectionist theory, he seeks for plausible psychological hypothesis elsewhere – in an innate linguistic structure of the Language of Thought (LOT; Fodor 1975, 2008). The general problem for the computational theory of mind then is explaining how symbolic representations can be grounded in the lower-level cognitive processes (such as perception and action) – domain, where connectionist modeling showed relative success. And, as many studies have shown (some will be discussed in later chapters) language acquisition and comprehension are strongly influenced by perceptual, social and cultural constraints.

There is no common ground for traditional paradigms to complement each other. While the proponents of symbolic approach argue for productivity and systematicity of language and thought (Fodor and Pylyshyn 1988, Pinker and Prince 1988) and hence for compositional semantics and syntactic constraints, connectionists claim that language processing is a "constraint-satisfaction process" sensitive to "local" semantic and contextual factors (Rumelhart and McClelland 1986). Nevertheless, by incorporating lower-level cognitive processes, context and environment, connectionism seems to offer psychologically more viable approach to language and cognition (e.g., Langacker 1974, Lakoff and Johnson 1980, Lakoff 1987, Cummins 1983, Smolensky 1988, Clark 1989, Clark 1991).

The problems of symbolic and connectionist modeling cannot be solved by merely developing or expanding upon more advanced cognitive models under the current theoretical assumptions of any of the two paradigms. None of the approaches offers an all-in-one theory. Missing from both is a credible semantic theory that could mediate between both levels in a true hybrid fashion. First, the symbolic and connectionist approach are solving problems on different levels and should really be

taken as complementary, rather than competitive cognitive theories (see Gärdenfors 2000). Second, neither of them nor their hybrid variations can fully account for essential aspects of human cognition, e.g. the semantics of natural languages. Third, over the years, cognitive psychology research on categorization and concept formation has uncovered aspects of natural language and cognition that cannot be fully supported by any of the traditional views[30]. For example, connectionist model might explain certain aspects of language acquisition, such as learning past tense forms of English words, and therefore to some extent successfully simulate some of the emergent properties of the language-learning process. But it cannot account for conceptual information, formation and structure of concepts and categories, etc. One reason being, connectionist representations operate on the lower, subconceptual level. The other, connectionist representations are typically implicit and not available to the system. Classical symbolic approach, on the other hand, is notoriously poor at explaining language acquisition and comprehension, especially in regard to individual's subjectivity, and social and cultural influences and constraints. As will be argued in Part III, the symbolic approach fails because concepts are not meaningless symbolic structures, but grounded in individual's experience (both perceptual and conceptual) and environment. *Meaning* is a conceptual beast that cannot be harnessed by any of the two traditional approaches alone.

[30] Also, see Deacon (1997) for a brilliant evolutionary account of language and cognition

PART III: SEMANTICS

Section 6: Realist semantics

8 Introduction: the notion of meaning and semantics

In previous chapters we have discussed two traditional computational paradigms of modeling language and cognition. Here, we discuss theoretical intuitions about language and semantics that underlie these models, and argue for an alternative semantic theory. To stay within the frame of the thesis, I will mostly focus on general notions, not on intricate details of individual theory[31].

According to Lewis (1970), there are two general approaches in modern semantics:

> I distinguish two topics: first, the description of possible languages or grammars as abstract semantic systems whereby symbols are associated with aspects of the world; and second, the description of the psychological and sociological facts whereby a particular one of these abstract semantic systems is the one used by a person or a population. Only confusion comes of mixing these two topics. (p. 170)

Unlike Lewis, I argue that for modeling natural language semantics, the two topics cannot be, and should not be, separated. To state my case, I start off with the discussion about meaning and semantics as proposed by Gärdenfors (1999a). According to Gärdenfors (ibid., p. 209), a theory of semantics should be able to answer four basic questions:

> (1) What are meanings? (the ontological question)
>
> (2) What is the relation between linguistic expressions and their meanings? (the semantic question)
>
> (3) How can the meanings of linguistic expressions be learned? (the learnability question)
>
> (4) How do we communicate meanings? (the communicative question)

[31] Much of the discussion from philosophy of language and linguistic theory is being omitted, here the focus is on basic building blocks and functional aspects that a particular theory brings to modeling semantic representations.

The ontological question gives us two competing semantic paradigms: *realist* semantics and *cognitive* semantics. Realist semantics is not about psychological validity and the linguistic system as used by an individual (and hence in accordance with Lewis' warning), but rather about the relationship between abstract linguistic system and aspects of the world. The main hypothesis of the realist semantics is that this relationship is independent of the meanings grasped by individual minds. According to the realist semanticist, "the meaning of an expression is something out there in the world" (ibid., p. 209); it is determined by the state of the world and truth-conditions. Hence, realist semanticist argues for an objective view of the world, emphasizing the relationship between linguistic expression and reality. In general, this relationship is explicated by the model-theoretic semantics of Montague, with its core in set-theoretical approach of modern logic. The realist semantics is truth-conditional: the meaning determines whether particular linguistic assertion rightly corresponds to the object or an observation in the world, i.e. under what conditions a particular sentence is true or false. The dependency between constituents is further defined by set-theoretical functions.

The realist approach to semantics comes in two flavors: the *extensional* semantics of Frege and Tarski, and the *intensional* semantics of Kripke and Montague. Common to both are notions of *reference*, *truth* and *inference*. The main difference between the two is in their interpretation of the *meaning* and *reference* to the world. In extensional semantics the reference is direct and unmediated, i.e. linguistic expressions correspond directly to objects in the world. In intensional semantics, linguistic expressions get their meaning indirectly via intensions.

Cognitive semantics, on the other hand, argues that meanings are mental entities – the relation to the external world and truth-condition are of secondary importance, what counts is the relation between natural language expressions and individual's conceptual structure. Arguably, the overall agenda of cognitive semanticist is best described by the slogan "meanings are in the head". Hence, cognitive semantics is dependent on conceptual structures of individual language user and context, not on objective atomic facts and logical calculus.

The two theories differ in all of the four questions above. In what follows, I will mostly focus on answering the semantic question; answers to the learnability and

communicative questions will be given indirectly. The view adopted is that of cognitive semantics. The topic is what Lewis labeled "psychological facts" or how "one of these abstract semantic systems is ... used by a person or a population". Unlike Lewis, and most proponents of realist view, I argue that the psychological or cognitive aspects of language should be incorporated into semantic theory. Contra Lewis, I argue that realist semantics, describing meaning in terms of purely abstract symbol system and its reference to the world, cannot successfully answer questions (2) and (3) above precisely because it tries to separate the semantic theory from the cognitive psychology. Moreover, I will argue that the realist view could be psychologically plausible only if it adopted the conceptual aspect of cognitive semantics. Or more to the point, the meaning of an expression (or sentence) should be determined via the conceptual structure of individual language user, not through direct, truth-conditional mapping between language and external world or possible worlds.

9 Realist semantics

The realist approach to semantics is rooted in the philosophical movement that existed since Leibnitz (1646-1716) and became predominant with the development of modern logic and philosophy of Frege (1848-1925), Russell (1872–1970), Carnap (1891–1970) and early Wittgenstein (1889-1951), among others[32]. Along with the development of modern logic came also the ideas for propositional analysis of formal languages, i.e. how natural language could be formalized to conform to logical

[32] The connections between logic and philosophy go way back to the philosophy of Greeks, with notable examples in Pythagoras' theorem, Euclides' Elements and Aristoteles' writing on logical thinking (Organon). In Organon, for example, Aristoteles laid three fundamental laws of logic: a) the law of excluded middle (an object cannot have both a property and the opposite property), b) modus ponens (if all B's are C's and all A's are B's then all A's are C's), and c) modus tollens (if all B's are C's and no A's are C's then no A's are B's).
Later, logic reemerged occasionally through the centuries in philosophy of Ockham (1300 a.d.), Francis Bacon's Novum Organum (1620) and Descartes' Discours de la Methode (1637), but finally found more stable ground in Leibnitz's *De Arte Combinatoria* (1676), Newton's calculus, Euler's (1761) system of logic diagrams, Mill's System of Logic (1843), and Boole's work on symbolic logic in *The Laws Of Thought* (1854). In the latter, Boole argued that, besides solving mathematical problems, logic could be applied to thought in general. His ideas contributed to the evolution of modern 'propositional' and 'predicate' logic and to philosophy, inspired by logical formalization of thought.

calculus. In *Foundations of Arithmetic* (1884/1974) and *Sense and Reference* (1892/1980), Frege laid the fundamentals by providing logical formalism which constituted the first predicate calculus (representing the internal structure of propositions) and the truth-functional propositional calculus. In *Principia Mathematica* (1903), Russell further refined propositional and predicate calculus, arguing for all aspects of meaning to be explicit: the language and the world were seen as logical structures, constructions of atomic facts based on logical primitives and truth-functional semantics. The content of a sentence is the proposition expressed. The focus is on propositional semantics, i.e. meanings of sentences, not words. In *Foundations of Arithmetic*, Frege (1884/1974) claims:

> Only in a proposition have the words really a meaning … It is enough if the proposition taken as a whole has a sense; it is this that confers on its parts their content (p. 71).

Realist position emphasized the objective reality of the world where linguistic categories are defined as sets of necessary and sufficient conditions strictly dependent on the reference to the things in the world. Its aim, in Russell's words, is "axiomatization" of thought.

9.1 Sense and reference

In general semantics, the meaning of an expression is a certain kind of entity, and the fundamental concern of semantic theory is the nature of such entity and the formalization of the relationship between the two: the expression and its meaning. The underlying hypothesis common to all variations of realist semantics is Frege's theory of reference (Frege 1892/1980). In its simplest form, the theory of reference defines the meaning of an expression to be its extension. All meaning is grounded in extension: the extension of a sentence is its truth-value and the extension of an expression is its referent. The notion of truth takes a central place of logic and semantics[33]. It is reflected in his truth-functional semantics and the *Principle of Compositionality*:

[33] Frege's aim was to avoid psychologism. For Frege, as Boden points out, "[p]sychologism is any approach which confuses formal logic (or norms of rational thinking) with empirical facts about how

> If our supposition that the reference of a sentence is its truth-value is correct, then the latter must remain unchanged when a part of the sentence is replaced by an expression with the same reference. And this is in fact the case... . If we are dealing with sentences for which the meaning of their component parts is at all relevant, then what feature except the truth-value can be found that belongs to such sentences quite generally and remains unchanged by substitutions of the kind just mentioned? (Frege 1892/1980, p. 35)

Principle of Compositionality explains how the reference of a complex expression is determined by the reference of its parts. Sentences have compositional structure and are determined by their logical form and the extensions of their parts: simple expressions compose complex expressions, which compose sentences. The focus is not on the meaning of an expression as some kind of special entity, but rather on the truth-value; i.e. on the contribution of individual expression (via reference) to the determination of the truth-value of a sentence in which it occurs. To appropriately formalize the propositional semantics, some further constraints need to apply.

The general problem arises when there are exceptions to such simple definition, and in natural languages, there are many. The following is just a short overview of some of the issues faced by theory of reference, upon which various realist theories of semantics are built.

First obvious problem is how to deal with expressions that have no referent (e.g. 'Santa Claus'). Here, the truth-value cannot be established. Another problem arises when two expressions share the same reference, but differ in meaning. In "Über Sinn und Bedeutung" (1892/1980), Frege argues that expressions 'Phosphorus' (an ancient term for the morning star) and 'Hesperus' (the evening star) have the same extension, i.e. the same referent, a planet Venus, but not the same meaning. Similar case is put forward by Putnam's 'Twin Earth experiment' (Putnam 1975), with terms 'water' and 'H_2O'. In such cases, the reference of an expression alone does not explain the contribution or role of individual expression in determining the truth-value of all sentences it occurs in. Frege argued that there is more to the theory than just reference. To know the difference between Phosphorus and Hesperus requires

people think. Frege's point was not that people don't—or don't always—think logically. It was that whether they do or not—and how they do, when they do—is of no interest to the logician. The logical should always be distinguished from the psychological, since the (normative) laws of logic are not the (empirical) laws of thought. In modern philosophical jargon, Frege's position was that logic, or rationality, can't be "naturalized"" (Boden 2006, p. 121). Similar sentiment comes from quote of Lewis at the beginning of this chapter.

cognitive effort, i.e. to *know* the meaning. "Hesperus is Phosphorus" is *cognitively significant* whereas "Hesperus is Hesperus" is not. Frege's solution to the problem is in defining two components of meaning: *sense* (Sinn) and *reference* (Begriffsschrift).

What then, is the relationship between *sense* and *reference*? Frege argued that the sense is the "cognitive value" or "mode of presentation" (Art des Gegebenseins) of the referent (1892/1980, p. 56). Every expression that has an extension also has a *sense*, and the *difference in cognitive significance is a difference in sense* – sense reflects cognitive significance (Chalmers 2002). Besides reference then, expressions also have *sense* or *content*, i.e. a non-extensional aspect that also affects the truth-value of a sentence. Thus, *sense determines the reference*. By this definition, two sentences can express different propositions (have different content) while having the same truth-value, but not vice-versa, i.e. two sentences expressing the same proposition (have the same content) cannot have different truth-values. For example, Sentences 'Hesperus is a planet' and 'Phosphorus is a planet' have different senses, but the same referent. Similarly, 'sense determines the reference' also applies to expressions: two expressions with the same referent can differ in content or sense (recall morning and evening star), but not vice versa (two expressions with the same content cannot differ in reference).

9.1.1 The Fregean notion of *sense* for natural language

Still, existing definition of *sense* seems opaque and does not offer a plausible semantics for natural languages. The problem with Fregean extensional semantics is it ascribes the 'sense determines reference' in a manner of *rigid designator*[34]. *Senses* are not mental entities, but propositions as primary bearers of truth. Moreover, the truth-value of a proposition is defined absolutely, in terms of *true* and *false*. For Frege, *senses* are objective:

> The reference of a proper name is the object itself which we designate by its means; the idea, which we have in that case, is wholly subjective; in between lies the sense, which is indeed no longer subjective like the idea, but is yet not the object itself. The following analogy will perhaps clarify these relationships. Somebody observes the Moon through a telescope. I compare the Moon itself to the reference; it is the object of the observation, mediated by the real image projected by the object glass in the interior of the telescope,

[34] The latter is Kripke's term (1972) for defining proper names, but Frege's approach concerns any kind of expression.

> and by the retinal image of the observer. The former I compare to the sense, the latter is like the idea or experience. The optical image in the telescope is indeed one-sided and dependent upon the standpoint of observation; but it is still objective, inasmuch as it can be used by several observers. At any rate it could be arranged for several to use it simultaneously. But each one would have his own retinal image. (Frege 1892/1980, p. 60)

Such objectivity does not fare well with the semantics of natural language. In natural language, most expressions are not rigid; they can have different references in different situations. There are sentences with truth-value different given the occasion. For example, the sense of 'It is really cold here now' depends on the occasion of use. Hence, there is no objective notion of *sense* (of an expression) that could determine the same reference to every possible situation – the sense can vary between occasions of use.

Frege was aware of this problem[35] and argued that *sense* is not a universal feature of an expression. According to Chalmers:

> ... Frege's view entails that one cannot always attach sense to expression *types*. To handle cases like this, one has to attach sense to expression *tokens* (or to expression types as used in specific contexts, or to something else that is tied to an occasion of use). It follows that on Frege's understanding, the sense of an expression should not be identified with its *linguistic meaning*, where the latter is required to be common to all tokens of an expression type. (Chalmers 2002, pp. 141-142)

On such interpretation, at least the thesis that 'sense of a sentence has an absolute truth-value' should be rejected. Sentence (or an expression) cannot be true or false absolutely, sentence is true or false relative to a subject, context and time. Related to the discussion, it is important to note there seems to be an implicit paradox in Frege's Principle of Compositionality, as is evident from the following quotes:

[35] Using the distinction between *sense* and *reference*, Frege's semantics goes beyond the formalism of general theory of reference: "[f]rom the standpoint of logic as such, we need an account of the working of language only as it relates to truth ... Frege's philosophical concerns go a long way beyond anything that is the proper concern of the logician" (Dummett 1973: 83). Moreover, while he insists that "the truth-value of a sentence . . . is true or false. There are no further truth values" (1892/1952, p. 63), he nevertheless admits that some cases, for example literary texts (in a sense of a Dichtung) are beyond reference, and hence, truth value.
"In hearing an epic poem ... we are interested only in the sense of the sentences and the images and feelings thereby aroused. The question of truth would cause us to abandon aesthetic delight for an attitude of scientific investigation" (ibid.). The problem is, within the Fregean framework, the latter cannot be semantically formulated.

> Frege says that we are interested in the significance of any part of a sentence only insofar as we are interested in the truth-value of the sentence. *Is this not to say that the significance of the parts of sentences, and in particular of names, consists in their contribution to the truth-value of the sentences into which they may enter? In this case we should have to take the significance of sentences as primary*. …it was Frege himself who had opened a new approach with the famous dictum in his Grundlagen: 'Only in the context of a sentence does a word signify anything'. It is this statement which points to the conception of significance as truth-value potential. (Tugendhat 1970, pp. 180-182; italics added)

> … Tugendhat [1970] is surely right in the substantial point he is making. It is that the semantics of sense and reference is primarily a semantics of whole sentences and not of sentence parts. ... *Even when Frege expresses himself in terms that seem at first sight barely compatible with the thesis that sentence meaning is primary, the justification for assigning sense to names turns out to be that they must have a sense because they must make a contribution to the sense expressed by the whole sentence*. (Sluga 1980, p.158; italics added)

The problem of above, arguably genuine interpretation of Frege's Principle of Compositionality, did not slip under the radar of most scholars. While Fodor and Lepore acknowledged the difficulty of the situation, Haaparanta (1985) and Dummet (1981) were concrete:

> The compositionality principle says that the senses of the ingredients of a sentence S are more basic than the sense of S, for the sense of S is compounded out of them. Now, if Frege holds the view that in order to understand the sentence, we must understand the senses of the words it contains, he cannot demand that in order to understand the senses of words, we must know the sentences in which the words occur. (Haaparanta 1985, p.90)

> This is a difficulty which faces most readers of Frege. … The thesis that a thought is compounded out of parts comes into apparent conflict, not only with the context principle, but also with the priority thesis; but Sluga takes no notice of either conflict. (Dummett 1981, p.547)

> Where Frege himself stands is a little unclear. On the one hand, it's a famous Fregean view that words have meaning only as constituents of…sentences…; but on the other hand Frege certainly thought that the semantics of sentences is compositionally determined by the semantics of the words they contain (plus their syntax)… Whether, and in exactly what way, these doctrines can be reconciled is a notorious crux in Frege interpretation. (Fodor and Lepore 1992, p. 210)

Furthermore, there is a problem with reference. According to Frege, only expressions with an extension also have *sense*. How then, do we explain cases where there are expressions without a referent (recall 'Santa Claus')?

9.2 Possible worlds

A semantic theory for natural language should be able to explain how sense determines reference according to a particular situation or context, also in situations where referent is not present. According to one view (see (Kaplan 1989)), rules should be employed to determine content of an expression in a given situation. Rules would act as functions from contexts to contents and thus add additional component to the theory of meaning: a *character* of an expression that defines the content relative to a given situation. Kaplan argued, that in many cases the sole context of a given situation, "the context of utterance", is not sufficient – the reference of an expression must be relativized also to a *circumstance of evaluation*, i.e. "the possible state of the world relevant to the determination of the truth or falsity of the sentence"[36] (Speaks 2010).

While this is intuitively plausible, it further adds to the complexity of meaning as explained by propositional semantics. To account for situational and circumstantial evaluation we need to redefine the semantic theory in terms of *possible worlds*:

> The idea is that the meaning of an expression is not what the expression stands for in the relevant circumstance, but rather a rule which tells you what the expression would stand for were the world a certain way. So, on this view, the content of an expression like 'the tallest man in the world' is not simply the man who happens to be tallest, but rather a function from ways the world might be to men—namely, that function which, for any way the world might be, returns as a referent the tallest man in that world (if there is one, and nothing otherwise). This fits nicely with the intuitive idea that to understand such an expression one needn't know what the expression actually refers to—after all, one can understand 'the tallest man' without knowing who the tallest man is—but must know how to tell what the expression would refer to, given certain information about the world (namely, the heights of all the men in it). (Speaks 2010)

[36] This two-way approach of using both *character* and *circumstance of evaluation* to define meaning of an expression is also called double indexing semantics (proposed by Kamp (1971) and Kaplan (1989)).

Such functions are called *intensions*, hence *intensional semantics*. The classic examples of intensional semantics are Kripke's (1959) semantics for modal logics and set-theoretical semantics of Montague (1974). Both are built upon extensional semantics of Frege and Tarski. Kripke expanded Tarski's model-theoretic semantics by adding "possible" and "necessary" to the existing values of "true" and "false". In Tarski's refinement of extensional semantics, the focus is on formalization of semantics based on the theory of truth, and on the problem of how semantic theories could be consistently developed, rather than on semantics itself. In "The Concept of Truth in Formalized Languages" (Tarski 1958), Tarski set the foundations of "model-theoretic" semantics[37]. Inspired by Gödel's incompleteness theorem (1931), Tarski argued that truth cannot be defined within the language itself – instead, to avoid contradictions, the truth in the "object language" should be observed from the "meta-language". Tarski's theory of truth is extensional in a sense that the truth of a predicate is determined by a definite set of objects and properties. For natural language, such definition is impossible, since the set of objects is infinite. Kripke's modal logic, on the other hand, is intensional: we can refer to something without having the whole set of properties of that thing. Instead of mapping the language onto a *single primary world*, as is the case in *extensional semantics*, the *intensional semantics* maps the language onto a set of *possible worlds.* To a degree, this solves some of the problems of extensional semantics, including cases where there is no referent in the actual world. For Kripke (1980),

> A possible world isn't a distant country that we are coming across, or viewing through a telescope. ... *A* possible world is *given by the descriptive conditions we associate with it*. What do we mean when we say 'In some other possible world I would not have given this lecture today?' We just imagine the situation where I didn't decide to give this lecture or decided to give it on some other day. Of course, we don't imagine everything that is true or false, but only those things relevant to my giving the lecture; but, in theory, everything needs to be decided to make a total description of the world. We can't really imagine that except in part: that, then, is a 'possible world'. ... 'Possible worlds' are *stipulated*, not *discovered* by powerful telescopes. (p. 44)
>
> ... Most important, even when we *can* replace questions about an object by questions about its parts, we *need* not do so. We can refer to the object and ask what might have happened to *it*. So, we do not begin with worlds (which

[37] Later, Tarski's model theoretic semantics has been expanded by Kripke (1956; 1975) and Davidson's truth-conditional semantics natural languages (1967).

> are supposed somehow to be real, and whose qualities, but not whose objects, are perceptible to us), and then ask about criteria of transworld identification; on the contrary, we begin with the objects, which we *have*, and can identify, in the actual world. We can then ask whether certain things might have been true of the objects. (p. 53)

Kripke rejects Fregean view on sense and reference: proper names and natural kinds have a referent, but not in Fregean sense. The property cannot determine the reference as the object might not have that property in all worlds. Kripke distinguishes between designation (the mode of a reference) and the way reference is determined. For example, proper names and natural kinds are "rigid designators" for they designate the same object in every possible world. Non-rigid designators, on the other hand, can have different reference relative to possible worlds.

Inspired by Kripke's amodal logic, and by Frege's, Russell's and Carnap's work on mathematical logic, Montague, like his predecessors, treats language as a purely formal system, not as psychological phenomenon.

> A central working premise of Montague's theory … is that the syntactic rules that determine how a sentence is built up out of smaller syntactic parts should correspond one-to-one with the semantic rules that tell how the meaning of a sentence is a function of the meanings of its parts. (Partee 1975, p.203)

Montague's theory of semantics is arguably one of the most sophisticated achievements in the field of intensional logic. For Montague, logic and psychology are seen as two separate disciplines, the focus is on logic and philosophy of language.

To conclude, essential to all realist theories of semantics is to provide the logical, truth-conditional account of language. The relationship between the two is efficiently summarized by Lewis (1970):

> We call *the truth-value of a sentence the extension of that sentence*; we call the thing named by a name the extension of that name; we call the set of things to which a common noun applies the extension of that common noun. The extension of something in one of these three categories depends on its meaning ... It is the meaning which determines how the extension depends upon the combination of other relevant factors. *What sort of things determines how something depends on something else? Functions, of course; functions in the most general set-theoretic sense*, in which the domain of arguments and the range of values may consist of entities of any sort whatever, and in which it is not required that the function be specifiable by

> any simple rule. *We have now found something to do at least part of what a meaning for a sentence, name, or common noun does: a function which yields as output an appropriate extension when given as input a package of the various factors on which the extension may depend.* We will call such an input package of relevant factors an index; and we will call any function from indices to appropriate extensions for a sentence, name, or common noun an intension.
>
> Thus *an appropriate intension for a sentence is any function from indices to truth-values*; an appropriate intension for a name is any function from indices to things; an appropriate intension for a common noun is any function from indices to sets." (p. 23; italics added)

The problem with realist semantics is its primary liability to truth conditional view of the world, or *possible worlds*. In this sense, both theories succumb to a common underlying problem when faced with psychological aspects of meaning. If all about the meaning of a sentence (a proposition) is its truth condition, i.e., an account of the state of affairs in the real world that would make the proposition true, then such meaning is semantically anchored to the world indifferent of human language understanding.

Following chapter argues that maximization of truth is neither sufficient nor necessary condition for psychologically plausible theory of semantics.

10 Problems with realist view

10.1 Objectivist metaphysics

The implications that a realist view makes about language and cognition face some serious problems when viewed from the perspective of cognitive psychology. In what follows, I shortly revise some of the facts of realist semantics with a more general interpretation, and point to the problems it carries.

To begin with, the main problem lies in metaphysical realism or, in Lakoff's terms, "objectivist metaphysics" of the realist approach. For realist, reality comes with a unique deterministic structure in terms of entities (or sets of entities defined by the

common properties of the members), properties, and the relations holding among those properties.

> The world, as objectivist doctrine envisions it, is extremely well-behaved. It is made up of discrete entities with discrete logical combinations of atomic properties and relations holding among those entities. Some properties are essential; others are accidental. Properties define categories, and categories defined by essential properties correspond to the kinds of things that there are [i.e., natural kinds]. And the existence of classical categories provides logical relations that hold objectively in the world. (Lakoff 1987, p. 161)

This set-theoretical framework exists independent of any human understanding – it is logically independent of the human mind, and something which is, in its basic character, metaphysically fundamental. The upshot of such structure is

> [c]lassical categorization: All the entities that have a given property, or collection of properties in common, form a category. Such properties are necessary and sufficient to define the category. All categories are of this kind. ...
>
> There are natural kinds of entities in the world, each kind being a category based on shared essential properties, that is, properties that things have by virtue of their very nature. (ibid., p. 160)

Realist approach to language and cognition assumes that the mind functions as a mirror of nature: the language and thought correspond to entities and categories in the world via symbols, and that the world is structured in a way that makes symbol-to-world correspondences possible – that is, structured in a way that can be modeled by set-theoretical models. What makes such correspondence to the world possible, i.e. objectively definable and amenable to set-theoretical modeling, are 'natural kinds'. "Natural kinds", according to Putnam, have "some 'essential nature' which the thing shares with other members of the natural kind. What the essential nature is, is not a matter of language analysis but of scientific theory construction" (Putnam 1975, p. 104). Natural kinds therefore represent something that exists in the world independently of human cognition and is attainable only through manipulation of symbolic structures. The realist semantics is objectively referential.

Already in his early writings (Putnam 1975, 1975b), Putnam departs from traditional realist view of cognition. A much stronger argument, that shakes foundations of classical theory, comes some years later in form of what is called Putnam's theorem

(Putnam, 1981). There, Putnam argues that model-theoretic semantics fails as a theory of meaning. The following tree tenets of Metaphysical realism came under Putnam's critique:

(1) that "the world consists of a fixed totality of mind-independent objects",

(2) that "there is exactly one true and complete description of the way the world is", and

(3) that "truth involves some sort of correspondence" (Putnam 1990, p. 30).

The casual realist reasoning could be interpreted as follows: thesis (1) is an underlying requirement for validity of thesis (2), both naturally suggest thesis (3), but thesis (3) requires a predefined or "ready-made world", thus thesis (3) suggests thesis (1) (Putnam 1983, p. 211). Putnam (1981) argues that such approach, where its constituents are objectively determined to provide "exactly one true and complete description of the world", is not only wrong but unintelligible.

> What does it mean ... to speak of mind independency? Human minds did not create the stars or the mountains, but this "flat" remark is hardly enough to settle the philosophical question of realism versus anti-realism. What does it mean to speak of a unique "true and complete description of the world?" (p. 52.)
>
> ... think of the world as consisting of objects that are at one and the same time mind-independent and Self-Identifying. This is what one cannot do (ibid., p. 54)

In (Putnam 1975, 1981), Putnam develops the view of "internal realism". Contrary to traditional view, internal realism is consistent with conceptual relativity, i.e. the observation that truth primarily depends on the conceptual scheme that we employ, not on the "God's Eye View" of realist metaphysics. The conceptual relativity is in the hands of the individuals, it comes from our cognition and interaction with the world, not from some presupposed definite reality. Hence, the metaphysical realist cannot accept conceptual relativity without the above theses falling apart.

And as Lakoff (1987) point out,

> [m]odel theory is, of course, the natural mathematization of objectivist semantics. What Putnam is suggesting is that there can be no such possible mathematization. That is, objectivist semantics cannot be made precise without contradiction. (pp. 230-231)

10.2 Referential representations

The objectivist view is strongly reflected in the classical computational approach to cognition. Within the realist tradition, two different theories (extensional vs. intensional) are divided by the question whether our perceptual access to the physical world is direct or mediated; with the former defending ontologically immediate and non-representational reference to the world, while the latter constitutes the reference through the representational system, i.e. via symbol system registering the presence of the object or the relevant aspects of its character. The latter underlies classical computationalism, since it aims to explain our contact with physical items as mediated by some form of mental representation. As Foster (2000, p. 1) points out: "[i]n place of the claim that our perceptual access to the physical world is direct, it insists that the perceiving of a physical item is always mediated by the occurrence of something in the mind which represents its presence to us" – our conceptual symbol system.

We will call this a representational realism, where external world is mediated by (internal, mental) representations, and cognition itself is described as manipulation of abstract symbols. By this view, concepts are discrete symbols that correspond to entities and categories in the world. Our conceptual symbol system is innate and made meaningful via its capacity to correspond correctly to these entities and categories in the world. According to realist view, "mental representations must thus be 'semantically evaluable' – capable of being true or false, or referring correctly or failing to refer correctly" (Lakoff 1987, p. 163). Our representation is representation of external reality, a mirror of logical relations among entities and categories in the world, independent of belief, knowledge, perception, modes of understanding, or any other aspect of individual's cognition. The success of our interacting with the world depends on our ability to successfully represent this external reality: "[k]nowledge consists in correctly conceptualizing and categorizing things in the world and grasping the objective connections among thing in those categories" (Lakoff 1987, p. 163). Concepts and categories of mind are mental representations of objects and categories in the world, detached from any kind of nonobjective influences that could make our knowledge objectively inaccurate, such as products of imagination (metaphor, metonymy, mental imagery etc.). Meaning is based on truth: the meaning of a sentence is taken to be its truth conditions, the conditions under which the

sentence would be true. Thought then, becomes a manipulation of abstract symbols, which get their meanings via correspondence with entities and categories in the world (or possible worlds). And our language becomes the Language of Thought.

10.2.1 Symbolic formalism, objective categories and natural language

There are two general representational formalisms of realist approach to meaning: the feature list approach (Smith, Shoben and Rips 1974) and the propositional structure approach (Collins and Loftus 1975). Both are victims of the realist view of the world. The basic idea of the former is that the words and objects get their meaning by belonging to a category (e.g., furniture, animal, plant), which, by objectivist definition, requires having the right defining features (or properties). As Barasalou (1993) points out, this definition might perfectly sensible on the surface – at least In Western culture, we have gotten accustomed to such world view, which has prevailed from the times of Aristotel – but there are many logical and empirical problems with such approach. First, for realist approach to succeed, we need to have the set of all the necessary features and appropriate relations established beforehand. By such account, it doesn't seem possible for us to know or recognize an object if neither the categories in the world nor the categories in our head have defining features. And therein lays the problem. Take a simple object such as table for an example: it is not enough to define its constituents (e.g., a top and legs), but we also need to define right relations among them (i.e., the legs must be below the top and support the top). Moreover, tables come in all kinds of forms (e.g., different shapes, number of legs, without legs, different function etc.) and it would be impossible to have an exhaustive list of every single necessary feature for each case. Inversely, imagine the set of necessary features that define a general class, i.e. comprising all tables, dogs, cats, etc. in the world. Or, take Lakoff's example of a category mother: there is the prototypical "birth mother" that bears a child and nurtures it; a biological mother who provides genetic materials but does not bear the child or nurture it; a surrogate mother who bears the child but does not provide genetic material; adoptive mothers; the mother of invention, etc. The ambiguity of these examples makes the

notion of an objective category unrealistic. Most categories do not seem to have an ontological structure defined by a set of necessary and sufficient conditions[38].

10.3 Relations to Cognitive psychology

Most critically, there is not much experimental evidence from cognitive science supporting the realist view on categorization and concept formation (cf. critical contributions from different areas of cognitive science, e.g, Rosch and Mervis 1975, Mervis and Rosch 1981, Smith and Medin 1981, Medin 1989, Anderson 1991, Gelman 1996, Barsalou 1983, 1985, 1999, Grossman *et al.* 2002, Jäkel 2007 and Jäger 2007, among others). Moreover, experimental research in cognitive psychology has shown (especially in prototype theory introduced by Rosch (1975, 1978)) that categories do not conform to the rules of logic and ontological view of the world as one based on defining features. In most cases, the structure of a category is "radial"– that is, the category has some central or prototypical members with marginal members related to these central members, both, by an extent of shared features and by metaphorical extension (Lakoff 1987, Lakoff and Johnson 1980). The deterministic structure of categories, it seems, could be appropriate only in matters of mathematics and logic.

We can infer from the realist definition that essential features forming such categories are abstract, amodal, arbitrary elements that take on their meaning by a principle of compositionality. From a standpoint of cognitive psychology, such approach has serious logical and empirical problems. The logical problem includes the *symbol grounding* problem (Harnad 1990) and Chinese Room argument (Searle 1980, Harnad 1989), claiming that meaning cannot arise solely from syntactic relations between arbitrary symbols. We need to have access to mental content:

> [f]ormal symbols by themselves can never be enough for mental contents, because the symbols, by definition, have no meaning (or interpretation, or semantics) except insofar as someone outside the system gives it to them (Searle 1989, p. 45).

[38] For exmple, even experts in biology do not agree as to what are the proper criteria for classification. Here, Linnaean taxonomy comes to mind: though exceptionally systematic, the groupings of observable characteristics and relationships in Linné's hierarchical classification have significantly changed since their conception in 18th century, as have the principles behind them (but see (Lakoff 1987) for other examples).

Learning the meaning of words is not analogous to processing abstract symbol structures. The empirical problem, on the other hand, argues that human performance in category tasks is very much influenced by context and modality (Jacoby and Dallas 1981, Barsalou 1987, 1993, 1999, 2005, Wisniewski and Medin 1994, Hintzman, 1986, Hampton *et al.* 2006).

The propositional structure inherits the same problems. Propositions do give structure to categorical knowledge and account for reasoning; for example, explaining necessary relations between constituents of the category table: e.g. "Legs are located beneath to support the table top". But meaning is anchored in the same sets of necessary and sufficient conditions as is the feature list. And the tokens of the constituents are abstract symbolic structures manipulated by syntactic rules. This brings us back to the *symbol grounding* problem, and, indirectly, to the implicit paradox in Frege's Principle of Compositionality. Further, such representation of the state of affairs in external world is central to Fodor's Representational Theory of Mind and to computational approach to cognition in general. Fodor is quite clear:

> What I am selling is the Representational Theory of Mind ... At the heart of the theory is the postulation of a language of thought: an infinite set of 'mental representations' which function both as the immediate objects of propositional attitudes and as the domains of mental processes. (Fodor 1987, p. 16-17)

The inherent problem to the realist semantics is there is no plausible explanation for how meaning could be eventually introduced into a system of meaningless symbols. Alone, realist semantics cannot answer the learnability question, "a semantic mapping between a language and a world (or several worlds or a partial world) does not tell us anything about how individual users "grasp" the meanings determined by such a mapping" (Gärdenfors 2000, p. 214). Harnad (1987) argues that realist view is ungrounded:

> ... the meanings of the atomic terms of its sentences cannot simply be derived from still more sentences without infinite regress. ...the meanings of elementary symbols must be grounded in perceptual categories. That is, symbols, which are manipulated only on the basis of their form (i.e., syntactically) rather than their "meaning," must be reducible to nonsymbolic, *shape*-preserving representations. Semantics can only arise when the interpretations of elementary symbols are "fixed" by these nonsymbolic,

> iconic representations and their causal connections to input and output from the world. (p. 550)

Nor has Lewis' agenda of keeping the two topics separated (i.e., logical vs. psychological and sociological aspects of semantics), granted the realist approach a superior, or at least more plausible, theory of meaning in natural languages.

Glenberg (1997) sums it up nicely:

> How did we get ourselves into this mess? The problem stems from trying to develop psychological theories of meaning on the basis of philosophers' analyses of formal languages. Because natural language is messy, most philosophical accounts of meaning have been constructed within a formal language, such as predicate calculus. The symbols in a formal language are intended to be meaningless so that they can be operated on by formal syntactic procedures. These symbols and sentences are given meaning by mapping them onto elements in a formal model of the world. Not only is that mapping formidable (and perhaps impossible in principle, see Putnam, 1981), but also it requires the sorts of Aristotelian categories that do not appear to exist in the real world.
>
> Note that the philosopher's problem is very different from the psychologist's problem. The philosopher is dealing with a formal language the elements of which are designed to be meaningless, whereas the psychologist is dealing with a natural language with elements that are designed to convey meaning. The philosopher is attempting to discover the "universal" meaning of formal sentences, that is, what a given set of relations among elements will mean for all times and all places. In natural languages, however, the meaning of a sentence depends critically on its context as well as on the experiences of the individual hearing the sentence. The psychologist needs to discover how a natural language sentence can have a particular meaning for a particular individual. (p. 508)

Section 7: Cognitive semantics

11 Introduction: the rise of cognitive theories

In cognitive sciences the realist view prevailed until mid-70's of previous century, when alternative theories of cognitive semantics and linguistics emerged to challenge the classical realist approach to cognition and its modeling in AI. But the core criticism came earlier, from philosophy itself: starting with Wittgenstein's (1953) repudiation of realist approach to language and meaning (including his own previous work) as fundamentally misguided, Putnam (1975, 1981), and later from scholars from various fields of cognitive sciences, most notably from phenomenological philosopher Hubert Dreyfus (1965, 1972) and cognitive psychologist Elanor Rosch (1978a/b). The 'discovery' of *family resemblance* concepts (with the underlying hint that grammar is arbitrary) resulted in Kuhnian paradigm shift, and is arguably Wittgenstein's most influential contribution to the development of cognitive sciences, and consequently to the demise of classical theory[39]. In Philosophical Investigations (1953), Wittgenstein argued that concepts and categories are not of fixed necessary and sufficient definitions, as is the realist mantra, but are necessarily context-dependent. Using an example of concept 'game', Wittgenstein argued there is no single property common to all games in virtue of which we call them 'games'; instead, there is "a complicated network of similarities overlapping and criss-crossing: sometimes overall similarities, sometimes similarities of detail" (§ 66). Moreover, "… the term 'language-game' is meant to bring into prominence the fact that the speaking of language is part of an activity, or of a form of life" (§ 23). What Wittgenstein argues is that "grammar is not abstract, but situated within the regular activity with which language-games are interwoven ... It is through analyzing language's illusive power that the philosopher can expose the traps of meaningless philosophical formulations" (Anat and Anat 2011). While not discussing psychological dimension, nevertheless, Wittgenstein's investigation into everyday psychological concepts and language use had an important influence on the development of cognitive science.

[39] Later, Elanor Rosch, a cognitive psychologist strongly inspired by Wittgenstein, developed a series of very influential theories on concept formation and categorization.

12 Cognitive semantics

Cognitive semantics theory challenges the prevalent realist view in every possible aspect. First, the main tenet of cognitive semantics is: *meanings are mental entities*. As we have shown, semantics is traditionally understood as a relation between language and the world (see Figure 6: a) and b)). This approach is generally called *referential* semantics, because it claims that sentences get their meanings by referring to the real objects and events. The problem with referential semantics is its objectivist metaphysics. Such view of the world is not psychologically real – it does not explain language acquisition and comprehension nor nuances in conceptual and categorical structure. Moreover, in the real world there are no objectively determinate sets or ultimate lists of necessary and sufficient features or conditions – we don't know or have access to such sets or lists. Meanings have to be perceptually grounded (Harnad 1990, Gärdenfors 1997). Therefore, referential semantics is not acceptable as a cognitive theory.

In cognitive semantics, on the other hand, the emphasis is on the graded structure of concepts and categories of natural language. The cognitive answer to the semantic question (2) is words get their meanings by mappings onto conceptual structure (Figure 6 c)). The semantics becomes the relation (via set of associations) between expression and conceptual structure, i.e. mental representations of individual language users, not external world defined by objective facts. Instead of propositions, cognitive semantics operates on lexical meanings of words. As Gärdenfors (1999b, p. 21) puts it: "meaning is conceptualization in a cognitive model."

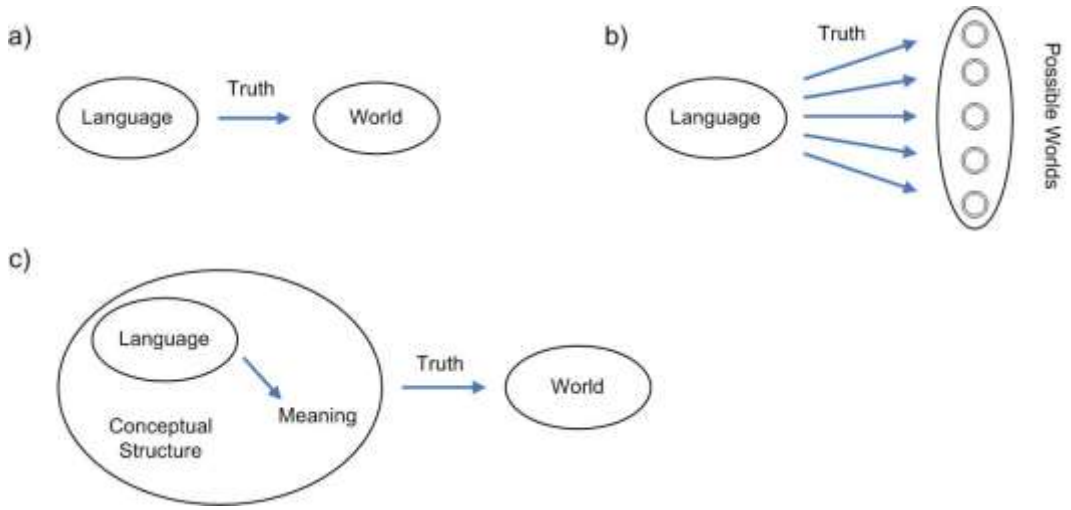


Figure 6: Semantics: a) the ontology of extensional semantics; b) the ontology of intensional semantics; c) the components of cognitive semantics (adapted after (Gärdenfors 1999a, p. 210))

In same vein goes the cognitivist answer to the learnability question (3): since meanings are conceptualizations, we get the meaning of an expression via associative link to our cognitive structure. And language itself is seen as a part of the cognitive structure, not a separate entity isolated from the user. Since meanings map onto individual's conceptual structures, the notion of truth and objective reference to the external world become secondary, what matters, is belief. Moreover, such meaning is *grounded*. Cognitive semantics offers a natural explanation of the relationship between perceptual and cognitive mechanisms and thus grounding of meaning:

> Since the cognitive structures in our heads are created mainly by our perceptual mechanisms, directly or indirectly, it follows that *meanings are*, at least partly, *perceptually grounded*. This, again, is in contrast to traditional realist versions of semantics which claim that since meaning is a mapping between the language and the external world (or several worlds), meaning has nothing to do with perception. A consequence of this is also that language and semantics is not seen as separated from other forms of cognition, but interacts with perception, memory, concept formation, etc. (Gärdenfors 1999a, p. 211; italics added).

The realist view has difficulties explaining any of the essential aspects of natural language related to cognition: e.g., how meanings are grasped by the individual language user, how perception influences categorization and cognition in general (i.e., how concepts and categories emerge, how we fill in partial information, etc.), the workings of semantic memory, etc. Ultimately, realists have a problem answering learnability question (3) precisely because of their exclusively formal semantics,

detached from perception and psychological and sociological aspects of language and its user. Hence, as Gärdenfors (1999a) points out, there are "insurmountable problems" for the realist explanation of learning new words:

> How could an associative link to the world function in these cases? I can't take seriously an answer from intensional semantics that presumes associative links between sounds and entities in merely "possible" worlds. How would such a link be physically realized? How could one learn something about a non-actual world? I conclude that realist brands of semantics have serious problems with the learnability question. These problems become particularly tangible if we consider the task of constructing a robot able to learn the meaning of new words that have no immediate reference in the environment of the robot. (p. 215)

The point is, philosophical 'possible worlds' imply ontological stances that, without grounding, cannot explain what is intuitively meaningful to individual language user (for a discussion of other implications of traditional philosophical account on language, see for example (Brandt 2005)).

12.1 Image-schematic representation of meaning

The unique approach of cognitive semantics is ultimately evident in its formalization. Unlike syntactic structure of Fodor's Language of Thought, cognitive models are *image-schematic*, based on *geometric* constructions, not propositions (Gärdenfors 1996, p. 164). Since cognitive semantics emphasizes the relation between language and cognitive structure, the semantic elements are being modeled as *spatial* or *topological* objects. While spatial or schema-like functions are common to all kinds of image schemas, general definitions of *what an image-schema is* are quite vague and heterogeneous, cf. (italics added):

> ... *dynamic pattern* that functions somewhat like the *abstract structure of an image*, and thereby connects up a vast range of different experiences that manifest this same *recurring structure* (Johnson 1987, p. 29)

> ... a recurring structure of or within our cognitive processes, which establishes *patterns of understanding and reasoning*. Image schemas emerge from our bodily interactions, linguistic experience and historical context (Johnson 1987, p. 256)

> The most useful way of understanding image schemas is to see them as *mental representations* of fundamental units *of sensory experience* (Grady 2005, p. 44)

> ... part of our *non-representational* coupling with our world, just as barn owls and squirrel monkeys have image schemas that define their types of sensorimotor experience ...
>
> ... the basis for our understanding of all aspects of our perception and motor activities. ...
>
> ... *activation patterns* (or "contours") in human *topological neural maps* (Johnson and Rohrer 2007, p.33)
>
> ... *dynamic analog representations* of spatial relations and movements in space. (Gibbs and Colston 1995, p. 349)
>
> ... *abstract mental pictures* with an inherent spatial structure, constructed from elementary topological and geometrical structures like "container," "link", and "source-path-goal." Such schemas are commonly assumed to *constitute the representational form common to perception, memory, and semantic meaning*. (Gärdenfors 2011, p. 1)

Image-schemas reflect the systematic relations of elements constituting a language, but are nothing like words in a language. Moreover, they are abstract mental pictures with inherent spatial structure that is schematic, not picturesque. They are

> ...that part of a picture which remains when all the structure is removed from the picture, except for that which belongs to a single morpheme, a sentence or a piece of text in a linguistic description of a picture ... (Holmqvist 1993, p. 31).

Some classic examples of image-schematic cognitive modeling are Fillmore's frames (1982, 1985), *image-schemas* in Langacker's theory of cognitive grammar (Langacker 1986, 1987), *metaphoric* and *metonymic mappings* in Lakoff (1987) and Lakoff and Johnson (1980), *mental spaces* in (Fauconnier 1985, 1997; Fauconnier and Sweetser 1996) and *conceptual spaces* in Gärdenfors (1988, 1991, 1996, 1997, 2000). All these different variations fall under a general term 'image schema'. A general characteristic of an image schema is its *inherent spatial structure*, composed of spaces or basic *domains*:

> It is however necessary to posit a number of 'basic domains', that is, cognitively irreducible representational spaces or fields of conceptual potential. Among these basic domains are the experience of time and our capacity for dealing with two and three-dimensional spatial configurations. There are basic domains associated with various senses: color space (an array of possible color sensations), coordinated with the extension of the visual field; the pitch scale; a range of possible temperature sensations (coordinated

> with positions on the body); and so on. Emotive domains must also be assumed. It is possible that certain linguistic predications are characterized solely in relation to one or more basic domains, for example time for (BEFORE), color space for (RED), or time and the pitch scale for (BEEP). However, most expressions pertain to higher levels of conceptual organization and presuppose nonbasic domains for their semantic characterization. (Langacker 1987, p. 5)

Further, Lakoff (1987), Johnson (1987) and Lakoff and Johnson (1980) argue that the main vehicles of conceptual interactions among these domains are metaphoric and metonymic operations (the aspects of natural language completely ignored by the realist approach):

> We discovered that the image-schema structure of the source domain is used in reasoning about the target domain. Moreover, by looking at hundreds of cases, we found that image-schema structure and image-schematic inferences seemed to be "preserved" by metaphors. That is, source domain containers (e.g., physical traps) are mapped to containers (e.g., metaphorical traps), with interiors mapped to interiors and exteriors mapped to exteriors. (Lakoff and Johnson 1980, p. 253)

An example of a general structure, with basic components of "container", "source-path-goal" and "link", is shown in Figure 7.

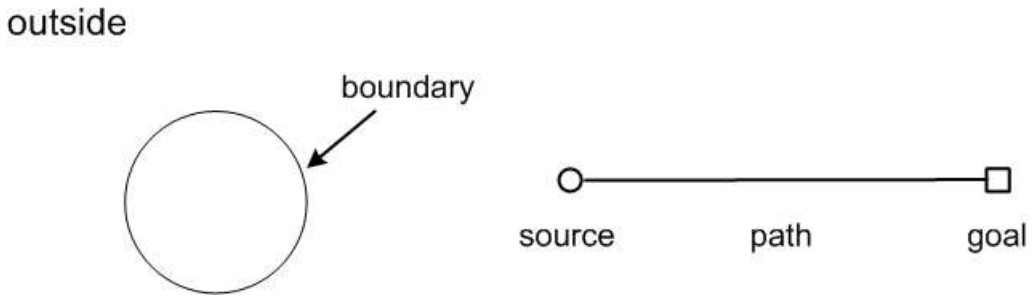


Figure 7: Image schemas: fundamental carriers of meaning (*container*, *source-path-goal* and *link*)

A more specific example is of an image-schema depicting a dynamic interpretation of English word *out* (Johnson 1987, p. 32-34) is shown below. In Figure 8, *out* is represented in various spatial senses by a metaphorical boundary. Possible interpretations are:

1. a case where a clearly defined trajectory (TR) leaves a spatially bounded landmark (LM) leaving a spatially bounded landmark (LM), as in "John went out of the room";

2. a case of where trajectory (TR) expands (LM), as in "She poured out the beans";
3. a case where containing landmark is implied and not defined at all, as in "The train started out for Chicago".

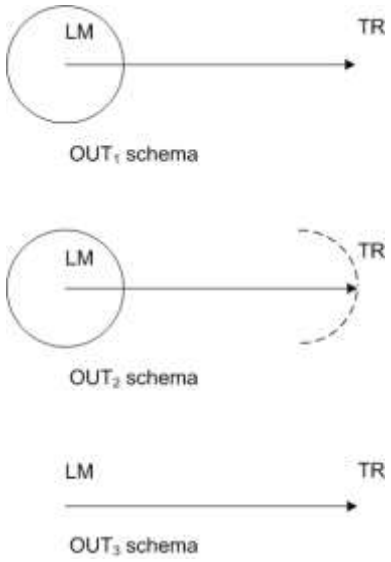


Figure 8: An image-schema of the English word *out* (adapted after (Johnson 1987, p. 32))

12.2 Problems with image schemas

The main problem with cognitive semantics is its variety of theories and different formulations of image schemas. Theories within cognitive semantics diverge on issues such as *representation* and *embodiment* (e.g., Varela, Thompson and Rosch 1991, Johnson and Lakoff 2002, Mandler 2004, Johnson and Rohrer 2007, Ziemke, Zlatev and Frank 2007, Gibbs 2006), *consciousness* (Johnson and Lakoff 1999, Talmy 2000, Thompson 2001, Evans 2003), level of *abstractness* (Grady 2005), *dynamicity* (Mandler 2004), *sensory modality* (Johnson 1987, Gibbs 2005, Ziemke, Zlatev and Frank 2007, cf. Piaget 1952) and *(inter)subjectivity* (Tomasello *et al.* (2005), Zlatev 2005, 2007, cf. Piaget 1962), among others. Furthermore, some researchers argue for neural foundation of image schematic mappings, for example, in studies on conceptual metaphors (Lakoff 1987, Johnson 1987, Lakoff and Johnson 1980, Gallese and Lakoff 2005), or in mappings between sensorimotor experiences and related brain regions. The latter has gotten much evidence from experimental

psychology and neuroimaging studies (e.g., Barsalou 1999, Pulvermüller 2001, Glenberg and Kaschak 2002, Gallese and Lakoff 2005, Rohrer 2005, Johnson and Rohrer 2007).

Overall, the lack of common underlying theory is pressing. The main reason for such heterogeneity of the field is in cognitive approach to semantics itself. In cognitive semantics, the semantic structures are not independent, but closely related to other cognitive mechanisms, particularly perception and memory. Hence, the study of semantics is inherent in the study of other, more general aspects of cognition.

> The idea is that since the acquisition and use of language rest on an experiential basis, and since experience of the world is filtered through extralinguistic faculties such as perception and memory, language will of necessity be influenced by such faculties. We can therefore expect the nature of human perceptual and cognitive systems to be of significant relevance to the study of language itself. One of the primary tasks of cognitive linguistics is the ferreting out of links between language and the rest of human cognition. (Regier 1996, p. 27)

Another problem, related to the issue above, is the lack of precise definition or formalism for image-schematic approach to modeling semantics. Without further constraints, the general notion of intrinsic topological or geometrical structure is too opaque. This became especially evident in various computational attempts of modeling cognitive semantics (e.g., a study by Holmqvist (1993)) where, due to the lack of mathematically defined parameters, the implementations of image-schematic representations have proven difficult.

And finally, linguistic meanings are not only individual, but public, *conventional* entities, a part of social environment affected by "linguistic power structure" emerging from the community (Gärdenfors 1993). Similarly, according to Tomassello (2003):

> Linguistic symbols are social conventions by means of which one individual attempts to share attention with another individual by directing the other's attentional or mental state to something in the outside world (p. 8)

In what follows, Gärdenfors' theory of *conceptual spaces* (Gärdenfors 2000) is proposed as an appropriate representational framework for modeling cognitive semantics. It purports to explain the interplay between individual and social aspects of meaning in relation to conceptual structure and category effects, and most

importantly (for the current topic), it defines mathematical foundation needed for computational modeling of cognitive semantics.

Section 8: Conceptual spaces

13 Conceptual Spaces: a framework for cognitive semantics

In (Gärdenfors 2000), Gärdenfors proposes a novel approach to modeling cognitive semantics: by using a notion of a *conceptual space*, concepts and properties are being represented in space geometrically based on their quality dimensions. Since concepts are represented spatially as vectors, the creation of conceptual spaces is mathematical and thus directly applicable for computer simulations. Conceptual spaces represent information by geometric structures rather than by symbols and propositions. Information is represented by points (standing for objects or individuals), and regions (standing for concepts, properties and relations) in multi-dimensional space. By exploiting distances in the space we can represent degrees of similarity between objects.

The general idea is based on cognitivist approach to semantics defining meanings as mental entities. Just like in an image schema, a notion of space "serves as a fundamental conceptual structuring device in language" (Regier 1996, p. 19). But unlike an image schema, conceptual spaces are a mathematically defined, based on "fundamental notions of geometry" (Gärdenfors 2000, p. 15). And unlike the realist approach to semantics and language, the conceptual spaces approach does account for psychologically plausible explanation of basic human cognitive processes of categorization and concept formation. As we have argued throughout the thesis, the fundamental issues of the realist approach come from its objectivist metaphysics, i.e., the view of the world defined in terms of necessary and sufficient conditions. The classical hypothesis of sets of defining features is not psychologically real (see, Rosch 1978, Smith and Medin 1981, Medin 1989 and Lakoff 1987). In general, there are no strict borders or ultimate lists of necessary and sufficient features, or any such criteria for category membership (in most psychological dimensions no clear-cut boundaries exists). Unlike the classic, realist approach then, the conceptual spaces do account for graded structure of categories as well as clears cases (e.g., of some scientific concepts).

13.1 Empirical evidence

There is an overwhelming empirical evidence underlying conceptual spaces approach to modeling cognition and, more precisely, for the conceptual basis of meaning. Here, I'll limit myself on some general, but essential examples that directly refer to the cognitivist theory and formulation of meaning through conceptual spaces.

13.1.1 Categorization and prototype theory

First, take for example the *category effect*. Various studies reveal the category structure in semantic memory in terms of *family resemblance*: items in the same category are more related or similar than items in different categories (Rosch 1973, Rosch and Mervis 1975, Smith and Medin 1981).

Second, in her studies of psychological principles of categorization (Rosch 1975, 1978), Rosch defined cognitive economy and structure in the perceived world as two general principles that underlie categorization system:

> The first has to do with the function of category systems and asserts that the task of category systems is to provide maximum information with the least cognitive effort; the second has to do with the structure of the information so provided and asserts that the perceived world comes as structured information rather than as arbitrary or unpredictable attributes. Thus maximum information with least cognitive effort is achieved if categories map the perceived world structructure as closely as possible. (Rosch 1978, p. 28)

These principles further affect both horizontal and vertical dimension of categorization. The former concerns prototype effects and "the segmentation of categories at the same level of inclusiveness - the dimension on which dog, cat, car, bus, chair, and sofa vary." (ibid., p. 30). The vertical dimension, on the other hand, "concerns the level of inclusiveness of the category - the dimension along which the terms collie, dog, mammal, animal, and living thing vary." (ibid., p. 30).

There are other important considerations. Strong empirical support has been shown for the prototype theory (e.g., in Rosch 1973, 1975, 1978, Rips, Shoben and Smith 1973, Mervis and Rosch 1981, Rosch and Mervis 1975, Rosch *et al.* 1976, and Lakoff 1987). The *prototype theory* and the studies on *typicality effect* have shown that within the same category not all category members are equal, some being prototypical and others less typical. Thus, the notion of prototypicality is defined as

'goodness of example', with the prototype seen as the best example or the "clearest case" among members of particular category. To borrow a classical example of prototypicality ratings (Rosch 1975), for a category 'bird', *robin*, *sparrow* and *bluebird* are seen as more representative than *chicken*, *penguin* or *emu*. What prototype theory purports to explain, are the asymmetries among category members and asymmetric structures within categories. The prototypicality effects and goodness-of-example ratings can be interpreted in terms of the internal structure of the category or a category membership[40]. To use just one example, some categories have "extendable boundaries", while others do not. Thus, category 'bird' has strict boundaries in the sense that all members belong to it in absolute terms, i.e. something is not a bird and a fish at the same time; a 'tall man', however, cannot be defined by itself, it needs a contrast class to compare with.

> In many cases, prototypes act as cognitive reference points of various sorts and form the basis for inferences. The study of human inference is part of the study of human reasoning and conceptual structure, hence, those prototypes used in making inferences must be part of conceptual structure. It is important to bear in mind that *prototype effects are superficial*. They may result from many factors. In the case of a graded category like tall man, which is *fuzzy* and does not have rigid *boundaries*, prototype effects may result from degree of category membership, while in the case of bird, which does have *rigid boundaries*, the prototype effects must result from some other aspect of internal category structure. (Lakoff 1987, p. 45; italics added)

13.1.2 Basic level categories

The vertical dimension of category system, on the other hand, gives an important insight into the relation between inclusiveness and abstractness of category structure, and the notion of *basic level*:

> ... not all possible levels of categorization are equally good or useful; rather, *the most basic level of categorization will be the most inclusive* (abstract) level at which the categories can mirror the structure of attributes perceived in the world. (Rosch 1978, p. 30; italics added).

[40] But see Lakoff (1987) for a different view. In short, Lakoff argues that the prototypicality ratings are not to be misunderstood in terms of constituting a graded membership (which has often been the case) in terms of some members being consequently less members of a category than others, but in terms of explaining the internal category structure.

The vertical dimension has taxonomical structure composed of various levels of abstraction: the greater the inclusiveness, the higher the level. The criteria for defining particular level of abstractness/inclusiveness are based on the probabilistic concept of cue validity (Rosch *et al.* 1976) and Tversky's set theoretical framework[41] (Tversky 1977). What Rosch and colleagues have found is that this *basic level* of categorization has an important cognitive significance:

> Superordinate categories have lower total cue validity and lower category resemblance than do basic-level categories, because they have fewer common attributes; in fact, the category resemblance measure of items within the superordinate can even be negative due to the high ratio of distinctive to common features. Subordinate categories have lower total cue validity than do basic categories, because they also share most attributes with contrasting subordinate categories; in Tversky's terms, they tend to be combined because the weight of the added common features tend to exceed the weight of the distinctive features. That basic objects are categories at the level of abstraction that maximizes cue validity and maximizes category resemblance is another way of asserting that *basic objects are the categories that best mirror the correlational structure of the environment*. (Rosch 1978, p. 31; italics added)

The *basic level* is further characterized by following conditions (cf. Rosch 1978, Lakoff and Johnson 1999, Lassaline, Wisniewski and Medin 1992):

(1) it is the highest level for similarly perceived overall shapes, i.e. the highest level of abstraction at which we have mental image for the entire category. For example, there is no overall shape for furniture or animal, but there is for a chair or a cat;
(2) it is the highest level for actions for interacting with category members – you have an idea about handling a chair or a table, but not furniture;
(3) it is the highest level at which subjects are fastest at identifying category members and most of our knowledge is organized. As many have argued, the basic level is not 'static' in the sense that it is the same across different subject groups – it highly depends on the individual knowledge and expertise

[41] Cue validity measures the probability of particular attribute x (cue) as a predictor of category y increases with the frequency of x being associated with this category y and vice versa. Tversky's criteria for "category resemblance" is defined as a difference between the weighted sum of all common features within a category subtracted by the sum of their distinctive features (also those belonging to other categories). In general, cue validity gives a more precise measure of the effect of contrast categories than Tversky's approach, but differences between the two are not relevant to our topic.

(see e.g., studies by Atran 1989, Tanaka and Taylor 1991, Mervis, Johnson and Scott 1993, Johnson and Mervis 1997, Medin *et al.* 1997, Proffitt, Coley and Medin 2000, Johnson 2001, Bailenson *et al.* 2002, Ross *et al.* 2003, Augustin and Leder 2006, Holt and Beilock 2006, Ballester *et al.* 2008);

(4) it is arguably the first level where one would expect names to evolve and therefore a first level used by children.

Numerous empirical studies support basic-level categorization, e.g.: general studies by (Berlin 1972, Rosch 1978, Lassaline, Wisniewski and Medin 1992), cross-cultural studies (Berlin 1972, Rosch 1974), cross-domain studies (Tversky and Hemenway 1984), studies on object categorization (Jolicoeur, Gluck and Kosslyn 1984, Rosch *et al.* 1976, Murphy and Wisniewski 1989), studies on free-naming tasks (Rosch *et al.* 1976), studies on children's language development and reasoning (Anglin 1977, Karmiloff-Smith 1986, 1992, Gopnik and Meltzoff 1992, Gelman 1996, Jones and Smith 1993, Plaut and Karmiloff-Smith 1993, Sloutsky and Fisher 2004, Medin and Waxman 2007); but see (Mandler and Bauer 1988, Markman 1991, Markman and Wisniewski 1997) for contrasting view.

The key psychological aspects of human categorization, i.e. category, prototype and basic-level effects, play an important role in language and semantics. In what follows, I will show how the theory of *conceptual spaces* can account for these findings by employing the geometrical structure of conceptual spaces and the inherent notion of similarity.

13.2 Architecture of conceptual spaces

In conceptual spaces, the meanings of words are represented by their mappings onto conceptual structures. The meaning is built out of concepts which are represented in space by regions of quality dimensions. A conceptual space then is a geometrical structure based on a number of quality dimensions with inherent connection between distances and similarity judgments.

13.2.1 Quality dimensions and similarity

Gärdenfors (1988, 1991, 1996, 1997, 2000) argues that the fundamental role of *quality dimensions* is to build up the domains needed for representing concepts. In conceptual spaces, the dimensions are the basic structural elements that represent

various "qualities" or *properties* of objects in different domains and specify *relations* among them; some examples are color, pitch, temperature, weight, spatial dimensions of *height*, *width* and *depth* etc.). The dimensions differ in their level of abstraction and kind. Some of the quality dimensions are innate and part of our perceptual-motor system (such as color, taste, smell, pitch and space), some are discrete, some are learned, and still others can be culturally dependent or introduced by science. Furthermore, some of the quality dimensions are psychological, abstract, and have non-sensory qualities. Others, such as color, pitch, temperature, weight and spatial dimensions are all sensory dimensions that are easily quantifiable. For example, *time* and *weight* are one dimensional structures, former is isomorphic to the line of real numbers, the latter to the line of non-negative numbers. More interesting are quality domains in color perception, especially since color theory involves both physical and psychological dimensions.

In conceptual spaces, each quality dimension has a certain geometrical structure, since it needs to satisfy particular structural constraints. In many cases, the structure of quality dimensions (e.g. sensory and physical) is metrical, meaning we can talk about distances along the dimensions. Since properties and objects are intimately tied to quality dimensions, the distances between representations (e.g. of particular instances of objects that are represented as points in space) offer similarity measures. Similarity becomes a function of distance in conceptual space; in psychological studies of categorization, the similarity is defined as exponentially decaying function of distance (Shepard 1987, Hahn and Chater 1997). And, as Gärdefnors (2000) points out,

> [t]here is a tight connection between *distances* in a conceptual space and *similarity* judgments: the smaller the distances between the representations of the two objects, the more similar they are. In this way, the similarity of two objects can be defined via the distance between their representing points in the space. Consequently, conceptual spaces provide us with a natural way of representing similarities (p. 5; italics added).

13.2.2 Convex regions and Voronoi tessellation

Conceptual spaces formulate new criteria of how properties and concepts are to be represented:

> CRITERION P A natural property is a convex region of a domain in a conceptual space (Gärdenfors 2000, p. 71).
>
> CRITERION C A natural concept is represented as a set of regions in a number of domains together with an assignment of salience weights to the domains and information about how the regions of different domains are correlated (Gärdenfors 2000, p. 105).

These criteria are based on the two essential notions given by geometrical or topological structure and its intimate connection to the notion of similarity: *connectedness* and *convexity*. In their most characteristic form, conceptual spaces are built up from convex regions. A region C in conceptual space S is *convex* if, for all points x and y in C, all points between x and y are also in C. Convexity in conceptual space is generated with Voronoi tessellation. Voronoi rule uses the prototypes as centers to define boundaries of individual regions and tessellate the space. What Voronoi rule does is partition the space with n points into convex regions such that each region contains exactly one generating point (a prototype) and every point in a given region is closer to its prototype than to any other. The convexity of conceptual spaces is thus directly expressed in representation of object's properties: in a region within which two objects sharing property P are represented at points x1 and x2, any objects between these two points will also share property P. Moreover, within the conceptual spaces the similarity of the objects can be measured according to their position in space as well as to the "center of gravity" of the individual region they are part of. The delimitation of properties in conceptual spaces comes naturally and is intimately related to the prototype theory: the criterion P accounts for both, asymmetries among category members in cases of graded membership and asymmetric structures within categories in cases where properties have distinct boundaries (Rosch 1975, Lakoff 1987). Thus, the tight connection between notion of similarity, centrality and property in criterion P carries psychological validity: "when natural properties are defined as convex regions of a conceptual space, prototype effects are indeed to be expected." (Gärdenfors 2000, p. 86)

Are regions in conceptual space always convex? It is important to note that convexity depends on the underlying metric space. Different rules generate different equidistances or 'betweenness relations'. Most commonly used Euclidean metric follows Pythagorean theorem and gives a constant distance regardless the rotation,

i.e. it is invariant to the rotation of the axis (in formula below, the exponent r = 2). In city-block metrics, as the name implies, the distance is measured by the sum of adjacent sides of the block (imagine a space filled with buildings; here, r = 1). They can be formulated as instances of Minkowski metric:

$$d(x,y) = \left[\sum_{i=1}^{n}(x_i - y_i)\right]^{\frac{1}{r}} \quad (1)$$

The shape of the regions in conceptual space changes accordingly: under Euclidean metric the regions are convex, whereas city-block metric is star-shaped (under the city-block metric not all areas are convex).

The selection of metric space is essentially an empirical question, since the choice of metrics reflects different assumptions about the psychological dimensions underlying the conceptual space. In general, the Euclidean metric is more appropriate for integral dimensions where we cannot selectively analyze individual dimensions, a good example are color dimensions of saturation and brightness. The city-block metric, on the other hand, seems a better solution for separable dimensions, such as color and size (see Gärdenfors 2000, Garner 1974, 1978, Goldstone 1998). In our case (the construction of computer model for generating conceptual spaces, presented in Part IV), the choice will be Euclidean metric, both because there are no readily available natural metric structures in the analysis of text corpora (and consequently no possibility to initially identify whether dimensions are discrete or separable), and because of specific statistical and probabilistic methods used in domain identification (i.e. in the analysis of latent topic structures within the corpora).

There are further arguments for promoting convexity in conceptual spaces. There is strong empirical evidence for convexity, e.g. universal convexity of natural color categories (shown in studies of color perception and focal colors by Berlin and Kay (1969), Rosch (1975, 1978), Sivik and Taft (1994)), in universal properties of color terms in natural languages (Berlin and Kay 1969, Jäger and van Rooij 2007), or in preference for convex meanings that are, according to (Jäger 2007), "the result of some process of (cultural) evolution" (p. 552). Similarly, Regier and Kay (2009) argue that language influences color perception and categorization, and further claim

that color naming across languages is shaped by both universal and language-specific forces. Also, the intimate link between criterion P and prototype theory makes many properties perceptually grounded. For example, evidence for psychological reality of focal colors (Rosch 1975, 1978) shows that "many fundamental quality dimensions are determined by our perceptual mechanisms, and in conceptual spaces there is a direct link between properties described as regions of such dimensions and perceptions" (Gärdenfors 2000, p. 77).

On a more speculative note, preferring convex regions instead of oddly shaped metrics could be attributed to the "*principle of cognitive economy*; handling convex sets puts less strain on learning, on your memory, and on your processing capacities than working with arbitrarily shaped regions" (Gärdenfors 2000, p. 70).

In Part IV, we answer some practical questions. First, the computer model for generating conceptual spaces *SpaceWalk* will be introduced, along with different methods used for dimensionality reduction and identification. The aim is to empirically test the functionality of these methods in constructing conceptual spaces. We close overall discussion by investigating some of the cognitive underpinnings inherent in similarity-space and probabilistic approach, possible future applications for conceptual spaces, as well as challenges that probabilistic approach brings to cognitive science.

PART IV: *SpaceWalk*: a computational model for conceptual spaces

Section 9: Methods

14 Introduction

This chapter discusses the construction and functionality of *SpaceWalk*: a computer model for representing semantics of conceptual spaces. The overall aim of *SpaceWalk* is to propose the basic architecture for modeling lexical semantics and to reflect upon, especially in terms of functionality, the interpretational value that conceptual spaces bring to the discussion of semantics. Unlike traditional symbolic and connectionist models, *SpaceWalk* uses *similarity-space* and *probabilistic* methods for measuring semantic similarity and generating the dimensions needed for construction of conceptual spaces. In what follows, these methods are being compared and evaluated.

15 Methods for dimension identification

The methods for dimension identification used in *SpaceWalk* are able to reveal latent semantic structure and similarity by employing different statistical and probabilistic learning mechanisms on natural language corpora. Our primary focus here is on lexical semantics, i.e. on the semantics of individual word meanings, rather than sentences. Traditional statistical methods, such as Latent Semantic Analysis (LSA; Landauer and Dumais 1997, Landauer, Foltz and Laham 1998, Landauer *et al.* 2006), essentially rely on vector calculations in high-dimensional semantic space. Here, similarity measures are defined by distances between word-vectors in the semantic space, hence the name *similarity-space models* or *semantic vector models*. These methods are related to connectionist research on natural language processing (McClelland and Kawamoto 1986, St. John and McClelland 1990, Miikkulainen 1993, Elman 1990, Elman *et al.* 1996, McClelland and Elman 1986, Smolensky 1990) and represent a bottom-up approach. The general idea is, semantics can be generated from language statistics by relying on simple approach of treating text

corpora as a 'bag of words', disregarding any sequential information, i.e. positioning of words in a sentence. Consequently, the grammatical structure is not preserved. This, again, is in stark contrast to symbolic approach that focuses primarily on compositionality and therefore on syntax and semantics of whole sentences.

As alternative, top-down *probabilistic* models of cognition have been proposed as a more advanced and effective approach to measuring semantic similarity. For example, *probabilistic topic models* (Hofmann 1999, 2001, Blei, Ng and Jordan 2003, Steyvers and Griffiths 2006, Blei and Lafferty 2009) are generative models that use probability distribution to identify the topics and gist of the collection. These models are based on computing probabilistic inference, encapsulated in the Bayes theorem. Bayes theorem aims to solve an inductive problem – a situation, where we cannot unambiguously identify the generating process from the observed data. The general premise of the Bayes theorem is going beyond the observed data to evaluate the probability of different hypotheses or assumptions about generating process, while maintaining uncertainty (Griffiths *et al.* 2010, p. 358). It is a formal characterization of a problem space, using *prior* and *posterior probability* to solve an inductive problem. As an example, consider a set of hypotheses H. First, we define *prior probability p(h)*, which reflects a probability distribution over each hypothesis $h \in H$ independent from the data d. *Prior probability* reflects human factor, i.e. one's beliefs or inductive biases regarding hypotheses H. Next, we define the *likelihood* $p(d|h)$, indicating the probability of each hypothesis h to be true in light of the information from data d., with the sum in the denominator ensuring the outcome sums to 1. The outcome of the Bayes theorem is the *posterior probability* $p(h|d)$,

$$p(h|d) = \frac{p(d|h)p(h)}{\Sigma_{h \in H}\, p(d|h)p(h)} \qquad (2)$$

a probability distribution reflecting the degree of one's belief in individual hypothesis based on the additional information gathered from the observed data. In probabilistic topic models the semantic ambiguity is represented through uncertainty over topics, thus we can discover hidden thematic structure of our text corpora. The main difference between the *similarity-space* and *probabilistic* approach is that *probabilistic topic models*, while still carrying many of the key statistical

assumptions behind LSA (e.g. dimensionality reduction), can identify and preserve interpretable topic structure rather than just an opaque semantic space of word associations.

In what follows, *similarity-space* and *probabilistic* methods will be compared and later evaluated in *SpaceWalk*. These methods are commonly used in areas of information retrieval, natural language processing, machine learning etc. In *SpaceWalk*, these methods are being used to compute the similarity relations between words and documents, their topical distributions, and the dimensions needed for construal of semantic representations. The latter are formalized by employing conceptual spaces. Conceptual spaces carry both explanatory and functional role in modeling lexical semantics, and also serve as the main criteria for the evaluation of the above-mentioned methods. For example, conceptual spaces put additional structural constraints on high-dimensional vector representations generated by the *similarity-space* and *probabilistic* models and can, to an extent, mitigate some compositional issues characteristic of traditional connectionist models. As we shall see, there are important differences between methods themselves. For example, unlike *similarity-space* approach, probabilistic approach can account for basic properties of natural language semantics, such as synonymy and polysemy. It does this by identifying the latent conceptual structure via probability distributions over topics. By employing conceptual spaces as an additional structural constraint (based on criterion P together with the notion of similarity and convexity), *SpaceWalk* can account for prototype theory, graded membership and asymmetries within and across concepts and categories.

15.1 Latent Semantic Analysis

Latent Semantic Analysis or LSA (Landauer and Dumais 1997) is one of the most known methods used in natural language processing. It is a computational technique based on associative approach to word meanings. Using large text corpora, LSA generates high-dimensional similarity-space representations of associations between words. By computing the distance between word vectors, LSA extracts the 'meaning' of individual word based on its proximity to other words in semantic space. The input to LSA is a word-by-document co-occurrence matrix, such as that shown in Figure 9.

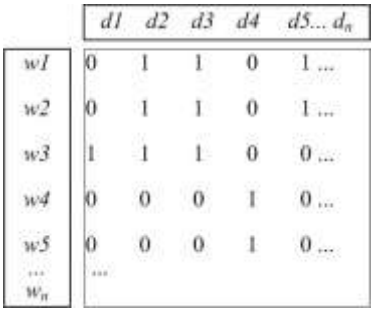


Figure 9: LSA: word-by-document co-occurrence matrix

The procedure to generate a semantic space in LSA goes as follows. LSA begins with a word-by-document co-occurrence matrix representation of a text corpus. Each row represents a word and each column represents a document, and the entries indicate the *frequency* with which that word occurred in that document. An association function is applied to dampen the importance of each word proportionate to its *entropy* over documents, by weighing "each word-type occurrence directly by an estimate of its importance in the passage and inversely with the degree to which knowing that a word occurs provides information about which passage it appeared in (Landauer et al 1998, p. 276). In general, words that appear together frequently over the documents get assigned smaller values. Next, the Singular Value Decomposition (SVD; Berry, Dumais and Obrien 1995) is applied to the co-occurrence matrix. This is the main point where LSA model differs from probabilistic models, as is evident from graphical representation of LSA and LDA matrix factorizations shown Figure 10 and Figure 12 respectively. The original matrix is generally sparse and contains many empty values. The role of SVD is to reduce the original matrix's dimensionality to a lower *latent semantic space* (typically between 100 – 300 dimensions; Deerwester *et al.* 1990, Dumais 1995, Martin and Berry 2006) to retain most of essential features and consequently remove statistical noise. The reduced dimensionality comes with SVD factorization of the original matrix into three smaller matrices, U, D, and V (Figure 10).

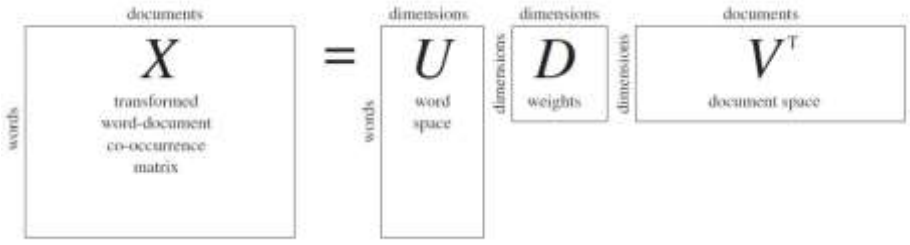


Figure 10: Graphical model of the matrix factorization in LSA

Each of these matrices has a different interpretation. U and V matrices provide orthonormal basis for a similarity-space where each word/document is represented as a point, with diagonal D matrix providing a set of weights for the dimensions of this space. The output from LSA is a *latent semantic space*. By re-multiplying these matrices we can get an approximation to the original matrix in a lower dimensional spatial representation. It is important to note that the number of dimension retained in the process of SVD is an empirical issue, but the overall aim is to retain the essence of the original matrix and emphasize the latent correlations among words. According to Landauer and Dumais (1997), the measure of similarity is computed as cosine (of the angle) between two word vectors in the semantic space. This is an effective measure of the semantic association between words across the dimensions, where cosine of 1 gives strongest similarity, while cosine of 0 (or negative value) shows dissimilarity or no similarity (for technical details see (Landauer *et al.* 2006)).

How does this procedure result in semantic similarity? Frequently, words that occur together have often no semantic similarity. However, LSA does not use only the information about how often word1 and word2 occur together but also how often they occur with all the other words in the corpus. LSA looks at the entire pattern of co-occurrences to define the similarity. For example, Landauer and Dumais (1997) tested LSA for a synonym test on Test of English as a Foreign Language (TOEFL), which is used as a college admission test for nonnative speakers of English and American universities, and received impressive results[42]. The model showed 64.4% accuracy, which at the time was almost identical to the performance of a large

[42] The test corpus was taken from Touchstone Applied Science Associates (TASA) corpus. TASA corpus is a collection of educational texts used in U.S. curricula from ground school up to first year of college, covering areas of arts, health, history and culture, home economics, natural sciences, social studies etc.

sample of college applicants who took the test. Landauer and Dumais note that such score would allow admission for many American universities.

LSA's success may indicate that conceptual information is simply not that necessary to explaining meaning. Because the only input of the model are words, meaning could just as well be represented based on associative links to other words, rather than through the knowledge underlying those words – that is, conceptual knowledge. Could that be the case?

The general problem with similarity-space models such as LSA is that knowing what words are associated to one another does not specify what the meaning of the individual word is. As Murphy (2002) points out, one cannot understand the meaning of a word only by its reference to other words: "If one only knows *dog* by its similarity to *cat* and *cow* and *bone* … and *cat* by its similarity to *dog* and *cow* and *bone* … and so on, one is caught in a circle of similar words" (p. 429). The point being, one needs primary conceptual and categorical knowledge to be able to evaluate different associative relations. Further problem is that relations between words are extremely different and the overall word similarity generated by LSA does neither specify different meanings of these associations nor their gist[43]. To avoid ambiguity, we need to represent some context: words must be connected to our (conceptual) knowledge, not just other words.

And these are exactly the things that conceptual spaces approach does well: having the concept explains why particular words are related. Since concepts are mental entities – our non-linguistic representation of the world, by connecting words to conceptual structure, we can explain how people can connect words to objects and events in the world. Thus, by hooking up words to concepts, we can break out of the circle of words connected to words and tie language to perception and action.

The biggest problem for similarity-space models is that the spatial representations of similarities between words are relatively unstructured, lacking the conceptual information needed for representing various semantic relations. In what follows, I present two alternatives to LSA. Both were inspired by LSA, but are based on the notion of probability and make different statistical assumptions.

[43] As I will argue later, LSA cannot account for polysemy, one of the most common effects of word meanings. Moreover, LSA cannot express topical variety implicit in text corpora.

15.2 Probabilistic Latent Semantic Analysis

As the name implies, Probabilistic Latent Semantic Analysis (pLSA; Hofmann 1999) is closely related to standard LSA approach, with one essential difference. pLSA is a *topic model*. Instead of using SVD to compute word and document similarity in the semantic space, pLSA is a *generative* data model (Hofmann 2001) and uses simple probabilistic procedure to generate documents.

The general idea behind the *topic models* is that

> ... documents are mixtures of topics, where a topic is a probability distribution over words. A topic model is a generative model for documents: it specifies a simple probabilistic procedure by which documents can be generated. To make a new document, one chooses a distribution over topics. Then, for each word in that document, one chooses a topic at random according to this distribution, and draws a word from that topic. (Steyvers and Griffiths 2006, p. 427)

The upshot of such an approach is that topics are individually interpretable and that words can be part of more than one topic. Thus, the topic model can account for context.

A more rigorous treatment of statistical assumptions behind each approach has been given in (Hofmann 1999, 2001, Griffiths and Steyvers 2002, Blei et al. 2003, Steyvers and Griffiths 2006). Here, we discuss the general notation for generative topic model,

$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)\, P(z_i = j) \qquad (3)$$

where T is the number of topics, P(z) is the distribution over topics z in a document and $P(w \mid z)$ is the probability distribution over word w given topic z. The first part of notation $P(w_i|z_i = j)$ refers to distribution over words for topic j, whereas $P(z_i = j)$ refers to a distribution over topics for document d. Accordingly, the first part indicates which words are important for which topic, whereas the second part of notation indicates which topics are important for a particular document.

Compared to Hofmann's pLSA model (1999, 2001), one important difference is evident. pLSA is limited in its definition concerning documents:

$$P(d, w_i) = p(d) \sum_z P(w_i|z)P(z|d) \qquad (4)$$

According to (Blei et al. 2003, Steyvers and Griffiths 2006), pLSA is not a complete generative model. While topic distributions over words are efficiently explained, pLSA provides no explicit probabilistic model at the level of documents: it can only learn the topic mixtures $P(z|d)$ for the documents in existing training set, not for new, previously untrained documents. We cannot assign probability to a document outside of the training set. Additional problem is that the number of parameters (in pLSA, these are being treated individually) grows linearly with the size of the corpus, which results in the overfitting of the model (for a detailed analysis, see Blei et al. 2003, chapter 7.1).

15.3 Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA; Blei et al. 2003) is a basic generative probabilistic topic model. LDA inherits the intuition of a general topic model: the documents exhibit multiple topics. The central question for LDA and topic modeling in general becomes, what is the hidden structure behind these documents? It tries to uncover the blend of latent topics as distributions over documents and words; the topics and topic distributions are hidden structure. Depending on a topic, words have different levels of probability; for example, word 'reason' will have high probability on topic about 'philosophy', but should get low probability value on topic about 'vegetables'. Thus, each document exhibits topics with different proportions and each word is drawn from one of the topics.

> This is the distinguishing characteristic of latent Dirichlet allocation – all the documents in the collection share the same set of topics, but each document exhibits those topics with different proportion. (Blei 2012, in print; [p. 4 in draft])

Formally, the LDA goes as follows (but see (Blei *et al.* 2003, Blei and Lafferty 2009, Blei 2012) for more detail). LDA is a generative probabilistic model. The data are part of generative process which defines a *joint probability distribution* over observed and hidden random variables, with the former being words and the latter being the topic structure. LDA uses joint distribution to compute the *posterior distribution* of the hidden variables given the documents.

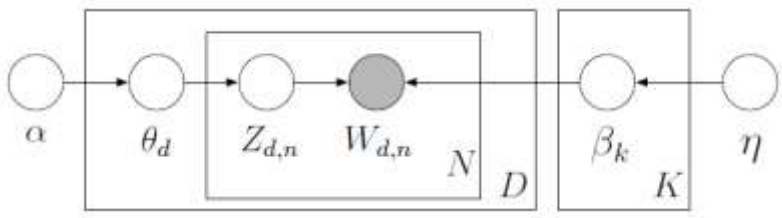
A more formal definition (Blei 2012) is expressed in the following notation,

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D,}\, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n}|\theta_d)\, p(w_{d,n}|\beta_{1:K}, z_{d,n})\right) \quad (5)$$

Suppose topics are defined as $\beta_{1:K}$ and each topic (β_k) is a distribution over words. Then, "[t]he topic proportions for the *d*th document are θ_d, where $\theta_{d,k}$ is the topic proportion for topic *k* in document *d* ... The topic assignments for the *d*th document are z_d, where $z_{d,n}$ is the topic assignment for the *n*th word in document *d* ... Finally, the observed words for document *d* are w_d, where $w_{d,n}$ is the *n*th word in document *d*, which is an element from the fixed vocabulary. ... With this notation, the generative process for LDA corresponds to the following joint distribution of the hidden and observed variables" (Blei 2012, p. 6)

The notation above specifies a number of dependencies: the topic assignment $z_{d,n}$ depends on per-document topic proportions θ_d; the observed word $w_{d,n}$ depends on the topic assignment $z_{d,n}$ and *all* of the topics $\beta_{1:K}$.

In LDA, the number of topics should be specified before any computation occurs (as we shall see later in our tests). Based on per-document topic distribution, the statistical inference algorithm called a Dirichlet distribution, LDA computes the hidden structure that generated the documents in corpus. The graphical model (Figure 11 below) shows the structure of LDA.



> Figure 11: The graphical model for LDA's hidden and observed variables (adapted from Blei *et al.* (2003, p. 997)). The words of the document $w_{d,n}$ are the only observed variable (shaded node). The rectangles denote replication for each level (*N* for words and *D* for documents).

In Figure 11, the three levels are represented by rectangles which denote replication: *corpus* level (variables α and β, sampled once per document), *document* level (variables θ_d) and *word* level (variables $z_{d,n}$ *and* $w_{d,n}$ are sampled once for each word in a document).

15.4 Comparing LSA and topic models (Part 1)

All three models use a word-document co-occurrence matrix as an input, but differ in statistical assumptions (compare Figures 10 and 12). All use the *dimensionality reduction* to provide interpretable dimensions and the gist of the corpora, but the methods and the outcome are very different.

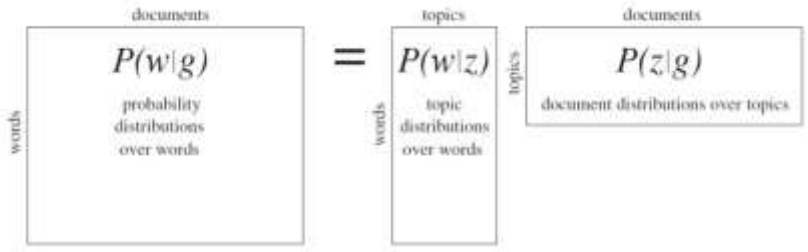


Figure 12: Matrix factorization in the topic model

For one, there is a difference between how topic distributions over documents are treated by pLSA and LDA. Whereas in pLSA the distribution of topics over documents is not intepretable, in LDA it is made explicit. As Blei *et al.* (2003) point out:

> LDA overcomes both of these problems by treating the topic mixture weights as a *k*-parameter hidden *random variable* rather than a large set of individual parameters which are explicitly linked to the training set. ... LDA is a well-defined generative model and generalizes easily to new documents. Furthermore, the *k*+*kV* parameters in a *k*-topic LDA model do not grow with the size of the training corpus. (p. 1001).

One essential advantage of topic models over similarity-space models is that statistical inference is flexible and can generate structured representations. LSA, on

the other hand, has several issues (e.g., cf. Hofmann 1999, 2001, Blei *et al.* 2003, Steyvers and Griffiths 2006, Griffiths *et al.* 2007): it cannot hypothesize about the underlying topics, the semantic space generated by SVD does not provide enough structural information, and it cannot account for polysemy. In generative topic model, the probability distribution of topics over words and documents over topics accounts naturally both for multiple lexical meanings (different word senses), as well as for topic structure implicit within documents. As Griffiths *et al.* (2007) show, this simple principle gives sufficient structure to "capture some of the qualitative features" and semantics of natural language. This, together with the ability to easily generalize to new documents, is the key advantage of probabilistic approach.

Further analysis and comparison of similarity-space and topic models is given in the following chapters and in Part 2 in the Discussion. There, I will focus on LSA and LDA, since they are most characteristic examples of respective approach. First, I present the creation and exploration of conceptual spaces in *SpaceWalk*.

Section 10: Creating and exploring conceptual spaces in SpaceWalk

16 Creating conceptual spaces

Following is the presentation of the corpus and additional methods and procedures used to compute conceptual space. *SpaceWalk's* architecture is reconstructed in MATLAB programming environment[44], and the order of presentation below follows sequences of computation.

16.1 The corpus

Our test corpus is a collection of articles and books spanning from 1950's to date, including most of the thesis' literature. The collection covers diverse topics related to cognitive science, and connects domains of cognitive science with philosophy, psychology, linguistics, computer science, artificial intelligence and neuroscience. The corpus contains 248 items altogether, of which there are 194 articles and 54 books with the average vocabulary size of 51,789 terms per document before the normalization and 14344 terms/document and 39,862 unique terms (for the whole corpus) after normalization (see Table 1)[45].

16.2 Methods and procedures

1. Text parsing: make TMG

   Text to Matrix Generator (TMG; Zeimpekis and Gallopoulos 2005) is being used to parse the corpus and generate word-document matrix. The TMG also returns word and document-titles dictionary for the collection, the vectors of global weights, and the normalization factor for each document. The stemming is used (words, smaller than 2 tokens and larger than 35 tokens are ignored) together with a standard stop-word list (functional words that algorithm should ignore, e.g. 'a', 'the', etc.). Additionally, the threshold for the minimum and maximum local (1; inf.) and global (3; inf.) frequencies is set to filter out high-frequency

[44] Matlab is commonly used for algorithm development, data analysis, visualization, and numerical computation, and offers various tools for these domains. Most of the tools presented here are part of Matlab's natural language processing toolbox.

[45] Note that this corpus is much smaller in size than TASA corpus. For comparison, the TASA corpus has a vocabulary of approximately 10 million words (92,409 word types).

words with low semantic content. Table 1 shows the results for text normalization.

Table 1: make TMG

Results:
Number of documents = 248
Number of terms = 39863
Average number of terms per document (before the normalization) = 51791.7
Average number of indexing terms per document = 26883.2
Sparsity = 7.89699%

Removed 128 stopwords...
Removed 3291 terms using the term-length thresholds...
Removed 115817 terms using the global thresholds...
Removed 0 elements using the local thresholds...
Removed 0 empty terms...
Removed 0 empty documents...

2. Run LSA, pLSA or LDA: set the number of topics, dimensions and maximum number of iterations

   Setting the number of topics is a trial and error process – finding an optimal number of topics depends on the size of the corpus and the algorithm in use. Furthermore, it is important not to confuse the probability distribution of topics in a topic model with dimensionality reduction used by LSA. The topic model is a generative model, whereas LSA is based on SVD. Depending on the size of the corpus, the optimal number of topics for LDA is characteristically low (in our case, between 10 and 50), whereas the optimal number of dimensions for dimensionality reduction in SVD is much higher, generally between 100 and 300 (see Landauer and Dumais 1997; in *SpaceWalk*, the SVD is set to 300 dimensions). The number-of-topics setting is straightforward: in case of topic models it defines the number of topics to start a generative process, whereas in case of LSA, it picks out first *n* most salient dimensions generated by SVD.

3. Make Self Organizing Map (SOM)

   We use SOM Toolbox (Alhoniemi, Himberg and Vesanto 2002) to calculate SOM. The SOM (Kohonen 1995) is a tool for visualizing high-dimensional data. In essence, SOM is an artificial neural network that produces a low-dimensional (typically two-dimensional) representation of the input space of the training samples, called a map. The goal is to discover the underlying structure of the data. SOM is a topology-preserving map, since it preserves the topological

properties of the input space, i.e. neighborhood relations between nodes. This makes SOMs useful for visualizing low-dimensional views of high-dimensional data, akin to multidimensional scaling (MDS). A common 2-dimensional representation of SOM is U-Matrix (Unified Distance Matrix). U-Matrix uses the codebook vector, depicting the relations between the neighboring neurons in a color or gray scale image. The light colors depict closely spaced nodes, and vice versa. Thus, U-Matrix gives a general insight into the structure of high-dimensional space: the group of light colors represents clusters while dark regions represent boundaries (Figure 13). We can set two general parameters to calculate SOM: size and input matrix. Setting the size of the SOM depends on the size of the corpus and number of topics. It's a trial and error process of achieving optimal solution. Additionally, the input data to compute SOM is chosen between normalized Pz_d matrix, which gives probability of documents over topics, or normalized Pw_z matrix, which gives the probability of words in topics.

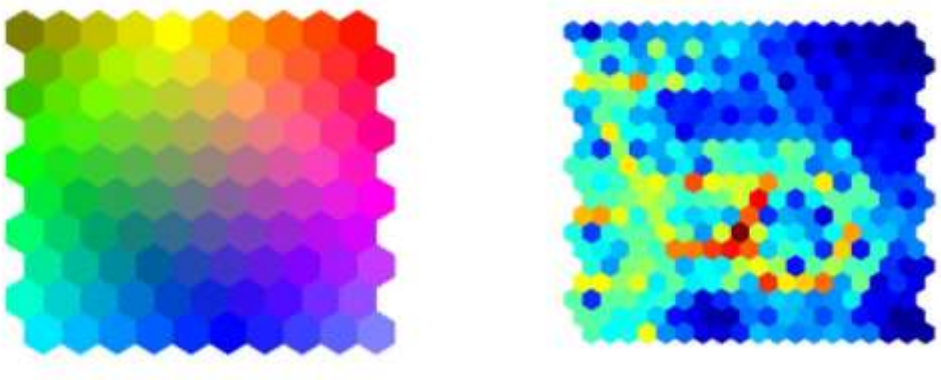


Figure 13: U-Matrix with color code

4. Project SOM: use PCA or Sammon's projection
   a. PCA projection

      Principal Component Analysis (PCA) is a statistical technique for dimensionality reduction that converts a set of observations into a set of linearly uncorrelated variables or principal components. Consequently, in PCA, variance decreases linearly: first principal component has the largest

possible variance, with each succeeding principal component having lower variance. The goal of PCA is to maximize the variance in the data.

a. Sammon's projection

While PCA does not preserve the structure of the dataset, Sammon's mapping (Sammon 1969) tries to preserve the distances and topological structure of dataset. A general comparison between PCA and Sammon's projection is given by (Henderson 1997).

5. Make plots
   a. Plot U-Matrix
   b. Plot 3D SOM Mesh

      Projection of SOM onto 3D-Mesh preserves structural relations between dimensions. Additionally, dimensions are color coded for easier identification.

   c. Plot centers of dimensions (topics) and most salient words or documents for each dimension

      Words and documents with highest values are projected onto individual dimension. The number of words/documents to plot can be set. To reflect context and multiple meanings, the words/documents can be plotted on more than one dimension.

6. Make Voronoi tessellation:

   In *SpaceWalk*, the partitioning of the space into convex regions is based on a set of prototypes, i.e. most salient vectors of words or documents that are most characteristic of individual domain. These are generated in advance by one of the natural language processing methods (LSA, pLSA or LDA). SOM and Voronoi rule are then used to exploit the geometrical properties of multi-dimensional space. Since these properties are generated by the topical distribution over words and documents[46], the space itself is naturally conceptual.

[46] This holds for generative models, such as LDA. I will discuss the shortcomings of LSA in later chapters.

When examining the space, it is important to compare Voronoi tessellation with 3D-Mesh projection. Because of the compression (and consequently transposition) of multiple dimensions onto 2D-Voronoi space, parts of 3D structure, such as proportions and distances between dimensions, cannot be faithfully preserved. Hence, it is common for two relatively distant dimensions (in 3D-Mesh) to neighbor each other in 2D-Voronoi projection and vice versa. Both projections should be taken as two different, but complementary representations of conceptual space. Whereas 3D-Mesh is a useful tool when we need to examine the underlying structure of multiple dimensions, Voronoi partitioning is best for analyzing graded membership, prototypes and category structure within and across individual regions of conceptual space.

We have already discussed different factors involved in modeling meaning through conceptual space. In next chapter, I focus on those that can be explored through *SpaceWalk*: prototype effects, probability distribution and graded membership (i.e., graded categorization), concepts and underlying quality dimensions.

17 Exploring conceptual spaces

This part focuses on exploration of conceptual spaces using LDA topic model. The data (top words and documents per topic) together with basic projections of the two alternative methods (LSA and pLSA) are in Appendix. All models (here and in Appendix) are trained on the corpus described in Chapter 16.1 and follow the calculation procedure described in Chapter 16.2. Results for LDA are presented in Table 2 below (see also Tables 3 and 4 in Appendix).

Table 2: Run LDA and SOM

run LDA LDA variational inference started with 248 documents and vocabulary of size 39863 using 10 topics and 0.000100 minimal relative change. Total number of words in data: 3.557313e+006 (14344.00 on average per document). **make SOM** Data to use: document data (normalized Pz_d) SOM size: [10 x 10] Projection: Sammon's projection (to preserve the distances and topological structure of dataset) **Results:** Final quantization error: 0.738 Final topographic error: 0.036 computing mutual distances: 100 iterations

Calculation procedure is the optimization, initialized randomly, and multiple calculations can yield slightly different results: each computation starts from a different position, picking out different aspects of conceptual space and quality dimensions involved. Thus, different computations yield different levels of analyticity. As consequence, in absolute terms, projections of conceptual space change accordingly, whereas relative relations among individual topics and internal structure (concepts and quality dimensions) of particular region generally remain consistent and coherent. For example, topics such as 'cognitive semantics', 'conceptual spaces' and 'quality dimensions' involve semantically similar or related sets of quality dimensions and hence consistently form a common region, whereas topics involving unrelated or contrasting sets of dimensions, such as 'conceptual spaces' and 'probabilistic approach', belong to separate regions of conceptual space even after multiple runs of the algorithm (see Figure 14).

These effects are strongly related to the number of topics. As a rule, smaller the number, more abstract or general are the topics, larger the number, more segmentation and detail is available in projections of conceptual space. Topics form regions of conceptual space, and, as already noted, the optimal number of topics is a trial and error process[47]. Here, I've decided to present a general view of the corpus and topics involved (e.g., philosophy, cognitive psychology, neuroscience, cognitive science, etc.; see corresponding regions in Figures 14 and 15), hence limiting the number of topics to 10. For a more detailed and segmented view of conceptual space 30 topics would be more appropriate (see Table 6 in Appendix); in such case, 'cognitive semantics' and 'conceptual spaces' would generally become separate regions of conceptual space, with more refined sets of quality dimensions, but would still remain strongly related semantically.

17.1 Projections

Following are various projections of conceptual space using *SpaceWalk*. Let us explore semantic relations and uncover the latent structure of the corpus with

[47] For this corpus, choosing a large number of around 50 topics results in overfitting – noise from functional (meaningless) words and topics start to overlap (multiple topics with the same sets of dimensions).

projections of topics, most salient dimensions (top words/concepts) and documents based on their probability distributions. We will do this by:

a) comparing Voronoi and 3D-Mesh projections of conceptual space (Figures 14 and 15), and

b) graphing the probability distributions (Figures 16-19).

Figures 14 and 15 show LDA projection of conceptual space in 3D-Mesh and Voronoi respectively.

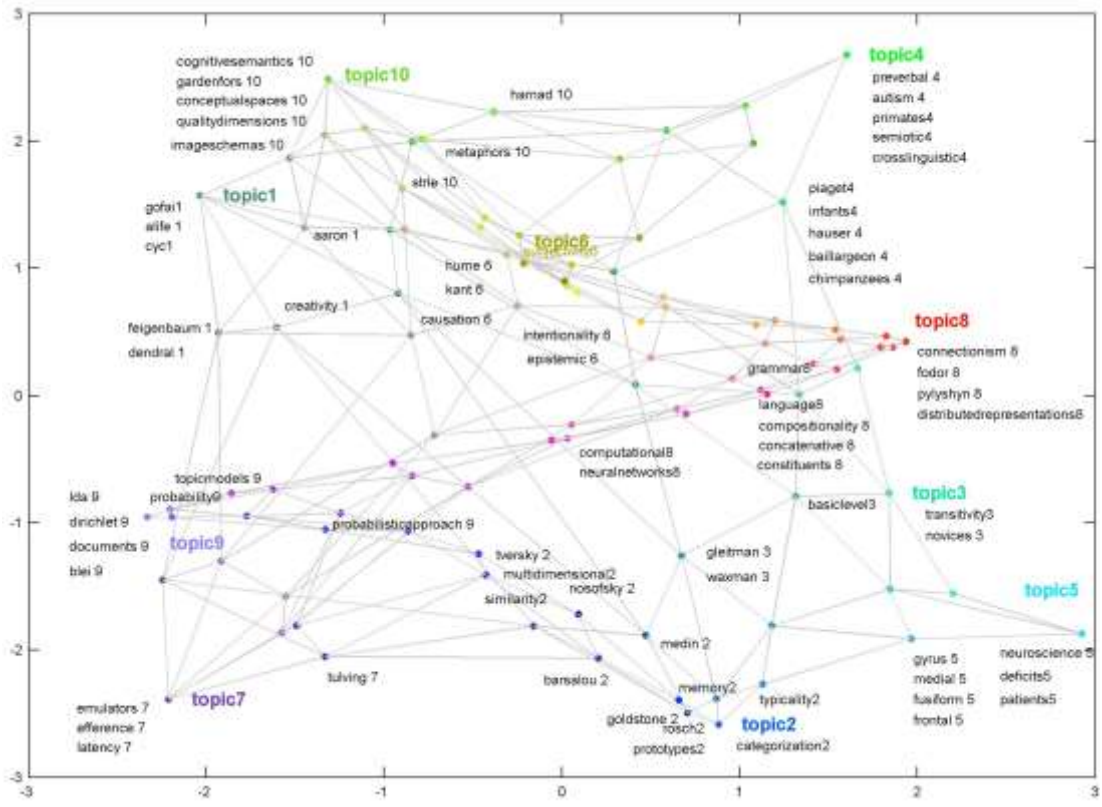


Figure 14: 3D-Mesh of conceptual space. Showing distribution of words over topics (Pw_z) with most salient words (top words) for each topic

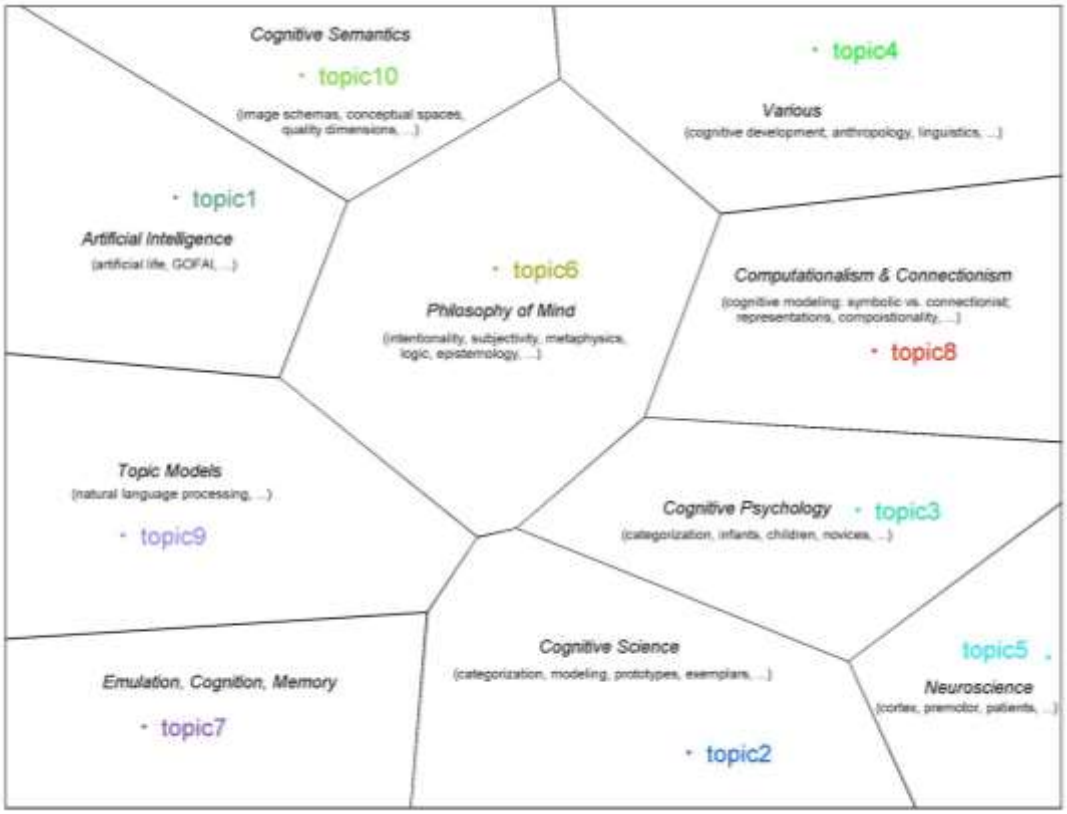


Figure 15: Voronoi tessellation of conceptual space

Figure 14 shows a 3D projection of topics and dimensions of conceptual space. The centers of individual regions are defined by a prototype (most salient member of individual topic) and later used for Voronoi tessellation of conceptual space (Figure 15). Nodes and edges of the 3D mesh represent shapes of underlying quality dimensions. Single or multiple quality dimensions can contribute to the formation of one or more regions of conceptual space. Each member of conceptual space (whether object, property or concept – these terms are being used interchangeably, depending on context) is plotted on a nearby node and gravitates towards the center of particular region based on the strength of its membership with some members being more prototypical than others. Furthermore, unlike properties (recall Criterion P) concepts share multiple dimensions and are not constrained by particular region of conceptual space. In *SpaceWalk*, concepts are represented by probability distribution over topics and hence, multiple regions of conceptual space (Criterion C). This enables a natural categorization of concepts into convex regions of conceptual space (Figure 15), a representation of graded membership for both, categories and concepts (Figures 16-18), as well as a differentiation of lexical meanings depending on context. For the latter, different regions of conceptual space activate different sets of quality

dimensions and thus different senses (different lexical meanings) with various degrees of semantic similarity. For example, in Figure 18 word 'neuron' strongly features in two different contexts: as the core component (primary cell type) of the nervous system (***topic5*** 'Neuroscience'), and as artificial neuron, the basic unit in an artificial neural network and the core component of connectionist modeling (***topic8*** 'Cognitive modeling: symbolic vs. connectionist').

Figures 16-19 present a series of probability distributions of topics over conceptual space[48]: distribution of topics (Figure 16), distribution of words/concepts over topics (Figure 17 and 18) and distribution of topics over documents (books and articles from the corpus; Figure 19). Tables 3-6 (in Appendix) list top words and documents for individual topic and the selection of documents for Figure 19 (see Table 5 in Appendix).

Each bar represents one topic in accordance with the prototype rule for discriminating conceptual space into individual regions (see Figure 15). Moreover, as prototype theory accounts for graded membership, each topic is to some extent related to other topics in conceptual space based on the probability distributions over underlying quality dimensions, and the strengths of these relations are visualized in Figures 16-19 by the lengths of bars. In effect, topics that share probability distributions over multiple quality dimensions (such as *topic2* 'Cognitive science' and *topic3* 'Cognitive psychology') are taken to be strongly related, whereas others, with a limited number of shared dimensions, are generally unrelated (e.g., compare *topic5* 'Neuroscience', *topic1* 'Classical Artificial Intelligence' and *topic9* 'Topic models'). What we get is a conceptual space of semantic relations based on the topic mixtures and probability distributions.

By comparing different aspects of conceptual space, we can get a feel for topic distributions, and thus for category structure of concepts, categories, and underlying quality dimensions. For example, Figure 19 shows probability distribution of topics over documents[49]. Consider the distribution of topics for this thesis, 'Strle-Semantics within', and compare Figure 19 (section 'Strle') with distributions in Figures 16-18. Among 10 general topics that represent main regions of conceptual space (Figure

[48] For purpose of clarity, different color codes are being used in bar charts (Figures 16-19) and other plots (U-Matrix, 3D-Mesh and Voronoi in Figures 13-15).
[49] Here, the probability distributions sum up to 1.

16), four topics prevail: *topic2* (cognitive secience), *topic8* (tradidional symbolic and connectionist approaches), *topic9* (topic models and natural language processing) and *topic10* (cognitive semantics, conceptual spaces). This shows where main focus of the thesis is regarding the general problem area of cognitive science: traditional models, cognitive semantics and methods for natural language processing. The activation of subregions of conceptual space (see Figures 16, 17 and 18) might show a more detailed analysis, also of different quality dimensions involved, and corresponds to the various topics covered in the thesis, as well as in thesis' literature. In Figure 17, the distribution of concepts corresponds to a more general view of the topics covered: 'similarity' and 'probability' are obviously related, but have different probability distributions. The former is strongly activated in area of categorization (*topic2*), whereas the latter is a part of topic modeling (*topic9*). Similarly, 'semantics' (Figure 17) is projected in the context of cognitive science (*topic2*), philosophy of mind (*topic6*), traditional symbolic and connectionist paradigms (*topic8*), natural language processing (*topic9*) and cognitive semantics (*topic10*), with strongest probability in *topic2*, *topic8* and *topic10*, corresponding to the focus of the thesis. There are also obvious cases of unrelated dimensions: whereas 'raam' is a purely connectionist system and belongs to *topic8*, 'lsa', 'plsa' and 'lda' are part of natural language processing approaches and belong to *topic9* (Figure 18). Nevertheless, as a rule, there are no strict borders and probability distribution naturally accounts for graded membership.

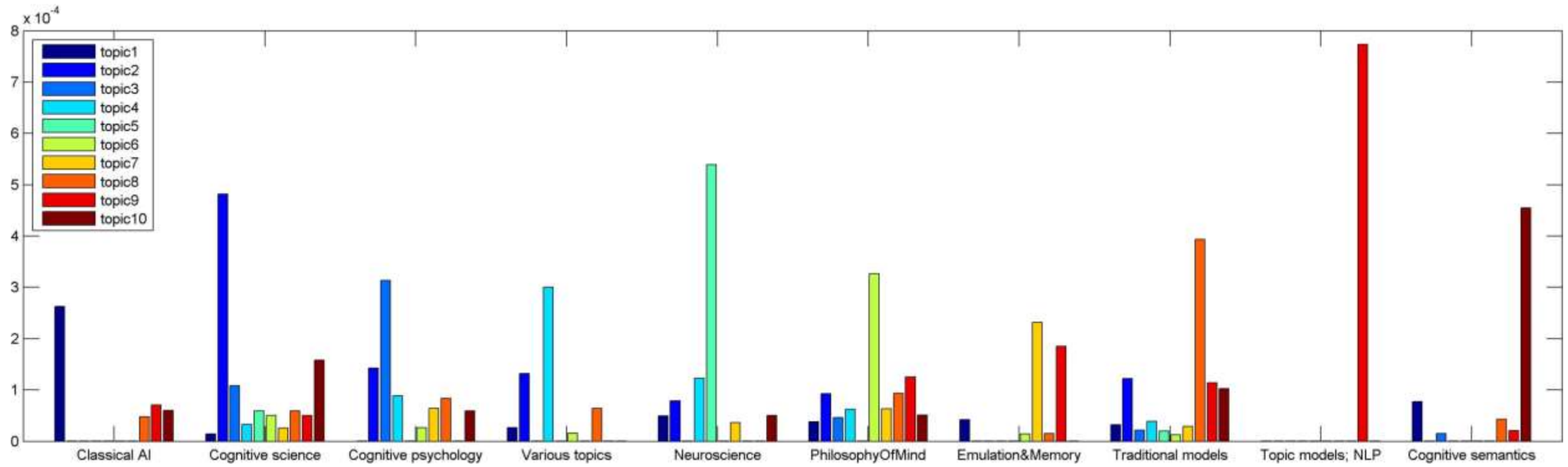


Figure 16: LDA distribution of 10 topics over conceptual space. Most salient member of individual topic is chosen for distribution.

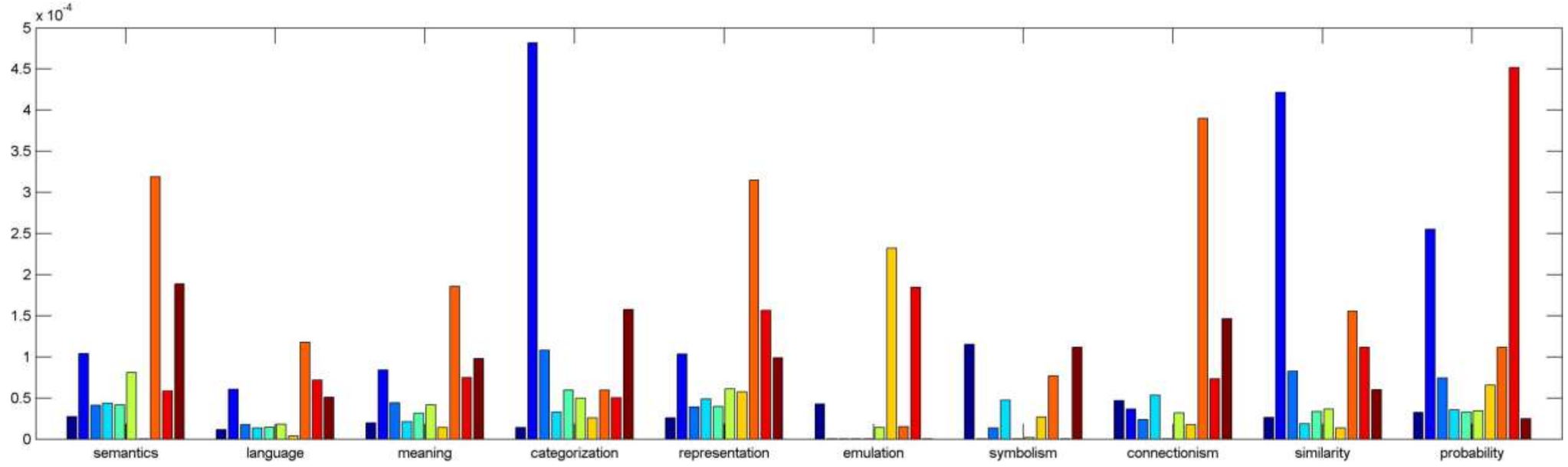


Figure 17: LDA distribution of words/concepts over topics (general)

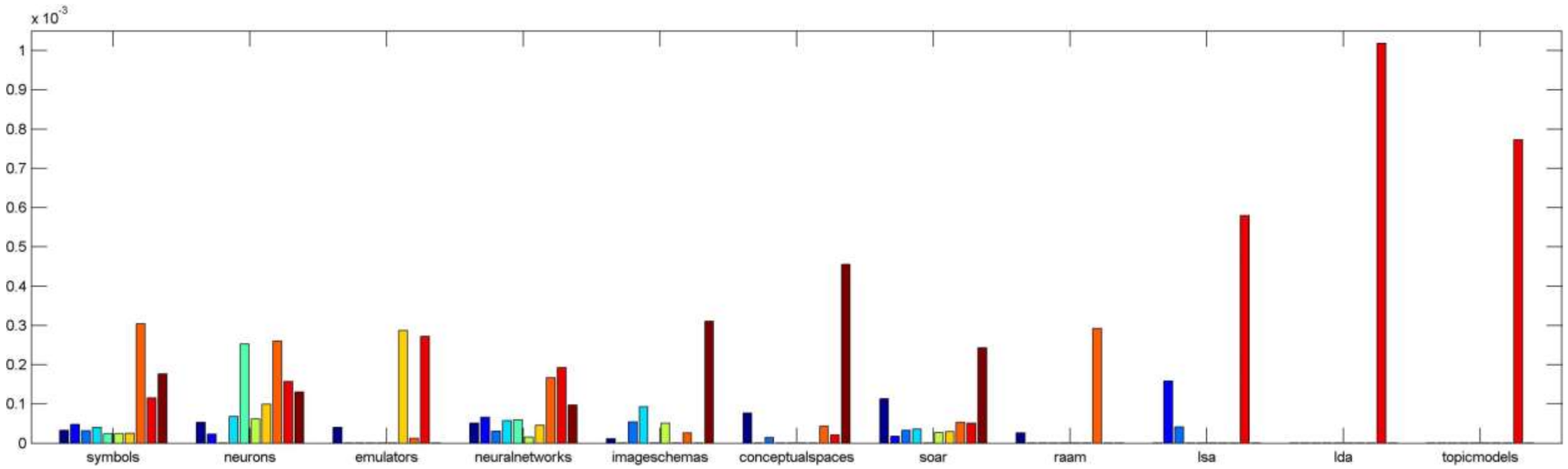


Figure 18: LDA distribution of words/concepts over topics (cognitive modeling)

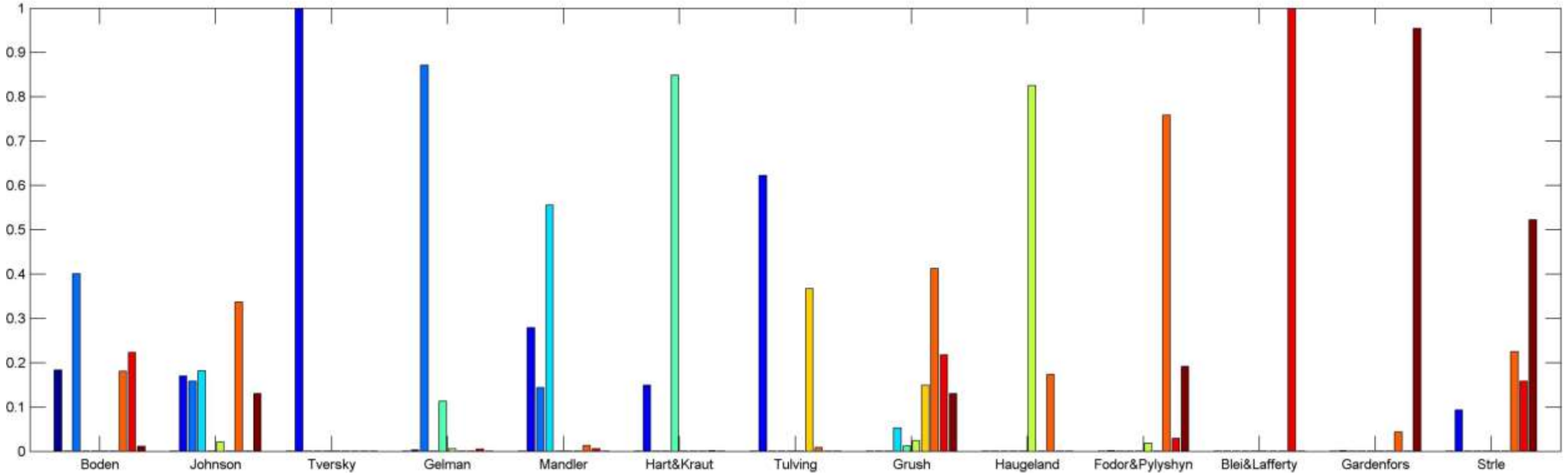


Figure 19: Probability distribution of topics over documents (here, documents are represented by authors' names, see *Table 6: LDA: List of document titles* in Appendix)

Section 11: Conclusion

18 Discussion

18.1 Comparing LSA and topic models (Part 2)

Previous chapters focused on the analysis of methods for natural language processing with the aim to answer the question which model is empirically and computationally more appropriate for generating semantic representations in conceptual spaces. One obvious outcome of the performed tests are different representational structures generated by similarity-space and probabilistic models. As shown, these representations significantly differ in their composition, with LSA generated space having in general more 'unbalanced' distribution compared to LDA[50] (compare Figures 14 and 15 with Figure 20 in Appendix). This difference is especially evident in the Voronoi tessellation of the semantic space, where salience of individual regions is highly disproportional (e.g. topics overlap) given the topical distributions of the corpus. This confirms the results from comparison studies, where (Hofmann 1999, 2001, Blei 2003, Landauer *et al.* 2006, Steyvers and Griffiths 2007, Griffiths *et al.* 2007) show that probabilistic models generally output more discriminative and hence interpretable semantic structures compared to similarity-space models.

In what follows, I shortly review some of the qualitative differences between both models by pointing out their underlying principles and the effects they have on the interpretation of semantics in natural languages. Here, similarity-space models face one of the more pressing criticisms: the classical criticism of similarity judgments in spatial models put forward by Tversky (Tversky 1977, Tversky and Gati 1978, 1982, Gati and Tversky 1984, Tversky & Hutchinson, 1986).

[50] By 'unbalanced', for want of a better name, I mean the regions of conceptual space generated by LSA are highly disproportional, with one region generally overruling. This is an effect of previously mentioned facts: a) the sum of the vector weights in LSA is not constrained by 1, and b) LSA cannot account for topic distribution and hence provides only a single gist of the corpora.

18.1.1 Similarity, probability and meaning

There are essential differences in the assumptions underlying similarity-space and probabilistic models of language and cognition.

First, LSA is not a generative model, but a connectionist network. It is a bottom-up approach based on SVD and uses dimensionality-reduced word association matrix as a measure for spatial representation of semantic similarity. Second, the similarity-space is computed as a cosine between vectors and based solely on the semantic relatedness between words (with words represented as points in undifferentiated Euclidean space). Such similarity-space representations follow basic metric axioms of Euclidean space (see below). Third, similarity-space representations are unstructured; the only measure of semantic relatedness given by LSA is based upon statistical regularities or co-occurrence effects of words within given corpus.

Related to the points above are some major flaws of LSA-like similarity-space representations. Using solely word-association matrix, LSA can generate only a single prevalent aspect (or gist) of the corpus. Such approach gives no hint about the underlying thematic composition of corpora and consequently cannot credibly account for different senses or meanings of words (synonymy and polysemy) nor their topical distribution within and across documents, or context. As a by-product, this makes it difficult for LSA to combat the data sparsity[51]. A further problem is that the computation of semantic space in LSA is not constrained by a collective sum; there are no causal constraints between clusters in the similarity-space, and since the parameters grow linearly with the size of the corpus, LSA is prone to overfitting.

LDA, on the other hand, is a generative probabilistic topic model. Unlike LSA, LDA uses a top-down approach and generates structured representations based on topic distributions: words are generated by topics and these topics are exchangeable within a document. This simple approach can account for basic properties of lexical meanings, such as polysemy and synonymy (Blei and Lafferty 2009). Since each topic is represented as a probability distribution over words, we can grasp multiple meanings of words, as well as their gist $P(w|g)$. Moreover, the probability

[51] Data sparsity refers to low density of data and connections. Lexical data sparsity occurs in cases where we don't have enough statistical information about certain words and/or their connections. Since LSA offers no other mechanisms for establishing relations between words, apart from word-association matrix, data sparsity is a common problem.

distribution over topics $P(z|w)$ is constrained by the collective sum of topic-mixture weights normalized to 1. This enforces a causal relation between topic distributions: the increase in probability of one topic consequently decreases the probability of other topics. The upshot of such approach is that by controlling the number of topics we can combat data sparsity.

18.1.2 Tversky's criticism

Even more pressing problems for similarity-space models, especially from cognitive science perspective, present the results of experimental studies on human similarity judgments carried out by Tversky (1977; Tversky & Gati 1978, 1982, Gati & Tversky 1984, Tversky & Hutchinson, 1986). Tversky *et al.* had argued, that human similarity judgments violate three basic metric axioms of Euclidean space (δ is a metric distance function):

a) *minimality*: $\delta(a, b) \geq \delta(a, a) = 0$,
b) *symmetry*: $(\delta(a, b) = \delta(b, a))$, and
c) *triangle inequality*: $\delta(a, b) + \delta(b,c) \geq \delta(a, c)$.

The *minimality* (self-similarity) assumes that the similarity between an object and itself is greater than or equal to the similarity between two distinct objects. The notion of *self-similarity* is hard to test empirically, but in data scaling, *self-dissimilarities* are commonly defined to be 0. More potent argument is hidden in the notion of *symmetry*, the assumption that the similarity between two points (or a point and itself) in Euclidean space is symmetric. Based on evidence from cognitive psychology experiments (e.g., Rosch 1975, 1978, Rosch and Mervis 1975, Mervis and Rosch 1981, Tversky 1977, Tversky and Gati 1978), this is often not the case. Human similarity judgments are generally asymmetric: the more prominent objects or prototypes are less similar compared to less prominent objects and vice versa. To use an example given by Tversky: the similarity of Tel Aviv to New York "exceeds the similarity" of New York to Tel Aviv. The point is "[w]e tend to select the more salient stimulus, or the prototype, as a referent, and the less salient stimulus, or the variant, as a subject" (Tversky 1977, p. 328). Thus, '*a* is like *b*' is generally not equivalent to converse similarity of '*b* is like *a*'. Since LSA and most geometric

models use cosine between vectors (which is symmetrical) as a similarity measure, such models cannot account for the asymmetries in judgments of similarity (but see the discussion on the more complex measures of similarity in Krumhansl 1978, Nosofsky 1988, 1991, Hahn and Chater 1997). A related concept is *triangle inequality*. By definition, the length of individual edge (of the triangle) cannot be larger than the length total of the two remaining edges. The triangle inequality constraints the similarity between *a* and *c* by the similarities between *a* and *b* and between *b* and *c*. This implies a certain similarity constraint on all points (vertexes) of the triangle: if *a* is similar to *b*, and *b* is quite similar to *c*, than, according to triangle inequality, *a* and *c* cannot be very dissimilar. Again, psychological validity of such assumption has been strongly criticized. Consider Tversky's example of the similarity between countries: Jamaica is similar to Cuba (taken in geographical context), Cuba is similar to Russia (taken in political context; at least in 1970's), but Russia and Jamaica are not similar. In effect, "... the triangle inequality is likely to fail when people shift their frame of reference from one judgment to another" (Tversky and Gati 1978, p. 149-150).

The three basic axioms of Euclidean metric characteristic for similarity-space models of cognition are arguably not psychologically valid criteria for modeling similarity judgements. Shift in the frame of reference also means shift in contextual or situational aspects of similarity judgments. In terms of cognitive semantics, a shift in reference is a conceptual shift, a shift in the underlying quality dimensions. In human similarity judgments (as in the example above), the perceived distances usually exceed the constraints of triangle inequality. Thus, similarity is not *transitive*. To use the analogy of conceptual spaces, by focusing on particular region (or part of region) in conceptual space, we can pick out the dimensions salient for the subject and not necessarily salient for the referent. Moreover, the position of particular object in conceptual space is further dependent on the 'goodness of example' in relation to the prototype. Using Lakoff's example, "Pope is a bachelor" holds a 'sort of' relation: Pope is technically a bachelor, has all the formal attributes ("an unmarried adult male"), but is neither the best nor typical example of a 'bachelor' (e.g., Pope cannot marry). Thus, different dimensions, as well as typicality and 'goodness of example' are involved in human similarity judgments (e.g., when comparing subject to its referent and vice versa).

It is important to note that Tversky's criticism doesn't hold for all geometric models of similarity (for a detailed review, see Hahn and Chater 1997). As Krumhansl (1978) and Nosofsky (1991) had shown, by refining or selectively tuning the basic similarity-space model, e.g. by employing additional flexible attention weighing of dimensions, geometric models can account for asymmetries. Nevertheless, such modifications are generally functional and serve to solve very specific issues (e.g., see Johannesson 2000), with the algorithms appropriated for the task at hand, but lack theoretical or explanatory commitment. To account for asymmetry, especially in terms of stimulus and inductive biases present in human similarity judgments, probabilistic approaches to modeling cognition seem more promising (Griffiths et al. 2008, 2010, Clark (in press)).

19 Conclusion

Admittedly, this thesis lacks rigorous analysis of special cases of symbolic and connectionist modeling, and more generally, a specific philosophical commitment. My overall aim has been to dismiss prevailing traditional accounts of modeling representations, especially in relation to meaning and semantics, and argue for an alternative approach to semantics by building a computer model.

Parts I and II of the thesis dealt with symbolic and connectionist approaches, emphasizing their essential characteristics and differences. Both accounts have been further evaluated from the perspective of *systematicity* and *productivity* of language, the two properties that exhibit *compositionality* and show the latter is not only characteristic of natural language, but of human cognition in general. The notion if *compositionality* presents a problem for connectionist approach. Arguably, some structure in the form of functional compositionality can be achieved by modifying a connectionist network (as shown by Pollack's RAAM) to simulate simple composition, e.g. of word tokens. But it is important to note that until now, connectionist functional compositionality has shown success on language tasks of a very limited scope, and does not scale up to more general properties of language. Unlike symbolic approach, neural networks simply do not have structural or methodological means to account for more abstract and hierarchical representations, and to use these same representations for further reasoning – the network does not

operate upon, in the sense of 'being detached from', as is the case for symbol systems, but within representational structures. In my view, this is the most pressing problem for connectionist approach. Hence, Fodor and Pylyshyn might be right in arguing connectionist compositionality is neither sufficient nor appropriate for modeling higher, more abstract aspects of language. Furthermore, but contrary to Fodor and Pylyshyn's criticism, some researchers argue connectionist models might be just too powerful to carry any empirical or explanatory value. For example, Massaro (1988) claims that the computational power of connectionist system seems to be unbounded: connectionist networks can simulate almost anything. To mitigate this problem, Regier (1996), inspired by Feldman's *structured connectionist models* (Feldman 1989), proposed *constrained* connectionism. Again, in practice, the constraints on the connectionist network had to be applied individually, to appropriately tackle each specific case, and not as a part of connectionist mechanism in general (see examples in Regier 1996). This opens a pressing question: to what extent is such connectionist system still seen as autonomuos, unlike symbol system exempt from pre-defined rules and procedures, reacting exclusively to the input and general setting of weights?

Classical symbolic approach, on the other hand, has its own set of problems. Reserving itself the domain of abstract thought and problem solving, symbolic approach has no answer to the challenges brought up by lower-level cognition, such as perception and bodily experience. In a classical computational system, to solve a specific problem, the decisions need to be hand-coded into the system as rules, in a top-down manner. As consequence, such system cannot represent the emergent properties of the environment, i.e. cognitive and bodily interactions with environment, or the bottom-up influences that lower (dynamic and distributed) cognitive processes have on higher-level cognition. By-products are well known: symbol-grounding problem and frame problem. And, as has been argued in the discussion on semantics (Part III), rigid, set-theoretical perception of the world faces difficulties in its own backyard: meaning is not defined by a set of necessary and sufficient conditions, nor is it a part of static, ontologically defined view of the world, rather, meaning is a conceptual entity, affected by individual's beliefs, background knowledge and context.

Common to both paradigms is the lack of cognitively plausible explanation of semantics. As argued throughout the thesis, concepts are the vehicles of meaning, hence the obvious need for conceptual level. Neither of traditional approaches nor existing hybrid formations can successfully account for conceptual representations. What is missing then is a proper conceptual level that could mediate between symbolic and sub-symbolic representations, and use conceptual representations to model graded structure of concepts and categories, and various influences of context.

The proposed alternative to the traditional models of semantics is based on the theory of *conceptual spaces* (Gärdenfors 2000). My main goal has been to build a computer model for representing semantics of natural languages by coupling conceptual spaces with methods for natural language processing. The latter are necessary for generating quality dimensions needed to construct conceptual spaces. Two different approaches to natural language processing, loosely mirroring the logic of their counterparts in symbolic and connectionist approaches to cognition, have been implemented in *SpaceWalk*. A similarity-space model LSA, arguably one of the most prominent examples of bottom-up (and hence, in spirit connectionist) approach to semantic analysis, has been proven inferior, both functionally and theoretically, to top-down (and hence, in spirit symbolic) probabilistic topic model LDA.

The effectiveness of LSA, pLSA and LDA algorithms has been widely discussed in scientific literature (e.g., Landauer and Dumais 1997, Seidenberg and MacDonald 1999, Landauer et. al 2006, Blei *et al.* 2003, Hofmann 1999, 2001, Steyvers and Griffiths 2007, Griffiths *et al.* 2010 and Blei 2011, among others). Based on the results gathered from these studies, probabilistic models in general fare better than similarity-space models (cf. Blei *et al.* 2003 and Hofmann 1999). However, how plausible are theoretical assumptions supporting respective approach is still an unsettled issue. Two issues of *Trends in Cognitive Science* (July 2006 and August 2010) have been devoted to connectionist and probabilistic modeling of cognition, with proponents of each approach generally split on the issue. Connectionists argue that bottom-up "emergentist accounts of cognition are more theoretically constraining than structured probability accounts" (Altmann (2010, p. 340), cf. McClelland *et al.* (same issue)). Proponents of probabilistic approach, on the other hand, argue that the top-down generative approach yields "greater flexibility for

exploring the representations and inductive biases that underlie human cognition" (Griffiths *et al.* 2010, p. 357; see also Lee (same issue)). Most of the scientific community, however, while pointing out the advantages and disadvantages of each approach, argues for a unified theory to modeling cognition. By connecting structural aspects with their emergent counterparts and perceptual substrata, the more abstract aspects of cognition could be grounded in perception and action (in the same issue, see for example, Feldman, Gopnik *et al.*, Kruschke, and Marcus, among others).

As has been shown, similarity-space models provide weak theoretical support for cognitively realistic modeling of natural language and cognition. Specifically, theoretical assumptions underlying similarity-space models are not supported by experimental evidence gathered from the studies in cognitive psychology, especially studies on human language and categorization (Rosch et al.) and studies on similarity judgments (Tversky et al.). From the perspective of human language processing, such models are fundamentally flawed. Admittedly, similarity-space models can lead to the discovery of general semantic patterns in large text corpora, and might even generate appropriate statistical approximations of the corpus data, or simulate some of the effects of language use, for example passing the synonym test (see (Landauer and Dumais 1997)). Nevertheless, while LSA might simulate effects of synonymy, these effects are generated solely through statistical operations on word association matrix, with little additional explanatory value. In reality, human conceptual structure is not the result of statistical inferences based exclusively on word associations, but is inherently affected by conceptual, categorical and contextual knowledge, let alone cultural and social influences and context (see e.g., Jäger and van Rooij 2007). As such, it is part of a larger abstract knowledge structure. By being constrained to the single view (gist of the corpora), LSA-like similarity-space models cannot account for these phenomena.

Probabilistic approach brings fresh air into traditional accounts of language and cognition. As a generative topic model, LDA is conceptually closer to the symbolic approach, but overcomes many of its vices. For one, it allows for hybridity and coupling of different representational architectures. LDA utilizes associative, approximating data structures and thus allows the 'environment' to influence the representational structure of the system. Furthermore, the notion of probability

represents a set of top-down constraints which, taken as *inductive biases*[52], can account for effects in human similarity judgments (Griffiths et al. 2008, 2010, Clark (in press)). Coupled with conceptual spaces, it offers a more flexible framework for creating and exploring semantic representations and aims to explain "how inductive biases – the constraints on learning and memory, which influence our conclusions from limited data – relate to the concepts ..." (Griffiths et al. 2008, p. 3503).

The role of conceptual spaces in modeling meaning and semantics of natural languages is significant. In a computational model such as *SpaceWalk*, conceptual spaces add another, conceptual level onto existing semantic representations (whether generated by LDA, LSA or pLSA). What we get, in machine-readable form, are not only representations of clusters of objects, concepts, properties and similarity relations, but the framework that exploits the underlying quality dimensions and projects them onto conceptual space according to the mode of graded categorization (see Rosch's prototype theory). Higher level symbolic operations can then be applied to manipulate these representations and form reasoning over conceptual space, e.g. for knowledge representation (e.g., in form of hierarchical dependencies between objects, concepts and categories; Strle and Marolt 2012), semantic web (Gärdenfors 2004), or as a means of communication (Gärdenfors and Warglien 2006, 2011, 2012). Moreover, from the top-down perspective, anchoring in conceptual spaces "could take advantage from top down information: high level, symbolic knowledge can constrain the possible shape of the (interpolated or extrapolated) trajectories" (Chella et al. 2003, p. 195). By connecting various levels of analyticity, e.g. by coupling conceptual space with the top-down and bottom-up approaches to natural language processing, such a system becomes truly hybrid.

What is currently lacking is a more integrated approach of using symbolic operations dynamically. For example, in current version of *SpaceWalk* only partial snapshots of conceptual space are possible and the projections of alternative sets of quality dimensions need repeated computations. Ideally, we should be able to explore the conceptual space by manipulating underlying quality dimensions (e.g., by changing the number of topics or by changing weights of individual topics or dimensions) and

[52] Here (in the context of machine learning), *inductive biases* are understood as different factors that affect human judgment, e.g. prior knowledge, prejudices, expectations, etc. In Bayesian framework they are expressed as a prior distribution over hypotheses.

discovering different semantic structures on the fly. There are numerous possible applications for such an approach, most obvious are the areas of machine learning, knowledge representation and semantic web, with one example being the discovery and exploration of latent semantic structures in digital text collections. This remains a motivation for future work.

References:

"Representation, *n.*" OED Online. March 2011. Oxford University Press. <http://www.oed.com/view/Entry/162997?rskey=FvhWq6&result=1&isAdvanced=false> (accessed September 05, 2011).

Akman, V. and Blackburn, P. (eds.). 2000. Special issue on Alan Turing and artificial intelligence. *Journal of Logic, Language and Information, vol.* 9, No. 4.

Alhoniemi, E. Himberg, J. and Vesanto, J. 2002. *SOM Toolbox Version 2.0beta*. http://www.cis.hut.fi/projects/somtoolbox/documentation/index.shtml#requirements

Altmann, G. T. M. 2010. Why emergentist accounts of cognition are more theoretically constraining than structured probability accounts: comment on Griffiths et al. and McClelland et al. *Trends in cognitive sciences*. *vol.* 14, Issue 8, August 2010, Page 340.

Anat, B. and Anat, M. 2012. Ludwig Wittgenstein. In Zalta, N. E. (ed.), *The Stanford Encyclopedia of Philosophy (Summer 2011 Edition)*. http://plato.stanford.edu/archives/sum2011/entries/wittgenstein/.

Anderson, J. R. 1983. *The architecture of cognition*. Cambridge, MA: Harvard Univ. Press.

Anderson, J. R. 1991. The adaptive nature of human categorization. *Psychological Review*, 98(3), 409–429.

Anellis, I.H. 1991. The first Russell paradox. In Drucker, T. (ed.), *Perspectives on the History of Mathematical Logic*. Cambridge, MA: Birkäuser Boston. pp. 33–46.

Anglin, J. M. 1977. *Word, object and conceptual development*. New York: Norton.

Atran, S. 1989. Basic Conceptual Domains. *Mind and Language*, 4(1-2), 7–16.

Augustin, D. and Leder, H. 2006. Art expertise: a study of concepts and conceptual spaces. *Psychology Science*, 48(2), pp. 135-156.

Aydede, M. 2010. The Language of Thought Hypothesis. Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy (Fall 2010 Edition)*. Available at: http://plato.stanford.edu/archives/fall2010/entries/language-thought

Bailenson, J. N., Shum, M. S., Atran, S., Medin, D. L., and Coley, J. D. 2002. A bird's eye view: Biological categorization and reasoning within and across cultures. *Cognition*, 84(1), 1–53.

Ballester, J., Patris, B., Symoneaux, R., and Valentin, D. 2008. Conceptual vs. perceptual wine spaces: does expertise matter. *Food quality and preference*, 19(3), 267–276.

Barsalou, L. W. 1993. Flexibility, structure, and linguistic vagary in concepts: Manifestations of a compositional system of perceptual symbols. In A. C. Collins, S. E. Gathercole, and M. A. Conway, (eds.), *Theories of memories*, (pp. 29-101). London: Erlbaum.

Barsalou, L. W. 1999. Perceptual symbol systems. *Behavioral and Brain sciences*, 22(04), pp. 577-660.

Barsalou, L. W. 2008. Grounded cognition. *Psychology*, 59(1), 617.

Barsalou, L.W. 1983. Ad hoc categories. *Memory and Cognition*, 11, 211-227.

Barsalou, L.W. 1985. Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology*: *Learning, Memory, and Cognition*, vol. 11, pp. 629-654.

Barsalou, L.W. 1987. The instability of graded structure: Implications for the nature of concepts. In U. Neisser, (ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization*, (pp. 101-140). Cambridge: Cambridge University Press.

Barsalou, L.W. 2005. Situated conceptualization. In H. Cohen, and C. Lefebvre, (eds.), *Handbook of categorization in cognitive science*, (pp. 619-650). St. Louis: Elsevier.

Barsalou, L.W., Solomon, K.O., and Wu, L.L. 1999. Perceptual simulation in conceptual tasks. In M.K. Hiraga, C. Sinha, and S. Wilcox (eds.), *Cultural*,

typological, and psychological perspectives in cognitive linguistics: The proceedings of the 4th conference of the International Cognitive Linguistics Association, *vol.* 3, (209-228). Amsterdam: John Benjamins.

Bechtel, W. 1998. Representations and cognitive explanations. *Cognitive Science,* vol. 22: 295–318.

Bechtel, W. 2001. Representations: from neural systems to cognitive systems. In W. Bechtel, P. Mandik, J. Mundale and R. Sufflebeam (eds.), *Philosophy and the Neurosciences*, pp. 332–348. Oxford: Blackwell Publishing.

Bechtel, W. and Abrahamsen, A. 1991. *Connectionism and the Mind. An Introduction to Parallel Processing in Networks*. Oxford: Basil Blackwell

Beer, R. D. 1995a. A dynamical systems perspective on autonomous agents. *Artificial Intelligence,* 72, 173–215.

Beer, R. D. 1995b. Computational and dynamical languages for autonomous agents. In R. F. Port, (ed.), *Mind as Motion Explorations in the Dynamics of Cognition*, (pp. 121–147). Cambridge, MA: MIT Press.

Beer, R. D. 2000. Dynamical approaches to cognitive science. *Trends in cognitive sciences*, 4(3), 91–99.

Beer, R. D. 2003. The dynamics of active categorical perception in an evolved model agent. *Adaptive Behavior*, 11(4), 209-243

Beer, R.D. and Gallagher, J. C. 1992. Evolving dynamical neural networks for adaptive behavior. *Adaptive behavior* 1 (1): 91–122.

Berlin, B. 1972. Speculations on the growth of ethnobotanical nomenclature. *Language in Society*, 1, 51-86.

Berlin, B. and Kay, P. 1969. *Basic Color Terms: Their Universality and Evolution*. University of California Press, Berkeley.

Bickhard, M. H. 1998. *Robots and Representations. From Animals to Animats*. Cambridge, MA: MIT Press.

Bickhard, M. H. 2000. Information and representation in autonomous agents. *Cognitive Systems Research*, 1, 65-75.

Bickhard, M. H. and Terveen, L. 1995. *Foundational issues in artificial intelligence and cognitive science: impasse and solution*. Amsterdam: North-Holland.

Blei, D. 2012. Introduction to probabilistic topic models. *Communications of the ACM*, (to appear).

Blei, D. and Lafferty, J. 2009. Topic Models. In A. Srivastava and M. Sahami, (eds.), *Text Mining: Classification, Clustering, and Applications*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series.

Blei, D. M., Ng, A. Y. and Jordan, M. I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5), pp. 993-1022.

Boden, M. A. (2006). *Mind as machine: A history of cognitive science*. Oxford: Clarendon Press.

Boden, M. and Niklasson, L. 2000. Semantic systematicity and context in connectionist networks. *Connection Science*, 12(2), pp. 111-142.

Brandt, P. A. 2005. Mental spaces and cognitive semantics: A critical comment. *Journal of Pragmatics*. *vol.*: 37(10), pp. 1578-1594. Publisher: Elsevier.

Brooks, R. A. 1990. Elephants don't play chess. *Robotics and autonomous systems*, 6(1-2), pp. 3-15.

Brooks, R. A. 1991. Intelligence without representation. *Artificial Intelligence*, 47(1-3), pp. 139-159.

Cabeza, R. and Nyberg, L. 2000. Neural bases of learning and memory: functional neuroimaging evidence. *Current Opinion in Neurology*, 13(4), p. 415.

Carnap, R. 1947. *Meaning and Necessity*. University of Chicago Press.

Chalmers, D. J. 1993a. Why Fodor and Pylyshyn Were Wrong: The Simplest Refutation. *Cognitive Science*, 6(3), pp. 340-347. Routledge.

Chalmers, D. J. 2002. On Sense And Intension. In J. Tomberlin (ed.), *Philosophical Perspectives 16: Language and Mind*, (pp. 135-82). Blackwell.

Chalmers, D.J. 1990a. Syntactic Transformations on Distributed Representation. *Connection Science*, 2 (2), pp. 53–62.

Chalmers, D.J. 1990b. Why Fodor and Pylyshyn Were Wrong. In *Proceedings of The Twelfth Annual Conference of the Cognitive Science Society*, Cambridge, MA, July 1990, pp. 340-347.

Chalmers, D.J. 1993b. Connectionism and Compositionality: Why Fodor and Pylyshyn Were Wrong. *Philosophical Psychology*, 6 (3), pp. 305–319.

Chalmers, D.J. 2012. A Computational Foundation for the Study of Cognition. *Journal of Cognitive Science*, forthcoming (2012). Available at: http://consc.net/papers/computation.html

Chater, N, Tenenbaum, J. B. and Yuille, A. 2006. Probabilistic models of cognition: Conceptual foundations. *Trends in cognitive sciences*, 10(7), pp. 287 – 291.

Chater, N. and Brown, G. D. 2008. From universal laws of cognition to specific cognitive models. *Cognitive Science*, 32(1), 36–67.

Chater, N., Oaksford, M., Hahn, U. and Heit, E. 2010. Bayesian models of cognition. *Cognitive Science*, 1, pp. 811–823. doi: 10.1002/wcs.79

Chella, A., Frixione, M. and Gaglio, S. 2003a. Conceptual spaces for anchoring. *Robotics and Autonomous Systems*, 43(2-3): 193-195.

Chella, A., Frixione, M. and Gaglio, S. 2003b. Anchoring symbols to conceptual spaces: the case of dynamic scenarios. *Robotics and Autonomous Systems* 43(2-3): 175-188.

Chemero, A. 2000a. Anti-Representationalism and the Dynamical Stance. *Philosophy of Science*, 67(4), p. 625.

Chemero, A. 2000b. Representation and 'Reliable Presence'. *Conceptus Studien 14: The New Computationalism*, pp. 9-25.

Chemero, A. 2009. *Radical embodied cognitive science*. Cambridge, MA: MIT Press

Chomsky, N. 1957. *Syntactic structures*. The Hague: Mouton.

Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press

Chomsky, N. 1968. *Language and Mind*. New York: Harcourt, Brace and World.

Chomsky, N. 1980. *Rules and representations*. Oxford: Basil Blackwell.

Chown, E. and Kaplan, S. 1992. Active symbols, limited storage and the power of natural intelligence. *Brain and Behavioral Sciences*, 15(3), 442-443.

Christiansen, M.H., and Chater, N. 1994., Generalization and Connectionist Language Learning. *Mind and Language*, 9(3), pp.273–287.

Church, A. 1936. An unsolvable problem of elementary number theory. *American Journal of Mathematics*, 58: 345-363.

Churchland, P. M. 1995. *The Engine of Reason, the Seat of the Soul: A Philosophical Journey into the Brain*. Cambridge, MA: MIT Press.

Clark, A. (in press). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain sciences*.

Clark, A. 1989. Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing. *Explorations in cognitive science*. Cambridge, MA: MIT Press.

Clark, A. 1991. Systematicity, structured representations andcognitive architecture: A reply to Fodor and Pylyshyn. In T. Horgan and J. Tienson (eds.), *Connectionism and the Philosophy of Mind*. Dordrecht: Kluwer, pp. 198–218.

Clark, A. 1997. The Dynamical Challenge. *Cognitive Science*, 21(4), pp. 461-481.

Clark, A. 1998. Philosophical issues in brain theory and connectionism. In M. A. Arbib (ed.), *The handbook of brain theory and neural networks*, pp. 738-741. Cambridge, MA: MIT Press.

Cleland, C. 1993. Is the Church-Turing thesis true? *Minds and Machines*, 3, 3, pp. 283-312.

Cleland, C. 1995. Effective procedures and computable functions. *Minds and Machines*, 5, pp. 9–23.

Cleland, C. 2004. The concept of computability. *Theoretical Computer Science*, 317, pp. 209-225.

Collins, A. M., and Loftus, E. F. 1975. A spreading-activation theory of semantic memory. *Psychological Review*, 82, pp. 407-428.

Cooper, R. and Franks, B. 1993. Interruptibility as a constraint on hybrid systems. *Minds and Machines*, 3:73–96

Copeland, B.J. 1998. Turing's O-machines, Penrose, Searle, and the Brain. *Analysis*, 58, pp. 128-38.

Copeland, B.J. 2002. The Church-Turing Thesis. In E. Zalta (ed.), *The Stanford Encyclopaedia of Philosophy*. Stanford University, http://plato.stanford.edu.

Crane, T. 2003. *The Mechanical Mind*. London: Routledge.

Cummins, R. 1983. *The Nature of Psychological Explanation*. Cambridge, MA: MIT Press

Cummins, R. 2002. Truth and meaning. In J. K. Campbell, M. O'Rourke and D. Shier (eds.), *Meaning and Truth: Investigations in Philosophical Semantics*. New York: Seven Bridges Press.

Cummins, R. and Schwarz, G. 1991. Connectionism, computation and cognition. In T. Horgan and J. Tienson (eds.), *Connectionism and the Philosophy of Mind*. Dordrecht: Kluwer.

Davidson, D. 1967. Truth and Meaning. *Synthese*, 17:304-323.

De Groot, A. D. 1965. *Thought and Choice in Chess*. The Hague: Mouton.

Deacon, T. W. 1997. *The symbolic species: the co-evolution of language and the brain*. W.W. Norton and Company.

Dell, G. S. 1986. A spreading-activation theory of retrieval in sentence production. *Psychological Review*, 93(3), pp. 283-321.

Dennet, D. C. 1986. *Brainstorms: Philosophical essays on mind and psychology*. Cambridge, MA: MIT Press.

Doug Edwards quote. http://www.jimgilliam.com/2005/11/xooglers.php

Dretske, F. I. 1988. *Explaining Behavior: Reasons in a World of Causes*. Cambridge, MA: MIT Press.

Dreyfus, H. 1972. *What Computers Can't Do*. New York: Harper and Row.

Dreyfus, H. 1992. *What Computers Still Can't Do*. Cambridge, MA: MIT Press.

Dummet , M. 1973. *Frege: Philosophy of Language*. (2nd ed.). Cambridge, MA: Harvard University Press

Dummett, M. 1981. *The Interpretation of Frege's Philosophy*. Cambridge, MA: Harvard University Press.

Elman, J. et al. (1996). *Rethinking innateness: A connectionist perspective on development.* Cambridge, MA: MIT Press.

Elman, J. L. 1989. Representation and Structure in Connectionist Models. In G. T. M. Altmann, (ed.) *Cognitive models of speech processing Psycholinguistic and computational perspectives*. Cambridge, MA: MIT Press

Elman, J. L. 1990a. Finding structure in time. *Cognitive Science*, 14, pp. 179–211.

Elman, J. L. 1990b. Structured representations and connectionist models. In Altmann, G. T. M. (ed.), *Computational and Psycholinguistic Approaches to Speech Processing*. New York: Academic Press.

Evans, V. 2003. *The Structure of Time. Language, Meaning and Temporal Cognition*. Amsterdam: Benjamins.

Fauconnier, G. 1985. *Mental Spaces*. Cambridge, MA: MIT Press.

Fauconnier, G. 1997. *Mappings in Thought and Language*. Cambridge University Press.

Fauconnier, G. and Sweetser, E. 1996. *Spaces, worlds, and grammar*. University of Chicago Press.

Feldman, J. 1989. Neural representation of conceptual knowledge. In L. Nadel, P. Culicover and R. M. Harnish (eds.), *Neural connections, mental computation*. Cambridge, MA: MIT Press.

Feldman, J. A. 2010. Cognitive Science should be unified: comment on Griffiths et al. and McClelland et al. *Trends in cognitive sciences*, vol. 14(8), August 2010, p. 341.

Fillmore, C. 1982. Frame Semantics. In Linguistic Society of Korea (ed.), *Linguistics in the Morning Calm*, pp. 111-38. Seoul: Hanshin.

Fillmore, C. 1985. Frames and the Semantics of Understanding. *Quaderni di Semantica*, 6(2), pp. 222-53.

Fodor, J. A. 1975. *The Language of Thought*. Cambridge, MA: MIT Press.

Fodor, J. A. 1981. *Representations*. Cambridge, MA: MIT Press.

Fodor, J. A. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press.

Fodor, J. A. 2008. *LOT 2: The Language of Thought Revisit*ed. Oxford University Press.

Fodor, J. A. and McLaughlin, B. P. (1990). Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition*, 35: 183-204.

Fodor, J. A. and Pylyshyn, Z. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28: 3-71.

Fodor, J. and Lepore, E. 1992. *Holism: A Shopper's Guide*. (Oxford: Blackwell).

Foster, J. A. 2000. *The Nature of Perception*. New York: Oxford University Press

Frank, R., Dirven, R. and Ziemke, T. 2008. *Body, Language and Mind – vol. 2: Sociocultural situatedness*. Berlin, Mouton de Gruyter.

Franks, B. and Cooper, R. 1995.Why Some Hybrid Solutions Aren't Really Solutions (and Why Others Aren't Really Hybrid). In H. Hallam (ed.), *Hybrid Problems, Hybrid Solutions*, pp. 61–71. Oxford: IOS Press.

Frege, G. 1884/1974. *Die Grundlagen der Arithmetik: eine logisch-mathematische Untersuchung über den Begriff der Zahl*. Breslau: w. Koebner, 1884. Austin, J. L. (transl.). 1974. *The Foundations of Arithmetic: A Logic-Mathematical Enquiry into the Concept of Number*. Oxford: Blackwell, second revised edition, 1974.

Frege, G. 1892/1980. Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100: 25-50. Geach, P. and Black, M. (transl.; eds.). 1980. *On Sense and Reference*. *Translations from the Philosophical Writings of Gottlob Frege*. 3rd ed. Oxford: Blackwell.

Gallese, V. and Lakoff, G. 2005. The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive Neuropsychology*, 22: 455-479.

Gärdenfors, P. 1988. Semantics, conceptual spaces and the dimensions of music. In V. Rantala, L. Rowell and E. Tarasti (eds.), *Essays on the Philosophy of Music*. *Acta Philosophica Fennica*, vol. 43, pp. 9-27. Helsinki.

Gärdenfors, P. 1990. Induction, conceptual spaces and AI. *Philosophy of Science*, 57, pp. 78-95.

Gärdenfors, P. 1991. Frameworks for properties: Possible worlds vs. conceptual spaces. In L. Haaparanta, M. Kusch and I. Niiniluoto (eds.), *Language, Knowledge and Intentionality. Acta Philosophica Fennica*, vol. 49, pp. 383-407. Helsinki.

Gärdenfors, P. 1993. The emergence of meaning. *Linguistics and Philosophy*, 16: 285-309.

Gärdenfors, P. 1996. Conceptual spaces as a basis for cognitive semantics. In: A. Clark et al. (eds.), *Philosophy and Cognitive Science*. Kluwer, Dordrecht, pp. 159-180.

Gärdenfors, P. 1997. Meanings as conceptual structures. *Mindscapes: Philosophy, Science, and the Mind*, pp. 61–86.

Gärdenfors, P. 1999a. Does semantics need reality? In A. Riegler, M. Peschl and A. Stein (eds.), *Understanding Representation in the Cognitive Sciences*. New York : Kluwer Academic/Plenum Publishers, pp. 209-217.

Gärdenfors, P. 1999b. Some tenets of Cognitive Semantics. In J. Allwood and P. Gärdenfors (eds.), *Cognitive Semantics: Meaning and Cognition*, (pp. 19-37). Amsterdam; Philadelphia: J. Benjamins.

Gärdenfors, P. 2000. *Conceptual Spaces: The Geometry of Thought*. Cambridge, MA: MIT Press.

Gärdenfors, P. 2004. How to make the semantic web more semantic. *Formal Ontology in Information Systems*, pp. 19–36.

Gärdenfors, P. 2011. Semantics Based on Conceptual Spaces. In M. Banerjee and A. Seth (eds.), *Logic and Its Applications. Lecture Notes in Computer Science*, *vol.* 6521. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1-11.

Gärdenfors, P. and Warglien, M. 2006. Cooperation, conceptual spaces and the evolution of semantics. In P. Vogt et al. (eds), *Symbol Grounding and Beyond*. Berlin, Heidelberg: Springer, pp. 16-30.

Gärdenfors, P. and Warglien, M. 2011. Semantics, conceptual spaces and the meeting of minds. *Synthese* (8 June 2011), pp. 1-29.

Gärdenfors, P. and Warglien, M. 2012. The development of semantic space for pointing and verbal communication. In J. Hudson, U. Magnusson and C. Paradis (eds.), *Conceptual Spaces and the Construal of Spatial Meaning. Empirical Evidence from Human Communication.* Cambridge: Cambridge University Press.

Garner, W. R. 1974. *The Processing of Information and Structure*. Potomac, MD: Erlbaum.

Garner, W. R. 1978. Selective attention to attributes and to stimuli. *Journal of Experimental Psychology: General*, vol. 107 (3), pp. 287-308

Gelman, S. A. 1996. Concepts and Theories. In Gelman, R. and Au, T. K. (eds.), *Perceptual and Cognitive Development. Handbook of Perception and Cognition* (2nd ed.). San Diego: Academic Press.

Gibbs, R. 2005. The psychological status of image schemas. In B. Hampe (ed.), *From perception to meaning: Image schemas in cognitive linguistics*, pp. 23-28. Berlin: Mouton.

Gibbs, R. 2006. *Embodiment and cognitive science*. New York: Cambridge University Press.

Gibbs, R. W. and Colston, H. L. 1995. The cognitive psychological reality of image schemas and their transformations. *Cognitive Linguistic* 6, pp. 347-378.

Gibson, J. J. 1979. *The ecological approach to visual perception*. Houghton-Mifflin.

Glenberg, A. M. 1997. Mental models, space, and embodied cognition. In T. B. Ward, S. M. Smith and J. Vaid (eds.), *Creative thought: An investigation of conceptual structures and processes.* Washington, DC: American Psychological Association, pp. 495-522.

Glenberg, A. M. and Kaschak, M. P. 2002. Grounding laguage in action. *Psychonomic Bulletin and Review* 9, pp. 558-565.

Gödel, K. 1931. Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme, I. *Monatshefte für Mathematik und Physik* 38: 173-98. DOI 10.1007/BF01700692 Available online via SpringerLink.

Goldstone, R. 1998. Perceptual Learning. *Annual Review of Psychology*, 49, pp. 585-612.

Goldstone, R. L., Schyns, P. G., and Medin, D. L. 1997. *Perceptual learning*. Academic Press.

Gopnik, A. and Meltzoff, A. N. 1992. Categorization and Naming: Basic-Level Sorting in Eighteen-Month-Olds and Its Relation to Language. *Child Development*, 63(5), pp. 1091–1103.

Gopnik, A., Wellman, H. M., Gelman, S., and Meltzoff, N. 2010. A computational foundation for cognitive development: comment on Griffths et al. and McLelland et al. *Trends in cognitive sciences*. *vol.* 14, Issue 8, August 2010, pp. 342-343.

Grady, E. J. 2005. Image schemas and perception: Refining a definition. In B. Hampe (ed.), *From perception to meaning: Image schemas in cognitive linguistics*, pp. 21-35. Berlin: Mouton.

Gregory R. L. 1969. On how so little information controls so much behaviour. In C.H. Waddington (ed.), *Towards a theoretical biology 2*. Edinburgh: University of Edinburgh Press.

Griffiths, T. L. and Steyvers, M. 2004. Finding scientific topics. *Proceedings of the National Academy of Science*, 101, pp. 5228-5235.

Griffiths, T. L., and Steyvers, M. 2002. A probabilistic approach to semantic representation. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*.

Griffiths, T. L., Chater, N., Kemp, C. Perfors, A., Tenenbaum, J. B. 2010. Probabilistic models of cognition: exploring representations and inductive biases. *Trends in cognitive sciences*. *vol.* 14, Issue 8, August 2010, pp. 357-364.

Griffiths, T. L., Kalish, M. L., and Lewandowsky, S. 2008. Theoretical and experimental evidence for the impact of inductive biases on cultural evolution. *Philosophical Transactions of the Royal Society*, 363, pp. 3503-3514.

Griffiths, T. L., Steyvers, M., Blei, D. M., and Tenenbaum, J. B. 2005. Integrating topics and syntax. *Advances in Neural Information Processing*, 17. Cambridge, MA: MIT Press.

Griffiths, T. L., Steyvers, M., Tenenbaum, J. B. 2007. Topics in semantic representation. *Psychological Review*, vol. 114(2), pp. 211-244.

Grossman, M. et al. 2002. The Neural Basis for Categorization in Semantic Memory. *Neuroimage*, 17(3), pp. 1549–1561.

Haaparanta, L. 1985. Frege's Doctrine of Being. *Acta Philosophica Fennica* 39. Helsinki.

Hahn, U. and Chater, N. 1997. Concepts and similarity. In K. Lamberts & D. Shanks (eds.), *Knowledge, concepts and categories*. Hove, England: Psychology Press, pp. 43-92.

Hallam, J. (ed.). 1995. *Hybrid Problems, Hybrid Solutions.* Oxford: IOS Press.

Hampton, J. A., Dubois, D. and Yeh, W. 2006. Effects of classification context on categorization in natural categories. *Memory and Cognition*, 34(7), pp. 1431-1443. doi:10.3758/BF03195908

Harnad, S. 1987. *Categorical Perception*. Cambridge: Cambridge University Press.

Harnad, S. 1989. Minds, Machines and Searle. *Journal of Experimental and Theoretical Artificial Intelligence*, 1: 5–25.

Harnad, S. 1990. The Symbol Grounding Problem. *Physica*, 42: 335-346.

Harnad, S. 1992. Connecting Object to Symbol in Modeling Cognition. In A. Clarke and R. Lutz (eds.), *Connectionism in Context*. Springer Verlag, pp. 75-90.

Haugeland, J. 1991. Representational Genera. In W. M. Ramsey, S. P. Stich, and D. E. Rumelhart (eds.), *Philosophy and connectionist theory*. Lawrence Erlbaum, pp. 171-206.

Haugeland, J. 2000. *Having thought: essays in metaphysics of mind*. Harvard University Press

Henderson, P. 1997. Sammon Mapping. *Pattern Recognition Letters* 18 (11-13), pp. 1307-1316.

Hinton, G. E. 1988. Representing part-whole hierarchies in connectionist networks. In *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*. Montreal, Canada, pp. 48-54.

Hinton, G. E. 1989. Connectionist Learning Procedures. *Artificial Intelligence*, 40: 185–234.

Hinton, G. E., and Anderson, J. A. (eds.). 1981. *Parallel models of associative memory*. Hillsdale, NJ: Erlbaum.

Hinton, G. E., and Sejnowski, T. J. 1986. Learning and relearning in Boltzmann machines. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group (eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. vol. 1: Foundations*. Cambridge, MA: MIT Press, p. 282-317.

Hinton, G. E., McClelland, J. L., and Rumelhart, D. E. 1986. Distributed representations. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group (eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. vol. 1: Foundations*. Cambridge, MA: MIT Press, p. 77-109.

Hintzman, D. 1. 1986. "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, 93, pp. 411-428.

Hofmann, T. 1999. Probabilistic Latent Semantic Analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*.

Hofmann, T. 2001. Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning Journal*, 42(1), pp. 177-196.

Holmqvist, K. 1993. Implementing Cognitive Semantics. *Lund University Cognitive Studies 17*. Lund: LUCS .

Holt, L. E., and Beilock, S. L. 2006. Expertise and its embodiment: Examining the impact of sensorimotor skill expertise on the representation of action-related text. *Psychonomic bulletin and review*, 13(4), pp. 694–701.

http://www.cis.hut.fi/projects/somtoolbox/theory/somalgorithm.shtml

Jacoby, L. L. and Dallas, M. 1981. On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General*, 110, pp. 306-340.

Jäger, G. 2007. The evolution of convex categories. *Linguistics and Philosophy*, 30(5), pp. 551-564.

Jäger, G. 2010. Natural colour categories are convex sets. In M. Aloni et al. (eds.), *Logic, Language and Meaning*. LNCS (LNAI), vol. 6042. Springer, Heidelberg, pp. 11-20.

Jäger, G. and van Rooij, R. 2007. Language stucture: Psychological and social constraints. *Synthese*, 159(1), pp. 99–130.

Jäkel, F. 2007. *Some theoretical aspects of human categorization behavior : similarity and generalization*. [Thesis/Dissertation]. Tübingen: Univ. of Tübingen.

Johannesson, M. 2000. Modelling asymmetric similarity with prominence. *British Journal of Mathematical and Statistical Psychology*, 53: 121–139.

Johnson, K. E. 2001. Impact of varying levels of expertise on decisions of category typicality. *Memory and cognition*, 29 (7), pp. 1036-1050.

Johnson, K. E., and Mervis, C. B. 1997. Effects of varying levels of expertise on the basic level of categorization. *Journal of Experimental Psychology-General*, 126(3), 248–276.

Johnson, M. 1987. *The Body in the Mind: The Bodily Basis of Reason and Imagination*. Chicago, IL: University of Chicago Press.

Johnson, M. and Rohrer, T. 2007. We are live creatures: Embodiment, American Pragmatism and the cognitive organism. In: T. Ziemke, J. Zlatev and R. Frank (eds.), *Body, Language and Mind. vol. 1. Embodiment*. Berlin: Mouton, pp. 17-54.

Johnson, M., and Lakoff, M. 2002. Why cognitive linguistics requires embodied realism. *Cognitive Linguistics* 13: 245-263.

Johnson-Laird, P. N. 1980. Mental models in cognitive science. *Cognitive Science: A Multidisciplinary Journal*, 4(1), pp. 71–115.

Johnson-Laird, P. N. 1983. *Mental models*. Cambridge, MA: Harvard University Press.

Jolicoeur, P., Gluck, M. A. and Kosslyn, S. M. 1984. Pictures and names: Making the connection. *Cognitive Psychology*, 16, pp. 243-275.

Jones, S. S. and Smith, L. B. 1993. The place of perception in children's concepts. *Cognitive Development*, 8(2), pp. 113–139.

Karmiloff-Smith, A. 1986. Some fundamental aspects of language acquisition after five. In P.Fletcher and M.Garman (eds.), *Studies in Language Acquisition, Second Revised Edition*. Cambridge: Cambridge University Press.

Karmiloff-Smith, A. 1992. *Beyond Modularity: A Developmental Perspective on Cognitive Science*. Cambridge, MA: MIT Press.

Katz, J., and Fodor, J. 1963. The structure of semantic theory. *Language* 39: 120–210.

Kleene, S.C. 1967. *Mathematical Logic*. New York: Wiley.

Kohonen, T. 1995. Self-Organizing Maps. *Series in Information Sciences*, *vol.* 30. Springer, Heidelberg.

Kohonen, T. *Intro to SOM*. SOM Toolbox.

Kosslyn, S. M. 1981. The medium and the message in mental imagery: A theory. *Psychological Review*, 88(1), pp. 46–66.

Kosslyn, S.M. and Hatfield, G. 1984. Representation without symbol systems. *Social Research*, 51, pp. 1019-1054.

Kripke, S. 1959. A completeness theorem in modal logic. *Journal of Symbolic Logic* 24: 1-24

Kripke, S. 1975. Outline of a theory of truth. *Journal of Philosophy* 72: 690-716.

Kripke, S.A. 1980. *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Krumhansl, C. 1978. Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review*, 85, pp. 450–463.

Kruschke, J. K. 2010. Bridging levels of analysis: comment on McClelland et al. and Griffiths et al. *Trends in cognitive sciences*. *vol.* 14, Issue 8, August 2010, pp. 344-345.

Laird, J. E., Newell, A., and Rosenbloom, P. S. 1987. SOAR: An architecture for general intelligence. *Artificial Intelligence*, 33, pp. 1-64.

Lakoff, G. 1987. *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago Press.

Lakoff, G. and Johnson, M. 1980. *Metaphors We Live By*. Chicago: University of Chicago Press.

Lakoff, G. and Johnson, M. 1999. *Philosophy in the flesh: the embodied mind and its challenge to Western thought*. New York: Basic Books.

Landauer, T. K. and Dumais, S. T. 1997. A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104, pp. 211-240.

Landauer, T. K., Foltz, P. W. and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2-3), pp. 259–284.

Landauer, T.K. et al (eds.). 2006. *Handbook of latent semantic analysis*. Erlbaum.

Langacker, R. 1987. *Foundations of Cognitive Grammar*. Stanford, CA: Stanford University Press.

Langacker, R. W. 1986. An Introduction to Cognitive Grammar. *Cognitive Science* 10, pp. 1-40.

Lassaline, M. E., Wisniewski, E. J., and Medin, D. L. 1992. Basic levels in artificial and natural categories: Are all basic levels created equal? In B. Bums (eds.), *Percepts, concepts and categories*. Amsterdam: Elsevier Science, pp. 327-378.

Lee, M. D. 2010. Emergent and structured cognition in Bayesian models: comment on Griffiths et al. and McClelland et al. *Trends in cognitive sciences*. *vol.* 14, Issue 8, August 2010, pp. 345-346.

Lewis, D. 1970. General Semantics. *Synthese* 22: 18-67.

Lewis, R. L. 1996. Architecture Matters: What SOAR has to say about modularity. In D. M. Steier, T. (ed.), *Mind Matters: Contributions to Cognitive and Computer Science in Honor of Allen Newell*. Hillsdale, NJ: Erlbaum

Lewis, R. L. 1999. Cognitive modeling, Symbolic. In R. A. Wilson and F. C. Keil (eds.), *The MIT encyclopedia of the cognitive sciences*. Cambridge, MA: MIT Press

MacLennan, B. 1994. Characteristics of Connectionist Knowledge Representation. *Information Sciences*, vol. 70, pp. 119-143

Mandler, J. 2004. *The Foundations of Mind: Origins of Conceptual Thought*. Oxford: Oxford University Press.

Mandler, J. M. and Bauer, P. J. 1988. The cradle of categorization: Is the basic level basic? *Cognitive Development*, *vol.* 3, Issue 3, pp. 247-264

Marcus, G. F. 2010. Neither size fits all: comment on McClelland et al. and Griffiths et al., *Trends in cognitive sciences*. *vol.* 14, Issue 8, August 2010, pp. 346-347.

Markman, A. B. 1999. *Knowledge representation*. Mahwah, NJ: L. Erlbaum.

Markman, A. B. and Wisniewski, E. J. 1997. Similar and different: The differentiation of basic-level categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23, pp. 54–70.

Markman, E. M. 1991. *Categorization and Naming in Children: Problems of Induction*. Cambridge, MA: MIT Press.

Marr, D. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: W.H. Freeman.

Martin A. and Simmons W.K. 2008. Structural Basis of Semantic Memory. In H. Eichenbaum (ed.), *Memory Systems. Vol. [3] of Learning and Memory: A Comprehensive Reference*. Oxford: Elsevier, pp. 113-130.

Martin, A., Wiggs, C. L., Ungerleider, L. G. and Haxby, J. V. 1996. Neural correlates of category-specific knowledge. *Nature*, 379 (6566), pp. 649–652.

Martin, D.I. and Berry, M.W. 2006. Mathematical foundations behind latent semantic analysis. In Landauer et al. (eds.), *Handbook of latent semantic analysis*. Erlbaum, p.35-55.

Massaro, D. 1988. Some criticisms of connectionist models of human performance. *Journal of Memory and Language*, 27, pp. 213-234.

MATLAB. Mathworks. http://www.mathworks.com/products/matlab/

McCarthy, J. 1968. Programs with Common Sense. In: M. Minsky (ed.), *Semantic Information Processing*, Cambridge, MA: MIT Press

McCarthy, J. and Hayes, P. J. 1969. Some Philosophical Problems from the Standpoint of Artificial Intelligence. In B. Meltzer. and D. Michie (eds.), *Machine Intelligence* 4. Edinburgh University Press, pp. 463-502.

McCauley, R. N. 1998. Levels of Explanation and Cognitive Architectures. In W. Bechtel and G. Graham (eds.), *Companion to Cognitive Science*. Oxford: Blackwell Publishers, pp. 611-624.

McClelland, J. L. 1985. Putting Knowledge in its Place: A Scheme for Programming Parallel Processing Structures on the Fly. *Cognitive Science*, 9, p. 115-128.

McClelland, J. L. 1988. Connectionist Models and Psychological Evidence. *Journal of Memory and Language*, 27, pp. 107-123.

McClelland, J. L. 1999. Cognitive modeling, connectionist. In R. A. Wilson and F. Keil (eds.), *The MIT Encyclopedia of the Cognitive Sciences*. Cambridge, MA: MIT Press, pp. 137-139.

McClelland, J. L. 2010. Emergence in Cognitive Science. *Topics in Cognitive Science*, vol. 2 (2010), pp. 751–770.

McClelland, J. L. and Kawamoto, A. H. 1986. Mechanisms of sentence processing: assigning roles to constituents of sentences. In J. McClelland, D. Rumelhart, and PDP Group (eds.). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 2*. Cambridge, MA: The MIT Press.

McClelland, J. L. and Rumelhart, D. E. 1985. Distributed Memory and the Representation of General and Specific Information. *Journal of Experimental Psychology*: General, 114, pp. 159-188.

McClelland, J. L. et al. 2010. Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in cognitive sciences*. *vol*. 14, Issue 8, August 2010, pp. 348-356.

McClelland, J. L. et al. 2010. Letting structure emerge: Connectionist and dynamical systems approaches to understanding cognition. *Trends in cognitive sciences*, 14, pp. 348-356.

McClelland, J. L., and Elman, J. 1986. The TRACE model of speech perception. *Cognitive Psychology*, 18, pp. 1–86.

McClelland, J. L., Rumelhart, D. E. and PDP Research Group (eds.). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 2*. Cambridge, MA: The MIT Press.

McClelland. J. L. and Rumelhart, D. E. 1981. An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, 88, pp. 375-407.

McDermott, D. 1987. We've Been Framed: Or, Why AI Is Innocent of the Frame Problem. In Z. W. Pylyshyn (ed), *The Robot's Dilemma*. Norwood, N. J.: Ablex.

Medin, D. L. 1989. Concepts and conceptual structure. *American psychologist*, 44(12), pp. 1469-1481.

Medin, D. L. and Waxman, S. 2007. Interpreting asymmetries of projection in children's inductive reasoning. In A. Feeney and E. Heit (eds.), *Inductive reasoning: experimental, developmental, and computational approaches*. Cambridge University Press, pp. 55–80.

Medin, D. L. et al. 1997. Categorization and reasoning among tree experts: Do all roads lead to Rome? *Cognitive Psychology*, 32, pp. 49–96.

Melara, R.D. 1992. The concept of perceptual similarity: from psychophysics to cognitive psychology. In Algom, D. (ed.), *Psychophysical Approaches to Cognition*. Elsevier, Amsterdam, pp. 303-388.

Mervis, C. B., and Rosch, E. 1981. Categorization of natural objects. *Annual review of psychology*, 32(1), pp. 89–115.

Mervis, C. B., Johnson, K. E., and Scott, P. 1993. Perceptual knowledge, conceptual knowledge, and expertise: Comment on Jones and Smith. *Cognitive Development*, 8(2), pp. 149–155.

Miikkulainen, R. 1993. *Subsymbolic Natural Language Processing*. Cambridge MA: The MIT Press.

Millikan, R. G. 1989. Biosemantics. *Journal of Philosophy*, 86, pp. 281-297.

Minsky, M. L. 1956. Some Universal Elements for Finite Automata. *Automata Studies*. Princeton, NJ: Princeton University Press, pp. 117-128.

Minsky, M. L. 1962. Size and Structure of Universal Turing Machines Using Tag Systems. *Proceedings of Symposium in Pure Math*, 5, pp. 229-238.

Minsky, M. L. 1967. *Computation: Finite and Infinite Machines*. Englewood Cliffs: Prentice-Hall.

Minsky, M. L. 1975. A framework for representing knowledge. In P. H. Winston (ed.), *The Psychology of Computer Vision*. New York: McGraw-Hill, pp. 211–277.

Minsky, M.L. 1968. *Semantic information processing*. Cambridge, MA: MIT Press

Montague, R. 1974. *Formal Philosophy: Selected Papers of Richard Montague*. New Haven: Yale University Press.

Moor, J.H. 2000. *Special issues on the Turing test: Past, present and future, Minds and Machines, vol.* 10, No. 4, and *vol.* 11, No. 1.

Murphy, G. L. and Wisniewski, E. J. 1989. Categorizing objects in isolation and in scenes: What a superordinate is good for. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15, pp. 572-586.

Murphy, G.L. 2002. *The Big Book of Concepts*. Cambridge, MA: The MIT Press, 2002

Newell, A. 1980. Physical symbol systems. *Cognitive Science* 4:135-183.

Newell, A. 1990. *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Newell, A. 1992. Precis of Unified theories of cognition. Behavioral and Brain sciences, vol. 15, pp. 425-492

Newell, A. and Simon, H.A. (1976). Computer science as empirical enquiry: Symbols and search. Communications of the Association for Computing Machinery 19(3):113-126.

Newell, A. and Simon, H.A. 1959. *The simulation of human thought*. Santa Monica, California: Rand Corp.

Newell, A. and Simon, H.A. 1972. *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.

Niklasson, L. and van Gelder, T. 1994. Can Connectionist Models Exhibit Non-Classical Structure Sensitivity? In, *Proceedings of the Sixteenth Annual Conference of the Cognitive Science.*

Nosofsky, R. M. 1988. Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(1), pp. 54-65.

Nosofsky, R. M. 1991. Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology*, 23, pp. 94–140.

Partee, B. 1995. Lexical Semantics and Compositionality. In Gleitman and M. Liberman (eds.), *Invitation to Cognitive Science. Part I: Language*, 2nd edition.. Cambridge, MA: MIT Press, pp. 311-360.

Partee, B. H. 1984. Compositionality. In F. Landman and F. Veltman (eds.), *Varieties of Formal Semantics*. Dordrecht: Foris.

Piaget, J. 1952. *The Origin of Intelligence in the Child*. New York: Basic Books.

Piaget, J. 1962. *Play, Dreams, and Imitation in Childhood*. English translation of La formation du symbole chez l'enfant by G. Gattegno and F. M. Hodgson. New York: Norton [French original 1945].

Pinker, S. and Prince, A. 1988. On Language and Connectionism: Analysis of a Parallel Distributed Processing Model of Language Acquisition. *Cognition*, vol. 28, pp. 73-193.

Plaut, D. C. 2003. Connectionist modeling of language: Examples and implications. In M. T. Banich and M. Mack (eds.), *Mind, brain, and language: Multidisciplinary perspectives*. Mahwah, NJ: Erlbaum, pp. 143-167.

Plaut, D. C. and Karmiloff-Smith, A. 1993. Representational development and theory-of-mind computations [Commentary on A. Gopnik, How we know our minds:

The illusion of first-person knowledge of intentionality]. *Behavioral and Brain sciences*, 16, pp. 70-71.

Pollack, J. 1990. Recursive distributed representations. *Artificial Intelligence*, 46, pp. 77-105.

Pollack, J. B. 1988. Recursive auto-associative memory: Devising compositional distributed representations. In *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*. Montreal, Canada, pp. 33-39.

Posner, M. I. Goldsmith, R. and Welton, K. E. 1967. Perceived distance and the classification of distorted patterns. *Journal of Experimental Psychology*, 73, pp. 28-38.

Principal component analysis. 2011, February 29. In Wikipedia, The Free Encyclopedia. Retrieved May 6, 2011, from: http://en.wikipedia.org/w/index.php?title=Principal_component_analysisandoldid=479531658

Proffitt, J. B., Coley, J. D., and Medin, D. L. 2000. Expertise and category-based induction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(4), pp. 811-828.

Pulvermüller, F. 1999. Words in the brain's language. *Behavioral and Brain sciences*, 22, pp. 253-336.

Pulvermüller, F. 2001. Brains reflections of words and their meanings. *Trends in Cognitive Science*, vol. 5, no. pp. 12, 512-524.

Putnam, H. 1975. The meaning of "Meaning". In H. Putnam (ed.), *Mind, Language and Reality*. Cambridge: Cambridge University Press.

Putnam, H. 1981. *Reason, truth, and history*. Cambridge: Cambridge University Press.

Putnam, H. 1988. *Representation and Reality*. Cambridge, MA: MIT Press.

Putnam, H. 1990. *Realism with a Human Face*. Cambridge, MA: Harvard University Press.

Pylyshyn, Z. W. 1984. *Computation and Cognition*. Cambridge, MA: Bradford/MIT Press.

Quillian, M. 1968. Semantic memory. In: M. L. Minsky (ed.), *Semantic Information Processing*, 216-260. Cambridge, MA: MIT Press

Ramsey, W. 2007. *Representation Reconsidered*. Cambridge: Cambridge University Press.

Reed, S. K. 1972. Pattern recognition and categorization. *Cognitive Psychology*, 3, pp. 382-407.

Regier, T. 1996. *The Human Semantic Potential: Spatial Language and Constrained Connectionism*. Cambridge, MA: MIT Press.

Regier, T. and Kay, P. 2009. Language, thought and color: Whorf was half right. Trends in Cognitive Science 13, pp. 439–446.

Rips, L. J., Shoben, E. J. and Smith, E. E. 1973. Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior*, 12, pp. 1-20.

Rogozhin, Yurii. 1996. Small Universal Turing Machines. *Theoretical Computer Science*, vol. 168 (2): 215–240.

Rohrer, T. Image schemata in the brain. In B. Hampe (ed.), *From perception to meaning: Image schemas in cognitive linguistics, Part II*. Berlin: Mouton, pp. 32-36.

Rosch, E. 1973. On the internal structure of perceptual and semantic categories. *Cognitive development and the acquisition of language*, 12, 308.

Rosch, E. 1974. Universals and cultural specifics in human categorization. In R. Breslin, W. Lonner, and S. Boehner (eds.), *Cross-culturai perspective on learning*. London: Sage.

Rosch, E. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General* 104, pp. 192–233.

Rosch, E. 1978a. Principles of categorization. In R. Rosch and B. B. Lloyd (eds.), *Cognition and categorization*. Hillsdale, NJ: Erlbaum.

Rosch, E. 1978b. Prototype classification and logical classification: The two systems. In E. Scholnik (ed), *New Trends in Cognitive Representation: Challenges to Piaget's Theory*. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 73-86.

Rosch, E. and Lloyd, B. B. 1978. *Cognition and categorization*. Lawrence Erlbaum Associates.

Rosch, E. and Mervis, C. B. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, 7(4), pp. 573–605.

Rosch, E. et al. 1976. Basic objects in natural categories. *Cognitive Psychology* 8, pp. 382-439.

Ross, N., Medin, D., Coley, J. D. and Atran, S. 2003. Cultural and experiential differences in the development of folkbiological induction. *Cognitive Development*, 18(1), pp. 25–47.

Rumelhart, D. E. and McClelland, J. L. and the PDP Research Group. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1: Foundations*. Cambridge, MA: MIT Press

Rumelhart, D. E., and McClelland, J. L. 1986. On learning the past tenses of English verbs. In J. L. McClelland, D. E. Rumelhart, and the PDP Research Group (eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. vol. 2: Psychological and biological models*. Cambridge, MA: MIT Press, pp. 216–271.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. 1986. Learning internal representations by error propagation. In J. L. McClelland, D. E. Rumelhart, and the PDP Research Group (eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. vol. 2: Psychological and biological models*. Cambridge, MA: MIT Press, (Chapter 8).

Russell, S., and Norvig, P. 1995. *Artificial Intelligence: A Modern Approach.* Prentice-Hall, Englewood Cliffs, NJ.

Sammon, J.W. Jr. 1969. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, vol. C-18, no. 5, pp. 401-409.

Schank, R. C. and R. P. Abelson. 1977. *Scripts, Plans, Goals and Understanding*. Hillsdale, NJ: Erlbaum.

Searle, J. 1980. Minds, brains and programs. *Behavioral and Brain sciences* 3: 417–457.

Searle, J. 1989. Artificial Intelligence and the Chinese Room: An Exchange. New York Review of Books, 36: 2.

Seidenberg, M. S. and MacDonald, M. C. 1999. A Probabilistic Constraints Approach to Language Acquisition and Processing. *Cognitive Science*. *vol.* 23, 4, pp. 569–588.

Shannon, C. E. 1956. A Universal Turing Machine with Two Internal States. *Automata Studies*. Princeton, NJ: Princeton University Press, pp. 157-165.

Sharkey, N. E. 1992. The ghost in the hybrid: a study of uniquely connectionist representations. *Artificial Intelligence and the Simulation of Behaviour Quarterly*, 79, pp. 10–16.

Shepard, R.N. 1987. Toward a universal law of generalization for psychological science. *Science* 237, pp. 1317–1323.

Simon, H. A. 1979. Models of Thought, New Haven: Yale University Press.

Sloutsky, V. M. and Fisher, A. V. 2004. Induction and Categorization in Young Children: A Similarity-Based Model. *Journal of Experimental Psychology: General*, 133(2), pp. 166–188.

Sluga, H. 1980. Gottlob Frege. London: Routledge and Kegan Paul.

Smith, E. and Medin, D. L. 1981. *Categories and concepts*. Cambridge, MA: Harvard University Press.

Smith, E. E., Shoben, E. J. and Rips, L. J. 1974. Structure and process in semantic memory: A featural model for semantic decision. *Psychological Review*, 81, pp. 214-241.

Smith, E.E., Osherson, D.N., Rips, L.J. and Keane, M. 1988. Combining prototypes: a selective modification model. *Cognitive Science* 12, pp. 485-527.

Smith, L. B. 1989. *From global similarities to kinds of similarities: the construction of dimensions in development, Similarity and analogical reasoning*. Cambridge University Press, New York, NY, 1989

Smolensky, P 1988. On the Proper Treatment of Connectionism. *Behavioural and Brain Sciences* 11, pp. 1–74.

Smolensky, P. 1987a. *On variable binding and the representation of symbolic structures in connectionist systems*. (Tech. Rep. No. CU-CS-355-87). Boulder: University of Colorado, Department of Computer Science.

Smolensky, P. 1987b. The constituent structure of connectionist mental states: A reply to Fodor and Pylyshyn. *Southern Journal of Philosophy*, 26: 137-163.

Smolensky, P. 1989. Connectionism and Constituent Structure. In Pfeifer R. et al. (eds.), *Connectionism in Perspective*, North, Holland, pp. 3-24.

Smolensky, P. 1990. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46, pp. 159-216.

Speaks, J. 2010. Theories of Meaning. In Zalta, E. (ed.), *The Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/archives/win2010/entries/meaning

St. John, M. F. and McClelland, J. L. 1990. Learning and applying contextual constraints in sentence processing. *Artificial Intelligence*, 46: pp. 217–257.

Stark, R. 1991. Does hybrid mean more than one? *Artificial Intelligence and the Simulation of Behaviour Quarterly*, 78, pp. 8–10.

Stern, D. G. 1995. *Wittgenstein on Mind and Language*. Oxford University Press, Oxford.

Steyvers, M. and Griffiths, T. 2006. Probabilistic topic models. In, Landauer et al. (eds.), *Handbook of Latent Semantic Analysis*, pp. 424–440.

Strle, G. and Marolt, M. 2012. The EthnoMuse digital library: conceptual representation and annotation of ethnomusicological materials. *International Journal On Digital Libraries*. Springer [Online ed.]. doi: 10.1007/s00799-012-0086-z.

Svensson, H. and Ziemke, T. 2005. Embodied representation: What are the issues. *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pp. 2116–2121.

Talmy, L. 1983. How language structures space. In H. Pick and L. Acredolo (eds.), *Spatial orientation: Theory, research, and application*. New York: Plenum Press, pp. 225-282.

Talmy, L. 1988. Force dynamics in language and cognition. *Cognitive Science* 12, pp. 49-100.

Talmy, L. 2000. *Toward a Cognitive Semantics, Vol. 1-2*. Cambridge, MA: MIT Press.

Tanaka, J. W., and Taylor, M. 1991. Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, 23(3), pp. 457–482.

Tannenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M. and Sedivy, J. C. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268, pp. 1632-1634.

Tarski, A. 1956. *The Concept of Truth in Formalized Languages. Logic, Semantics and Metamathematics*. Oxford: Oxford University Press.

Thelen, E. and Smith, L. 1994. *A dynamic systems approach to the development of cognition and action*. Cambridge, MA: MIT Press.

Thompson, E. 2001. Empathy and consciousness. *Journal of Consciousness Studies* 8: 1-32.

Tomasello, M. 2003. *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press

Tomasello, M., Carpenter, M., Call, J., Behne, T. and Moll, H. 2005. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain sciences* 28: 675-735.

Touretzky, D. S. 1986. Representing and transforming recursive objects in a neural network, or "Trees do grow on Boltzmann machines." In *Proceedings of the 1986 IEEE Conference on Systems, Man and Cybernetics*. Atlanta, GA.

Touretzky, D. S. and Hinton, G. E. 1988. A distributed connectionist production system. *Cognitive Science*, 12: 423-466.

Trends in cognitive sciences. August, 2010. Approaches to cognitive modeling. *vol.* 14, Issue 8, pp. 339-388.

Trends in cognitive sciences. July, 2006. Probabilistic models of cognition. Special issue, *vol.* 10, Issue 7, pp. 287-344.

Tugendhat, E. 1970. The Meaning of 'Bedeutung' in Frege. *Analysis* 30: 177-189.

Turing, A. M. 1938. Correction to: 'On Computable Numbers, with an Application to the Entscheidungsproblem'. In *Proceedings of the London Mathematical Society Series*, 2 (43), pp. 544-546.

Turing, A. M. 1948. 'Intelligent Machinery'. National Physical Laboratory Report. In: Meltzer, B. and Michie, D. (eds). 1969. *Machine Intelligence 5*. Edinburgh: Edinburgh University Press.

Turing, A.M. 1936. On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society Series*, 2 (42) (1936-37): 230-265.

Turing, A.M. 1947. Lecture to the London Mathematical Society on 20 February 1947. In: B. E. Carpenter and R. W. Doran (eds). 1986. *A.M.Turing's ACE Report of 1946 and Other Papers*. Cambridge, MA: The MIT Press.

Turing, A.M. 1950. Computing machinery and intelligence. *Mind* 59, pp. 433–460.

Tversky, B. and Hemenway, K. 1984. Objects, parts, and categories. *Journal of Experimental Psychology: General*, 113, pp. 169-197

van Gelder, T. 1990. Compositionality: A connectionist variation on a classical theme. *Cognitive Science*, 14(3), pp. 355-384.

van Gelder, T. 1995. What Might Cognition Be, If Not Computation? *Journal of Philosophy*, 92(7), pp. 345-381.

van Gelder. 1998. The Dynamical Hypothesis in Cognitive Science. *Behavioral and Brain sciences,* 21, pp. 615-665.

Varela, F. J., Thompson, E., and Rosch, E. 1991. *The embodied mind: Cognitive science and human experience.* MIT press Cambridge, MA.

Waskan, J. and Bechtel, W. 1997. Directions in Connectionist Research: Tractable Computations Without Syntactically Structured Representations. *Metaphilosophy*. 28 (1-2): 31-62.

Weisstein, E. W. Voronoi Diagram. From: *MathWorld--A Wolfram Web Resource*. http://mathworld.wolfram.com/VoronoiDiagram.html

Wheeler, M. 1994. From Activation to Activity: Representation, Computation and the Dynamics of Neural Network Control Systems. *Artificial Intelligence and Simulation of Behaviour Quarterly*, 87, 1994, pp. 36-42.

Wheeler, M. 2005. *Reconstructing the Cognitive World: The Next Step*. Cambridge, MA: MIT Press.

Wheeler, M. 2008. Cognition in Context: Phenomenology, Situated Robotics, and the Frame Problem. *International Journal of Philosophical Studies*, vol. 16(3), pp. 323–349.

Winograd, T. and Flores, F. 1987. *Understanding Computers and Cognition: A New Foundation for Design*. New Jersey: Addison-Wesley.

Wisniewski, E. J. and Medin, D. L. 1994. On the interaction of theory and data in concept learning. *Cognitive Science*, 18(2), pp. 221–281.

Wittgenstein, L. 1922. *Tractatus Logico-Philosophicus*. (English translation by C.K. Ogden and F.P. Ramsey). Routledge and Kegan Paul: London.

Wittgenstein, L. 1953. *Philosophical Investigations*. (English translation by G.E.M. Anscombe). Blackwell: Oxford.

Zeimpekis, D. and Gallopoulos, E. 2005. *TMG: A MATLAB Toolbox for Generating Term-Document Matrices from Text Collections*. Mathworks. http://www.mathworks.com/products/matlab/

Ziemke, T., Zlatev, J. and Frank, R. 2007. *Body, Language and Mind – vol. 1: Embodiment*. Berlin, Mouton de Gruyter.

Zlatev, J. 2005. What's in a schema? Bodily mimesis and the grounding of language. In Hampe B. (ed.), *From Perception to Meaning: Image Schemas in Cognitive Linguistics*. Berlin: Mouton, pp. 313-343.

Zlatev, J. 2007. Intersubjectivity, Mimetic Schemas and the Emergence of Language. *Intellectica* (46-47), pp. 123-151.

20 Appendix: Data and projections[53]

20.1 LDA: topics, top words and documents

Table 3: most salient words per topic (10 topics)

Topic 1: Classical Artificial Intelligence (artificial life, GOFAI)	gofai, feigenbaum, lenat, dendral, creativity, babbage, prolog, mccorduck, homo, aaron, cyc, alife, kurzweil, technological, eliza, boden, weizenbaum, weir, hodgson, humanlike, bipedalism, hacker, colby, artificialintelligence ...
Topic 2: Cognitive Science (categorization, modeling, prototypes, exemplars)	nosofsky, medin, tversky, goldstone, barsalou, exemplars, ratings, typicality, categorization, rips, shepard, mervis, judgments, prototype, kruschke, hampton, markman, metric, similarity, stimuli, exemplarbased, gentner, rated, trials, category, minda, multidimensional, rosch, prototypes, featural, ...
Topic 3: Cognitive Psychology (categorization, infants, children, novices)	gleitman, waxman, novices, chi, yearolds, coley, wellman, gelman, preferences, mervis, osherson, infants, montholds, thematic, superordinate, subcategorization, markman, basiclevel, transitivity, categorization ...
Topic 4: Various (cognitive development, anthropology, linguistics, ...)	hauser, baillargeon, chimpanzees, preverbal, autism, leslie, spelke, tomasello, wynn, creole, containment, boesch, societies, byrne, sperber, infants, crosslinguistic, tamarins, montholds, reiter, anthropology, ontogeny, infancy, gardner, hirschfeld, adulthood, semiotic, boroditsky, primates, lieberman, piaget, anthropologists, abductive, prosodic, ...
Topic 5: Neuroscience (cortex, premotor, patients)	gyrus, fusiform, neuroscience, medial, lateral, ventral, fmri, lobe, sulcus, prefrontal, categoryspecific, imaging, parietal, neuropsychologia, lesions, anterior, neuroimaging, cortex, cortical, premotor, bilateral, deficits, occipitotemporal, psychological, selectivity, patients, sts, simmons, ...
Topic 6: Philosophy of Mind (intentionality, subjectivity, metaphysics, logic, epistemology)	causation, intentionality, kant, hume, epistemic, phenomenal, metaphysical, intentional, constitutive, materialism, qualia, leibniz, crane, supervenience, objectivity, sellars, tye, reductive, descartes, quantum, constituted, subjectivity, intrinsic, logic, philosophyofmind, brentano, intersubjective, wittgenstein, consciousness, teleological, metaphysics, transcendental, dretske, ...
Topic 7: Emulation, Cognition, Memory	emulator, efference, latency, tulving, anterograde, egocentric, oscillations, spc, amnesia, emulation, grush, golgi, shock, autonoetic, recollection, circadian, paralysis, oscillatory, thalamocortical, broca, sherrington, oscillation, ephemeris, endogenous, occipital, navigation, proprioceptive, rotational, suppression, erent, kihlstrom, anesthetized, ...
Topic 8: Traditional models (Computationalism & Connectionism; Cognitive modeling: symbolic vs. connectionist; representations, compositionality)	connectionist, fodor, connectionism, constituents, pylyshyn, syntactic, sentences, distributedrepresentations, rumelhart, compositionality, systematicity, smolensky, semantics, churchland, elman, representational, neuralnetworks, symbol, network, grammar, tokens, symbols, output, architecture, machines, newell, computers, raam, mcclelland, connectionists, symbolic, turing, program, mental, hinton, turingmachine, rules, ...
Topic 9: Topic Models (natural language processing)	lda, dirichlet, blei, document, variational, topicmodels, latent, bayesian, griffiths, markov, posterior, topics, distributions, graphical, jordan, steyvers, corpus, lsa, distribution, multinomial, unsupervised, parameters, probabilistic, corpora, topic, generative, gibbs, algorithm, latentdirichletallocation, sampling, probability, hofmann, optimization, dumais, mixture, gaussian, latentsemanticanalysis, statistics, predictive, landauer, matrix, bayes, nonparametric, ...
Topic 10: Cognitive Semantics (image schemas, conceptual spaces, quality dimensions)	conceptualspaces, harnad, qualitydimensions, cognitivesemantics, metaphors, Gärdenfors, grounding, lakoff, symbolsystem, metaphorical, metaphor, imageschemas, voronoi, topological, symboliclevel, imageschematic, sensorimotor, soar, tessellation, newell, emotions, bickhard, grounded, intensional, rationality, metaphoric, neuralnetworks, operator, constructions, objectivist, turing, pragmatics, ungrounded, expressions, fauconnier, robotic, worlds, symbol, schemas, robot, ontology, invariants, functioning, ...

Table 4: Top documents per topic (10 topics)

Topic 1: Artificial Intelligence (artificial life, GOFAI)	b-Boden-Artificial Intelligence; b-Boden-Mind As Machine A History Of Cognitive Science; Henderson-Ai Mirrors For The Mind
Topic 2: Cognitive Science (categorization, modeling, prototypes, exemplars)	Barsalou-The instability of graded structure; Collins&Quillian-Does Category Size Affect Categorization Time; Goldstone&Medin&Halberstadt-Similarity in context; Gregory-Knowledge in perception; Hampton&Moss-Concepts and meaning; Zaki&Nosofsky-Prototype and Exemplar Accounts of Category Learning

[53] Lists of top words for LDA, LSA and pLSA (Table 3, 6-8) are presented, with more salient words in bold.

Topic 3: Cognitive Psychology (categorization, infants, children, novices)	b-Gelman-Perceptual and Cognitive Development; Mandler-Perceptual and Conceptual Processes in Infancy
Topic 4: Various (cognitive development, anthropology, linguistics, ...)	b-Mehler-Language, Brain, And Cognitive Development; b-Deacon-Symbolic Species; b-Mandler-The Foundations Of Mind Origins Of Conceptual Thought
Topic 5: Neuroscience (cortex, premotor, patients)	Caramazza-Domain-Specific Knowledge in the Brain; Martin-The Representation of Object Concepts in the Brain; Martin&Chao-Semantic memory and the brain-structure and processes; b-Neural Basis Of Semantic Memory; McNamara-Cognitive maps and the hippocampus
Topic 6: Philosophy of Mind (intentionality, subjectivity, metaphysics, logic, episteology)	b-Essential Sources in the Scientific Study of Consciousness; b-Introduction to the Science and Philosophy of Mental Imagery; b-Ramsey-Representation Reconsidered; Chalmers-Facing Up to the Problem of Consciousness; b-Crane-The Mechanical Mind; Dennett Intentional systems; Putnam-The Meaning of Meaning
Topic 7: Emulation, Cognition, Memory	b-Embodied Minds in Action; b-Waskan-Models And Cognition Prediction And Explanation In Everyday Life And In Science; Grush-Emulation and Cognition; Grush-The emulation theory of representation-Motor control, imagery, and perception; Menon-Relating semantic and episodic memory systems; Tulving-What is Episodic Memory
Topic 8: Traditional models (Computationalism & Connectionism; Cognitive modeling: symbolic vs. connectionist; representations, compositionality)	Bechtel-Levels of description and explanation in cognitive science; Chalmers-Syntactic Transformations on Distributed Representations; Chalmers-Why Fodor and Pylyshyn Were Wrong; Niklasson-Connectionism and the Issues of Compositionality and Systematicity; Fodor&Pylyshyn-Connectionism and Cognitive Architecture-A Critical Analysis; Rumelhart&Mcclelland-On learning past-tense of english verbs; Van Gelder_What Might Cognition Be, If Not Computation; Waskan&Bechtel-Connectionism and Cognitive Linguistics
Topic 9: Topic Models (natural language processing)	Blei&Lafferty-Topic Models; Blei&Lafferty-Dynamic Topic Models; Blei&McAuliffe-Supervised Topic Models; Blei-Introduction to Probabilistic Topic Models; Griffiths et al.-Topics in Semantic Representation; Hoffman et al.-Finding Latent Sources in Recorded Music; Steyvers&Griffiths-Probabilistic Topic Models
Topic 10: Cognitive Semantics (image schemas, conceptual spaces, quality dimensions)	b-Langacker-Cognitive Grammar An Introduction; b-Gärdenfors-Conceptual Spaces: The Geometry of Thought; Gärdenfors-Conceptual Spaces as a Basis for Cognitive Semantics; Gärdenfors-Meanings As Conceptual Structures; Gärdenfors-Mental Representation, Conceptual Spaces and Metaphors; Harnad-The Symbol Grounding Problem; Lakoff-Cognitive models of categorization; b-Lakoff-Women, Fire And Dangerous Things

Table 5: LDA: List of document titles (see Figure 19 in the main section)

Topic 1: Artificial Intelligence (artificial life, GOFAI)	Boden-Artificial Intelligence (book)
Topic 2: Cognitive Science (categorization, modeling, prototypes, exemplars)	Johnson-Prototype Theory, Cognitive Linguistics and Pedagogical Grammar (article); Tversky-Features of Similarity (article)
Topic 3: Cognitive Psychology (categorization, infants, children, novices)	Gelman-Perceptual and Cognitive Development (book)
Topic 4: Various (cognitive development, anthropology, linguistics, ...)	Mandler-The Foundations Of Mind: Origins Of Conceptual Thought (book)
Topic 5: Neuroscience (cortex, premotor, patients)	Hart&Kraut-Neural Basis Of Semantic Memory (book)
Topic 6: Philosophy of Mind (intentionality, subjectivity, metaphysics, logic, episteology)	Haugeland-Having Thought: Essays in the metaphysics of mind (book)
Topic 7: Emulation, Cognition, Memory	Tulving-What is Episodic Memory (article); Grush-Emulation and Cognition (article)
Topic 8: Traditional models (Computationalism & Connectionism; Cognitive modeling: symbolic vs. connectionist; representations, compositionality)	Fodor&Pylyshyn-Connectionism And Cognitive Architecture: A Critical Analysis (article)
Topic 9: Topic Models (natural language processing)	Blei&Lafferty-Topic Models (article)
Topic 10: Cognitive Semantics (image schemas, conceptual spaces, quality dimensions)	Gärdenfors-Conceptual Spaces: The Geometry Of Thought (book); Strle-Semantics within: Representation Of Meaning Through Conceptual Spaces (thesis)

Table 6: LDA: 30 top words per topic (30 topics)

Topic 1	**mindasmachine**, cricket, bourdieu, crickets, theyd, **babbage**, **nlp**, cussins, hed, **alife**, borges, grossberg, geertz, governments, gazdar, aisb, priests, fathers, **neuron**, jespersen, vaucanson, mays, awe, licklider, repr, wars, pask,

	hartley, behaviourism, humboldt
Topic 2	**lda**, **dirichlet**, **blei**, **variational**, **documents**, **document**, **topicmodels**, **topicmodel**, **latent**, **lsa**, **log**, **markov**, **steyvers**, **posterior**, **distributions**, **griffiths**, **multinomial**, **topics**, **corpus**, **bayesian**, **corpora**, **graphical**, **latentdirichletallocation**, **distribution**, **unsupervised**, **lafferty**, **jordan**, **landauer**, **dumais**, **topic**
Topic 3	**cognitivesemantics**, **imageschemas**, **lakoff**, **trajector**, **metonymic**, **imageschematic**, icm, **schemas**, **prototypetheory**, **imageschema**, zebra, **fauconnier**, **constructions**, **Gärdenfors**, **conceptualspaces**, mat, **metaphor**, backgrounded, **prototype**, **container**, icms, **langacker**, **hampton**, **grounding**, beep, sweetser, infinitival, **hybrid**, **metonymy**, **metaphorical**
Topic 4	reitman, halford, cannon, **actr**, quartercentury, **schizophrenia**, novick, **creativity**, perkins, bassok, jurisprudence, **clinicians**, polya, **judicial**, **individualistic**, salthouse, wason, duncker, disessa, **kahneman**, **lawyers**, barth, maier, **sloman**, bowden, hogarth, greenfield, kruglanski, frederick, diagrammatic
Topic 5	**handaxes**, **flakes**, flake, **erectus**, **handaxe**, **palaeolithic**, levallois, nests, habilis, knapping, bifaces, **archaeological**, **bones**, **archeological**, **sortals**, **homo**, **genus**, **australopithecines**, **hunting**, endocasts, **neanderthal**, afarensis, archaeology, mithen, mellars, **africa**, neanderthals, **lithic**, **nest**, oldowan
Topic 6	**creativity**, **mozart**, **minda**, mse, poincarã, coleridge, benzene, thagard, hamiltonian, stahl, **escher**, **prototypebased**, **galileo**, emmy, **orbit**, mckinley, boyle, passim, nls, lavoisier, serendipity, **arcs**, **bosons**, sellars, **kepler**, **schoenberg**, chs, riemannian, dunham, dalton
Topic 7	squire, oxidative, **tulving**, **episodic**, **hippocampus**, fermentation, **hippocampal**, atp, okeefe, phosphorylation, **ferrier**, **autonoetic**, **amnesia**, wimsatt, **lobe**, mishkin, endel, interlevel, parietal, enzymes, darden, roediger, dissociations, **medial**, hera, schaffner, **lesions**, inquiries, **navigation**, **maplike**
Topic 8	**categoryspecific**, warrington, **fusiform**, **caramazza**, **chao**, **gyrus**, **dementia**, **hodges**, **deficits**, **modalityspecific**, **impairment**, capitani, gainotti, haxby, **amodal**, impaired, **neuroimaging**, **damasio**, **patients**, **lobe**, naming, **nonliving**, **neuropsychologia**, **martin**, **medial**, **frontal**, **tranel**, **anterior**, **fmri**, shelton
Topic 9	**chemero**, **goldstone**, rooij, **markman**, **antirepresentationalism**, **gentner**, **grounding**, **harnad**, **varela**, **carello**, hutto, **medin**, **brooks**, bressler, **nonsymbolicrepresentations**, **icons**, **gelder**, **representationhungry**, **categorical**, **regier**, grossman, **gibson**, **pss**, **antirepresentationalists**, affordances, **haselager**, **cerebellum**, **contextindependent**, tuller
Topic 10	**raam**, **systematicity**, **distributedrepresentations**, **smolensky**, **pollack**, **structuresensitive**, connectionists, **connectionism**, **compositionality**, **microfeatures**, **symbolsystem**, **pylyshyn**, **distributedrepresentation**, **governor**, **connectionist**, **concatenative**, **fodor**, **gelder**, **backpropagation**, **robot**, hadley, **dynamical**, **symbolic**, **level**, **combinatorial**, **haugeland**, **chalmers**, **constituents**, **layer**, **connectionistnetworks**
Topic 11	**metric**, **tversky**, **convex**, **conceptualspaces**, **additivity**, **triangleinequality**, **qualitydimensions**, **conceptualspace**, **topological**, **voronoi**, krantz, beals, **multidimensional**, **additive**, canary, **ratings**, **dissimilarity**, **dimensions**, **euclidean**, **Gärdenfors**, dimension, msec, **scaling**, **ordinal**, **categorizations**, **judgments**, **betweenness**, **similarity**, **tessellation**, **rosch**,
Topic 12	**dualism**, **intentionality**, **qualia**, **searle**, **materialism**, **weizenbaum**, **materialists**, **desires**, **consciousness**, **soul**, **cyc**, **nonreductive**, **intentional**, **turing**, hunk, **physicalism**, url, **functionalism**, **sartre**, **eliza**, **selfconscious**, efficacious, **brentano**, **chessplaying**, **reductive**, **turingmachine**, **hollow**, **singularity**, **believes**, **desire**,
Topic 13	**rayner**, **prelexical**, **cutler**, **mcqueen**, **norris**, **frauenfelder**, marslenwilson, frazier, otake, **distractor**, **acoustical**, **dyslexia**, **highfrequency**, demuth, connine, **khz**, **toefl**, **lowfrequency**, **nonwords**, **dyslexics**, garnham, catalan, **priming**, deese, **seidenberg**, **sublexical**, christophe, basilar, wagemans, linebarger
Topic 14	**phenomenality**, tarahumara, **dennett**, whorf, firstperson, **soar**, **fluent**, heated, **qualia**, **experiential**, **tye**, zombies, **recognitional**, shouldnt, doesnt, **band**, **economics**, isnt, jam, mixtec, dont, **axioms**, paragons, comforting, didnt, **informationprocessing**, nicod, baars, ofthe, chapter
Topic 15	beneficiary, pfc, **antisocial**, **impulsive**, **heine**, ajh, acc, henrich, **ultimatum**, **forcedynamic**, **volitional**, **violent**, **punishment**, **impersonal**, **mood**, **liking**, ict, **weird**, **adjunct**, periphrastic, lcs, **serotonin**, ofc, kick, americans, **selfinterest**, **willful**, **sue**, **paraphrases**, **pleasures**
Topic 16	wcs, bestfit, ima, covis, **animat**, agr, ecg, **creole**, lefebvre, filoteo, tessellation, oracle, **kay**, informant, sing, kdd, cafã, **categorial**, yucatec, apriority, **amoeba**, aho, **expressiveness**, munsell, montrã, **usability**, turkish, **pronoun**, **conceptualspaces**, lillomartin
Topic 17	**mcclelland**, psychol, sci, **tics**, **hinton**, **rumelhart**, rev, **connectionist**, **units**, **processing**, ofthe, **nets**, **weights**, **pdp**, **minsky**, **backpropagation**, **papert**, **bayesian**, zerocrossings, **networks**, **trends**, fig, **rolespecific**, unit, **detectors**, **probabilistic**, **marr**, **graded**, **ullman**, **sejnowski**
Topic 18	**erent**, **reticular**, erences, laberge, **erence**, ect, baars, anosognosia, ective, **thalamocortical**, clin, ror, ects, fringe, steriade, cient, **codelets**, culty, **libet**, **neurophysiology**, spc, slowwave, **hypnotic**, **anaesthesia**, **anesthesia**, **analgesia**, scheibel, **thalamic**, **awakening**, **wakefulness**
Topic 19	**intransitivities**, wst, ect, **erent**, **subadditivity**, **erence**, **scalability**, sophia, erences, gambles, sattath, intransitivity, gamble, **reasonbased**, ects, redelmeier, **overconfidence**, gri, majors, quattrone, ered, wakker, slovic, ense, thaler, voters, manslaughter, **aversion**, culty, rottenstreich,
Topic 20	**glasersfeld**, **gallistel**, **oscillators**, **transmitter**, **oscillation**, **endogenous**, nectar, stein, lambda, **maze**, bearings, **anticipatory**, **enaction**, fourier, elapsed, bee, oscillator, **environment**, thesame, deg, circadian, css, onion, reckoning, scorpion, harth, hive, ephemeris, csus, **stabilities**
Topic 21	pelke, **gelman**, **baillargeon**, **carey**, **domaingeneral**, **preschool**, **preschoolers**, **infants**, landau, hirschfeld, **autism**, premack, **wellman**, **meltzoff**, tagerflusberg, bartsch, **conservation**, leslie, bellugi, **baroncohen**, flavell, baldwin, **domainspecific**, crain, hermer, **karmiloffsmith**, **montholds**, wynn, **reorientation**, **infant**
Topic 22	**stereogram**, morais, trehub, randomdot, drake, **eccentricity**, **roughness**, **epithelium**, slant, **nonspeech**, **directionally**, offcenter, redgreen, **grating**, **contour**, **deutsch**, **spectral**, **transparency**, julesz, jusczyk, monocular,

	mcadams, regan, **nonmusicians**, hildreth, acuity, sekuler, occluding, knudsen, **englishlearning**
Topic 23	**elman**, **tense**, **newport**, **irregulars**, **plunkett**, **symbolmanipulating**, **pinker**, **regulars**, **inflection**, **bates**, **recurrent**, **symbolmanipulation**, axons, **marchman**, terrence, spinal, **synaptic**, **subjectpredicate**, shastri, marcus, macwhinney, **turing**, **rewiring**, **savagerumbaugh**, christiansen, **multilayer**, **searle**, **kanzi**, **sherman**, **verbs**
Topic 24	**physicalsystem**, **quantum**, **interpreter**, **phenomenal**, **rationality**, **conscious**, **newell**, **searle**, ips, **algorithmic**, **metaphysical**, **successor**, **dretske**, **commonsense**, **laws**, **functioning**, **explanatory**, **discoveries**, **dreyfus**, **consciousness**, **desires**, **marr**, **humanlike**, **mechanics**, byandlarge, **intentional**, **administrative**, writers, chapter, causal
Topic 25	**emergentist**, provisos, **scalemodel**, **magnetosome**, inï, **sentential**, **explanans**, keplers, **nonsentential**, singer, forethought, icm, enlightening, bucket, **johnsonlaird**, **friston**, byrne, premises, doxastic, gauntlet, **modem**, **deductivenomological**, **explanandum**, **nonconcrete**, **imagery**, **kant**, **waskan**, engel, **hume**, **uncertainties**
Topic 26	**emulator**, emulators, **emulation**, **controller**, **grush**, **musculoskeletal**, kawato, **articulants**, **efferent**, **closedloop**, **offline**, **proprioceptive**, **effector**, **motorcontrol**, mel, **tendon**, **imagery**, decety, gon, **bickhard**, wolpert, mss, calvo, plant, **sensor**, jeannerod, **pseudoclosedloop**, **muscle**, mock, torque
Topic 27	**area**, **wernickes**, **stereotyping**, whmovement, **diachrony**, benthem, marcia, diss, degraff, **brocas**, **language**, sapir, dels, areas, **creole**, **creolization**, andrade, nrem, ltp, creoles, **grammar**, malinowski, raiffa, reinhart, albright, geschwind, thaler, **thornhill**, headdriven, **memory**
Topic 28	**altriciality**, **autistic**, **husserl**, **falsebelief**, **vowel**, **autism**, **archaic**, **abnormalities**, **psychologism**, **hobson**, credulous, **incomprehensible**, bracketing, **zahavi**, **wasons**, **syndrome**, lived, admissible, **intersubjectivity**, **interpretive**, **phenomenologists**, **neurogenesis**, **leibnitz**, lorenz, **dialogues**, **prepotent**, marian, **syndromes**, **heritability**, frith
Topic 29	**constituent**, **lepore**, **holism**, dogs, arent, **compositionality**, **constituents**, doorknob, weve, rtm, **redescription**, **individuation**, **compositional**, **primitive**, **frege**, fish, correspondingly, chairs, tendentious, doorknobs, **karmiloffsmith**, **evaluable**, cats, bachelor, dog, **syntactic**, **prototypes**, **containment**, facto, **productivity**
Topic 30	**heidegger**, **dreyfus**, **externalism**, **epistemic**, **merleauponty**, **intentionality**, **phenomenology**, **dennett**, **encodings**, **bickhard**, **varela**, notebook, **situated**, inga, **haugeland**, otto, **representationality**, **embodiment**, **morse**, **subpersonal**, **gallagher**, **potentialities**, **interactivism**, **encodingism**, **interfaces**, **scaffolding**, **hermeneutics**, **cyc**, campbell, **agent**

22.2 LSA and pLSA: dimensions/topics and top words

Table 7: LSA: top 30 words per topic (first 10 dimensions)

Topic 1	**patients**, **infants**, **frontal**, **deficits**, **cortical**, chapter, **emotional**, **lobe**, **monkeys**, **cortex**, **monkey**, **parietal**, **prefrontal**, executive, emotions, emotion, **disorders**, **clinical**, **cells**, **hemisphere**, **awareness**, **lesions**, **cultural**, **culture**, **deficit**, **kahneman**, **damasio**, **wilson**, **rats**, **cohen**
Topic 2	theyd, hed, **babbage**, passim, **mindasmachine**, masterman, mccorduck, **gofai**, hadnt, **lovelace**, **nlp**, goethe, weizenbaum, darcy, **alife**, werent, emmy, vaucanson, pask, **neokantian**, quartercentury, chs, shrdlu, **behaviourism**, hype, repr, **behaviourist**, **cyc**, gelernter, **feigenbaum**
Topic 3	chapter, scenario, **basiclevel**, chapters, **superordinate**, **barsalou**, dog, dogs, dont, **categorical**, **items**, **child**, ahn, event, **nominals**, water, car, **children**, **intentionally**, **agent**, someone, phase, **animals**, ball, **medin**, containers, **belief**, book, bird, count
Topic 4	**hypnotic**, **analgesia**, **erent**, soa, lemniscus, suprathreshold, mcadams, grossberg, baars, **neurons**, analgesic, fringe, referral, **anaesthesia**, refractory, pandemonium, masking, **metacontrast**, longuethiggins, pedestal, **hypnosis**, **harmonics**, erences, eriksen, repp, sperling, **melodies**, buds, grating, **cuing**
Topic 5	**embodiment**, **dretske**, **robot**, **kelso**, **churchland**, **bickhard**, **constitutive**, prosodic, clausal, **pronoun**, **robots**, **putnam**, **intentionality**, **sellars**, **haugeland**, **millikan**, **langacker**, **embodied**, **metaphysical**, surfaces, **disembodied**, roughness, **gibson**, **dynamical**, **controller**, **naturalism**, **robotics**, **varela**, **philosophyofmind**, locative
Topic 6	**erent**, erences, erence, clin, steriade, **neurophysiology**, **anosognosia**, ect, bogen, ects, slowwave, ective, **autonoetic**, awakenings, **libet**, **unconsciousness**, rightness, codelets, culty, **rem**, nmda, neuron, nig, commissurotomy, mangan, cholinergic, baars, **waking**, nrem, **abstract**
Topic 7	**categoryspecific**, **caramazza**, **warrington**, **fusiform**, **chao**, **modalityspecific**, forde, gainotti, capitani, hodges, laiacona, **dementia**, **haxby**, **gyrus**, **nonliving**, **shelton**, **impairment**, hillis, **moss**, **deficits**, **silveri**, **neurocase**, lambon, **barbarotto**, buxbaum, garrard, tranel, gornotempini, **aphasia**, mahon,
Topic 8	**hauser**, **chimpanzees**, **bickhard**, **thelen**, dunbar, klahr, **accumulator**, **autistic**, perner, boysen, **syndrome**, deloache, **chalmers**, ceci, **meltzoff**, novick, **vygotsky**, **primates**, **chimpanzee**, **affective**, **piaget**, numerosities, troglodytes, engle, brannon, **cortex**, **clancey**, bassok, siegal, cordes
Topic 9	**categoryspecific**, riddoch, garrard, lambon, **subadditivity**, hja, **moss**, **caramazza**, **barbarotto**, capitani, funnell, laiacona, hillis, **vegetables**, wordpicture, **nonliving**, **neurocase**, silveri, hse, artefacts, sft, overconfidence, breedin, vignettes, **hodges**, devlin, forde, sartori, cipolotti, fishburn
Topic 10	**spelke**, **montholds**, **sexual**, **wynn**, **mothers**, nongeometric, knudsen, **gelman**, **handaxes**, **yearolds**, baillargeon, **males**, **females**, **meat**, **erectus**, **foods**, **mandler**, **familiarization**, **cage**, yonas, meck, gould, **taxonomic**, **preschool**, **preschoolers**, tectum, **flakes**, **bipolar**, chamber, dow

Table 8: pLSA: top 30 words per topic (10 topics)

Topic 1	actually, **scientific**, arises, **complexity**, three, **abstract**, text, come, assumes, **component**, use, thus, values, ways, figure, **content**, **algorithms**, cases, terms, also, estimates, **solution**, **systems**, years, **topicmodel**, top, unknown, **classification**, **science**
Topic 2	**conceptual**, allow, done, cases, difference, ask, circumstances, element, **cognition**, **behavioral**, actual, **computational**, **direct**, **development**, clear, critical, correct, better, case, **animals**, **architecture**, content, beyond, comments, conclusions, corresponding, available, conditions, degree
Topic 3	thus, **positive**, **correctly**, **identity**, long, provide, note, **functional**, **wilson**, finally, **level**, **recall**, set, will, point, remains, second, **visual**, **model**, see, **studies**, show, take, place, **represented**, press, right, specify, **system**
Topic 4	without, thus, well, way, ways, two, **understanding**, **analysis**, together, able, words, used, suggests, work, account, york, university, whereas, also, using, something, **abstract**, **time**, **access**, argued, across, will, working, years
Topic 5	**process**, **position**, **notion**, others, **object**, press, **psychology**, rather, quite, **physical**, **perception**, provided, provide, proceedings, **reason**, number, precisely, potential, **problems**, **phenomena**, one, respect, possible, **research**, **processing**, **relative**, place, see, **relevant**
Topic 6	introduction, **inference**, **natural**, one, press, many, given, mit, instead, include, oxford, ieee, **models**, **maps**, note, **realistic**, point, makes, particular, problem, moreover, **patterns**, make, news, **generated**, **processing**, including, main, **matrix**
Topic 7	know, **internal**, **individual**, **models**, instead, end, many, highly, **interaction**, large, just, **intelligence**, likely, important, foundation, **mental**, **input**, indeed, next, however, **knowledge**, interesting, involved, make, **memory**, **levels**, general, **experience**, **language**
Topic 8	**stored**, **time**, **table**, thus, **subset**, taken, **tasks**, **simple**, **faster**, together, takes, seen, **test**, **understood**, **similar**, **response**, **finding**, various, six, still, **sets**, things, undivided, equal, use, set, unless, first, **structure**
Topic 9	**systems**, **understand**, way, **theories**, various, center, terms, either, work, though, underlying, cover, whether, support, give, will, also, like, two, correct, **analysis**, forbears, use, **time**, many, wrong, fact, thus, **architecture**
Topic 10	examples, detailed, established, **explain**, areas, fourth, come, directly, fundamental, hardly, getting, found, counts, described, form, enters, **fleshed**, essentially, may, **framework**, depends, details, **associated**, final, **functioning**, first, issues, force

22.3 Projections: LSA and pLSA

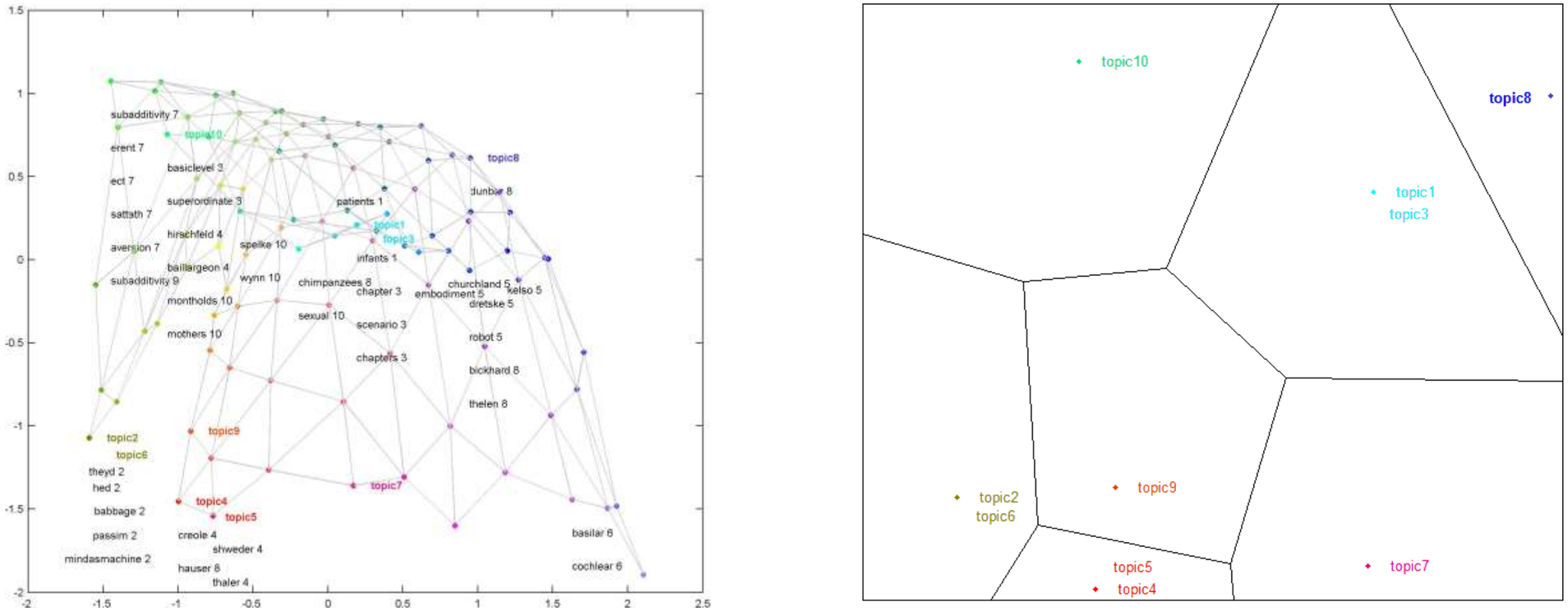


Figure 20: LSA distribution of topics in conceptual space: 3D Mesh (left) and Voronoi tessellation (right)

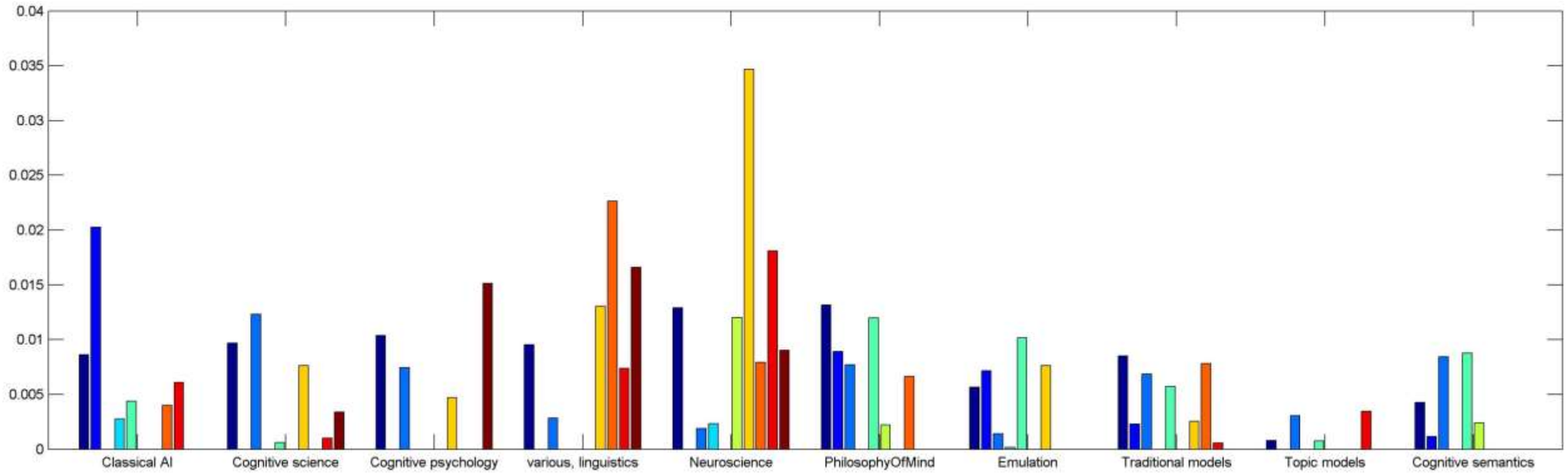


Figure 21: LSA distribution of 10 topics over conceptual space. Most salient member of individual topic is chosen for distribution.

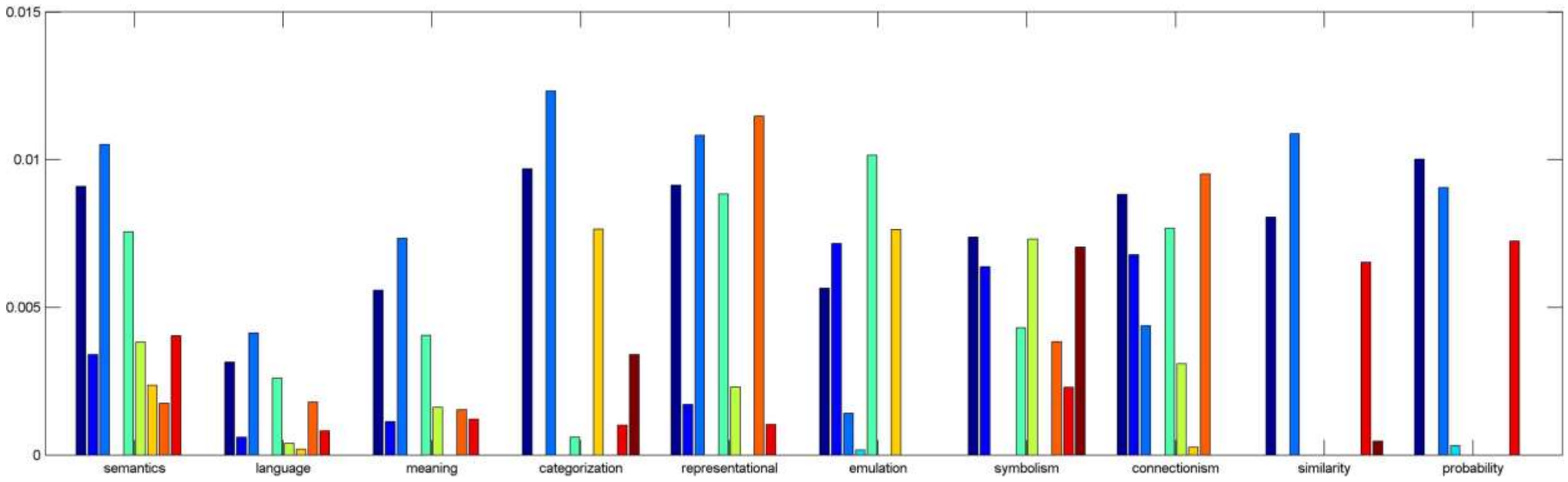


Figure 22: LSA distribution of words/concepts over topics (general)

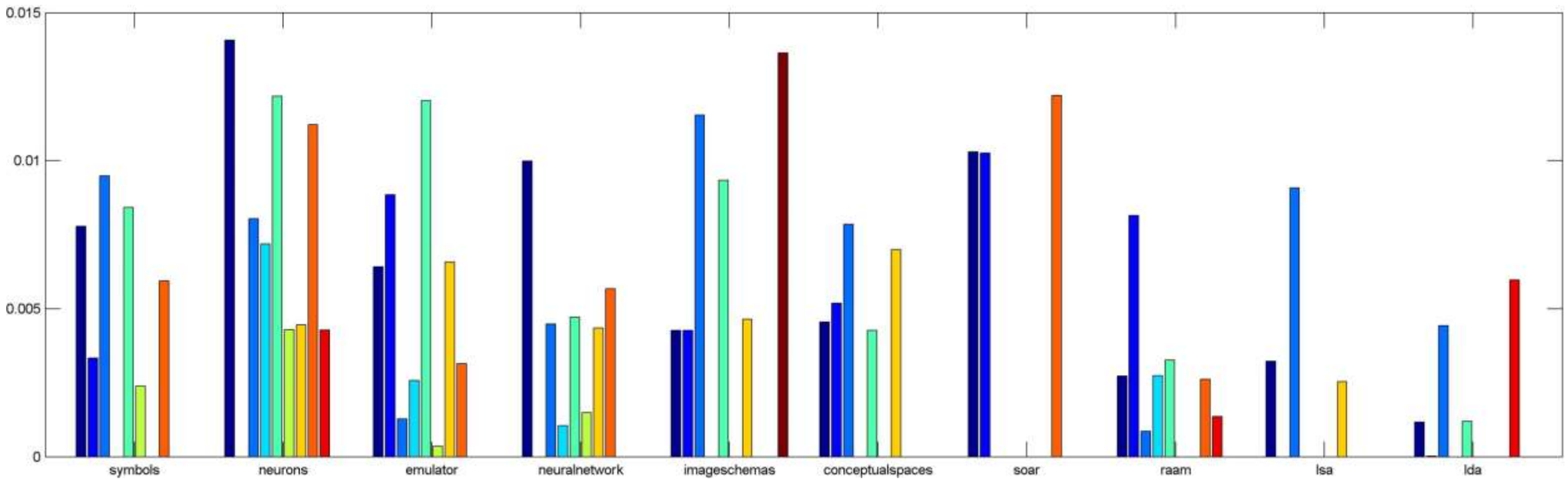


Figure 23: LSA distribution of words/concepts over topics (cognitive modeling)

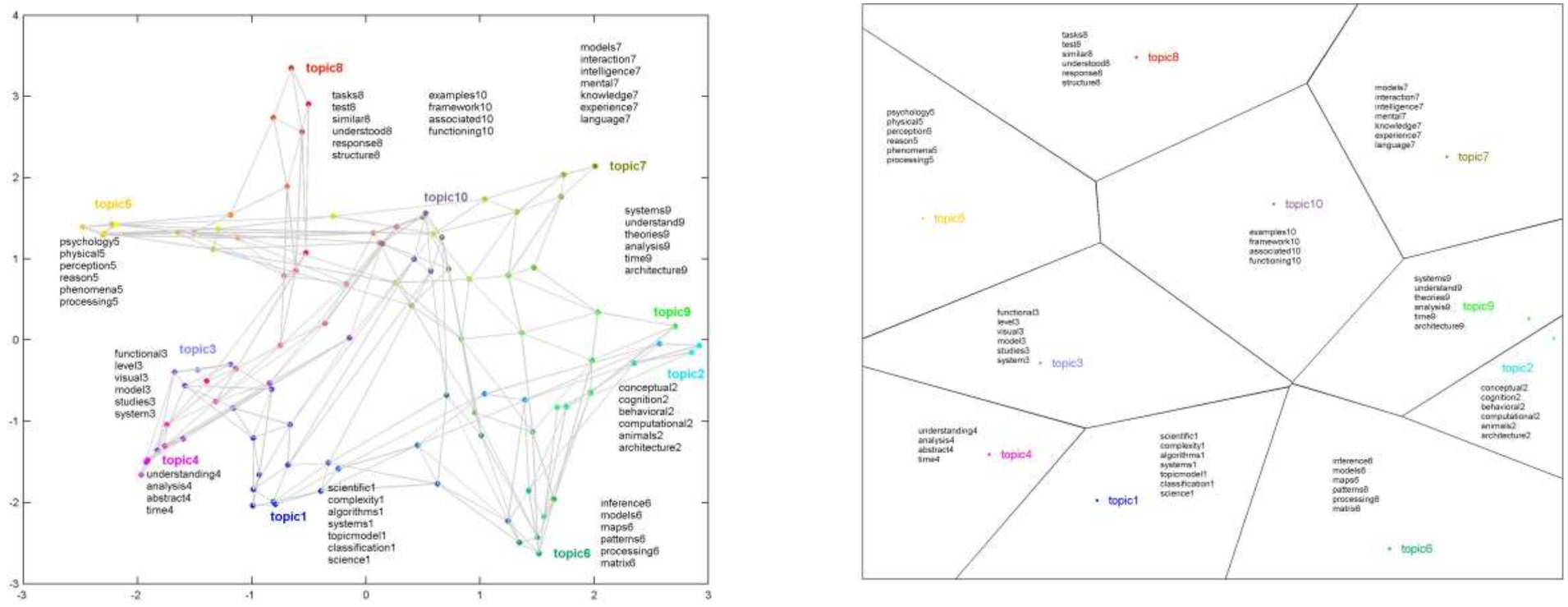


Figure 24: pLSA distribution of topics in conceptual space: 3D Mesh (left) and Voronoi tessellation (right)

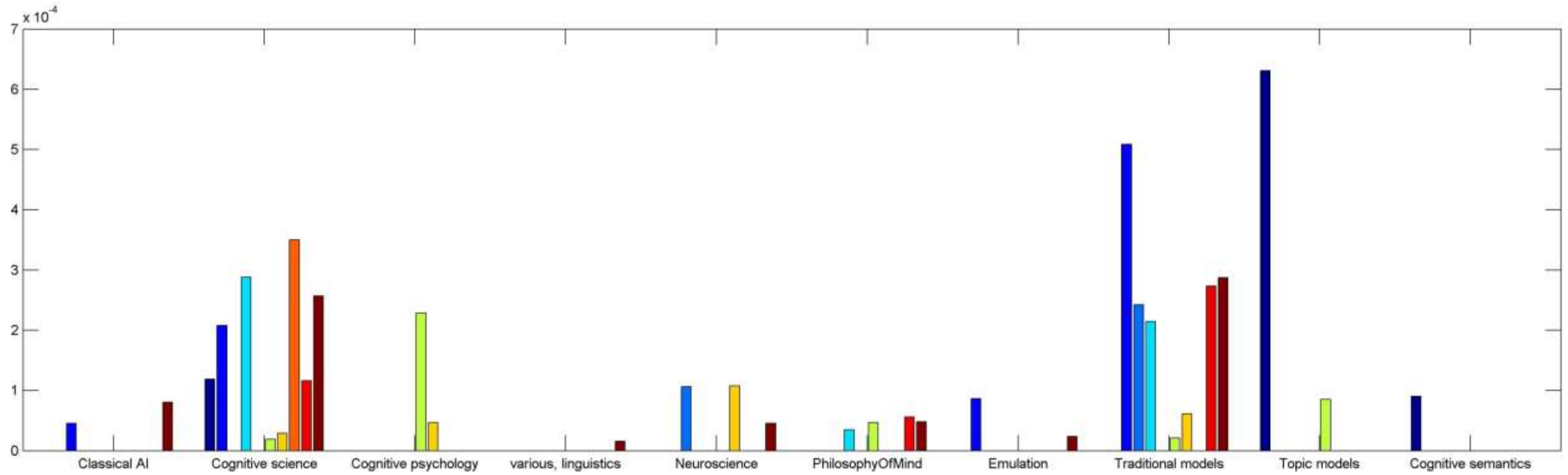


Figure 25: pLSA distribution of 10 topics over conceptual space. Most salient member of individual topic is chosen for distribution.

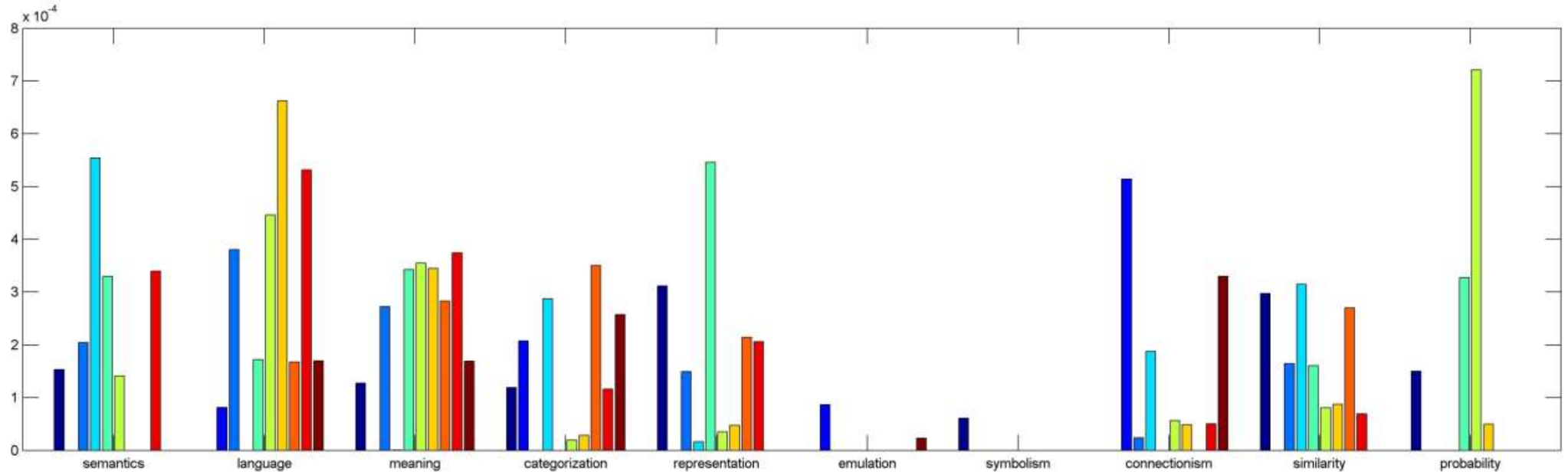


Figure 26: pLSA distribution of words/concepts over topics (general)

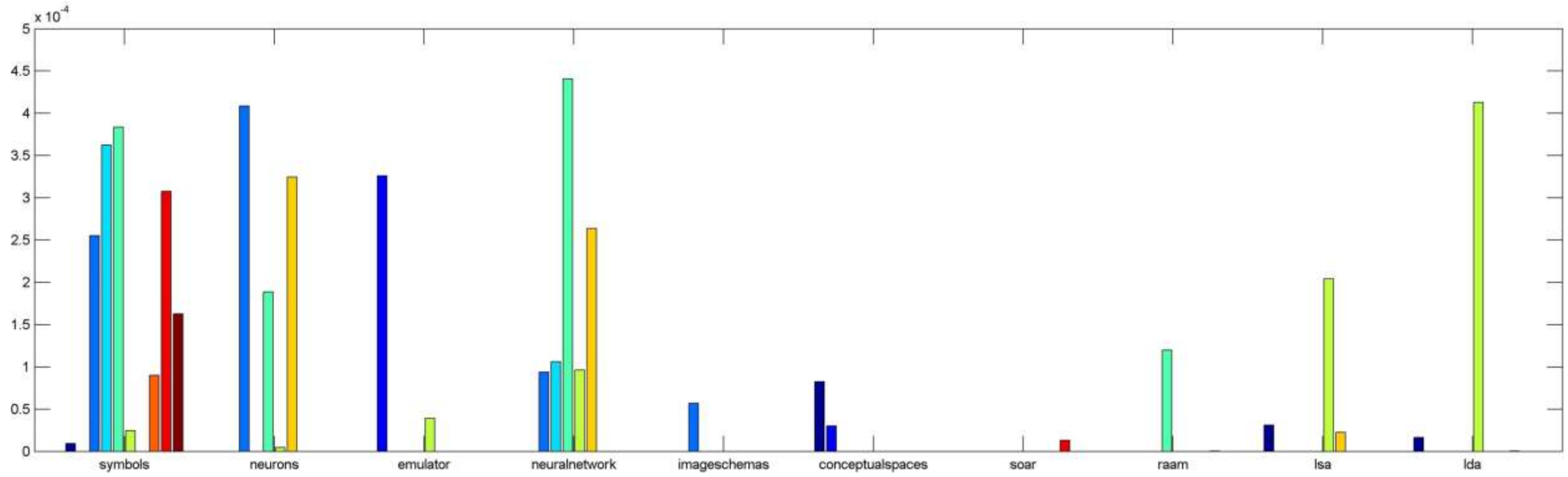


Figure 27: pLSA distribution of words/concepts over topics (cognitive modeling)

Povzetek

Semantika znotraj: Reprezentacija pomena v konceptualnih prostorih

1 Uvod

Doktorska naloga obravnava temeljno problematiko kognitivne znanosti, oblikovanje semantičnih reprezentacij. Vključuje tako teoretični kot tudi praktični oz. konstrukcijski vidik: proučuje strukturo in zahteve za oblikovanje reprezentacij na *konceptualnem* nivoju in izgradnjo računalniškega modela, ki bo omogočal semantično repezentacijo vsebin. Glavni problem večine obstoječih računalniških modelov je v tem, da temelje na simboličnem ali konekcionističnem/asociativnem pristopu, kjer učinkovito semantično oblikovanje reprezentacij ni možno, saj ta pristopa ne omogočata ustrezne konceptualne podlage za konstruiranje *pomena*. Kot rešitev oblikovanja semantičnih reprezentacij je predlagana teorija *konceptualnih prostorov* (Gärdenfors 2000), v navezavi z metodami in tehnikami za računalniško analizo jezika. *Konceptualni prostori* predstavljajo temeljno teoretsko podlago kakor tudi matematično strukturo potrebno za izgradnjo računalniškga modela. V nadaljevanju so predstavljena izhodišča, cilji, metode in rezultati doktorske disertacije.

2 Izhodišča

Glavna raziskovalna problematika zadeva izgradnjo računalniškega modela za oblikovanje semantičnih reprezentacij. Obstoječe rešitve večinoma temelje na tradicionalnih simboličnih in asociativnih pristopih k oblikovanju reprezentacij, ki pa so v disertaciji po kritični analizi zavrnjeni. Osnovno izhodišče in raziskovalna hipoteza je: *pomeni* so *konceptualne entitete*, ki jih je mogoče uspešno reprezentirati zgolj na konceptualnem nivoju. Slednji predstavlja vmesni nivo med obstoječima tradicionalnima pristopoma in, vsaj teoretično, omogoča hibridno povezovanje

arhitektur teh sistemov. Kot osnovna rešitev oblikovanja semantičnih reprezentacij je predlagana Gärdenforsova (2000) teorija *konceptualnih prostorov*. Ta, za razliko od teorestko sorodnih rešitev kognitivne semantike, omogoča matematično opredeljivo strukturo, ki je primerna za računalniško implementacijo in interpretacijo. Nadalje, teoreija *konceptualnih prostorov* ponuja verodostojno razlago nekaterih kognitivnih pojavov (npr. efekt *prototipičnosti*), ki so navzoči pri človekovem razumevanju pomena, nastajanja oz. formacije konceptov, kategorizaciji itn., a jih tradicionalna pristopa ne zmoreta uspešno razložiti.

Teorija *konceptualni prostorov* uvaja nov pristop k oblikovanju reprezentacij na podlagi semantičnih razmerij med koncepti: koncepti so repezentirani v prostoru geometrično, na podlagi kvalitativnih dimenzij in regij katerim pripadajo. To ima kar nekaj konkretnih posledic tako v praktičnem (npr. konstrukciji računalniških modelov in simulatorjev) kot tudi teoretičnem smislu (reševanju problematike simboličnega in asociativnega pristopa). Ker so kvalitativne dimenzije v prostoru reprezentirane vektorsko, je generiranje *konceptualnih prostorov* matematično in tako neposredno prevedljivo v računalniško simulacijo. Po drugi strani pa z vpeljavo konceptualnega nivoja rešimo številne probleme povezovanja reprezentacij simboličnega in asociativnega nivoja v nek hibriden sistem.

V nadaljevanju sta prvo na kratko predstavljeni glavni semantični teoriji znotraj kognitivne znanosti, *referenčna* in *kognitivna semantika*. Poseben poudarek je na kognitivnem pristopu, ki za razliko od tradicionalne referenčne semantike *pomene* razume kot *konceptualne* entitete. Nadalje so predstavljene prednosti in pomankljivosti tradicionalnega simboličnega in asociativnega pristopa k oblikovanju reprezentacij. Sledi predstavitev teorije *konceptualnih prostorov*, ki je temeljna teoretična podlaga disertacije. Teorija *konceptualnih prostorov*, in konceptualni vidik na splošno, igrata v tej disertaciji osrednjo vlogo – tako v odkrivanju številnih problematičnih vprašanj, ki nastanejo pri uporabi simboličnih in asociativnih modelov, kot tudi z vidika formalne reprezentacijske strukture in metodološkega okvira izgradnje računalniške aplikacije za semantično reprezentacijo vsebin (z delovnim imenom *SpaceWalk*). V ta namen, so v zadnjem delu naloge predlagane konkretne rešitve: z navezavo konceptualnih prostorov na različne metode

računalniške analize jezika dobimo nove inovativne rešitve oblikovanja semantičnih reprezentacij.

2.1 Problematika

Problem semantike in pomena v kognitivni znanosti obravnavata dva temeljna, a precej različna pristopa: tradicionalen *referenčni* pristop in *kognitivni* pristop. Znotraj tradicionalnega pristopa (predvsem v lingvistiki in filozofiji) je *semantika* razumljena kot razmerje med *jezikom* in *svetom* (oz. različnimi 'možnimi svetovi'). Poudarja objektivističen pogled na svet, kjer funkcioniranje semantičnega razmerja temelji na podlagi resničnostnih sodb. Ta pristop k semantiki je pogosto imenovan *referenčna semantika*, saj sklepa, da besede dobe svoje *pomene* z navezavo na konkretne objekte in dogodke v svetu. S stališča kognitivne psihologije je omenjeni vidik zelo problematičen. Kot prvo, ne vključuje uporabnike jezika. Ne pove nam ničesar o tem, kako posameznik zapopade *pomene* izražene v taki navezavi (Lakoff 1987, Gärdenfors 1997). Kot drugo, problem takega sklepanja (vsaj s kognitivnega stališča) je tudi v tem, da je *semantična relacija* definirana na podlagi resničnostnih sodb ter nujnih in zadostnih pogojev, ki konstituirajo nek koncept. Tretje, tak način ne zmore razložiti številnih psiholoških vplivov, npr. efekt *prototipičnosti*, s tem povezane delne (angl. 'graded') kategorizacije, ali vpliv konteksta na *pomen*, ki so navzoči v človekovem dojemanju jezika in sveta. Tudi s tega vidika referenčna semantika ni sprejemljiva kot kognitivna teorija.

Drug pristop predstavlja *kognitivna semantika*. Ta izpostavlja uporabnika s tem, ko se osredotoča na razmerje med *jezikovnim izrazom* in uporabnikovo *mentalno reprezentacijo pomena* tega izraza, večinoma v obliki 'slikovnih shem' (angl. 'image schema'). *Pomen* postane *konceptualna entiteta*. Kot primer lahko vzamemo različne modele kognitivne semantike (glej npr. Lakoff 1987, Langacker 1986, 1987), kjer so *slikovne sheme* temeljni nositelji *pomena*. Slikovna shema je v bistvu konceptualna struktura, ki pripada določenemu posamezniku. Navezava na zunanji svet tu ni bistvena, kakor tudi ne resničnost stavkov, ta je sedaj nadomeščena s *prepričanjem*. Posledica kognitivistične pozicije, ki jo spravlja v konflikt z večino ostalih, realistično obarvanih semantičnih teorij je v tem, da resničnostne predpostavke sedaj

niso več potrebne za določitev njegovega *pomena*. Resničnost izraza postane sekundarna, s tem ko se ukvarja z razmerjem med kognitivno strukturo in svetom. V kognitivni semantiki je *pomen* pred *resnico.* Vendar je problem kognitivne semantike predvsem v tem, da so matematične strukture slikovnih shem preveč abstraktne in malokdaj izpeljane, in posledično niso primerne za računalniško implementacijo. V ta namen so potrebni matematični parametri in izmerljive metrične strukture, ki jih uvaja teorija *konceptualnih prostorov* (Gärdenfors 2000).

2.1.1 Simbolični in asociacijski pristopi k oblikovanju reprezentacij

Raziskave v kognitivni znanosti lahko delimo glede na dva med seboj povezana cilja. Eden je *pojasnjevalen*: s proučevanjem kognitivnih aktivnosti ljudi in živali, lahko formuliramo teorije o različnih aspektih kognicije. Teorije se testirajo z eksperimenti in z računalniškimi simulacijami. Drugi cilj je praktičen oz. *konstrukcijski*: z izgradnjo artefaktov kot npr. programov za igranje šaha, robotov, animacij itd., poskušamo konstruirati sisteme, ki lahko opravijo različne kognitivne naloge. Za obe vrsti ciljev je temeljni problem v tem, kako oblikovati reprezentacije, ki jih določen kognitiven sistem uporablja.

V kognitivni znanosti prevladujeta dva temeljna pristopa k oblikovanju reprezentacij. *Simbolični* pristop temelji na predpostavki, da naj bi bili kognitivni sistemi oblikovani kot Turingovi stroji. S tega stališča, je kognicija v bistvu razumljena kot manipulacija abstraktnih simbolov. Drugi pristop je konekcionistični oz. *asociativni*, kjer asociacije med različnimi informacijami nosijo glavno težo reprezentacij (npr. umetna nevronska omrežja, angl. 'Artificial Neuron Networks' oz. ANNs). Oba vidika imata svoje prednosti in pomankljivosti. Pogosto sta razumljena kot konkurenčna, vendar rešujeta probleme na različnih nivojih in jih je bolj smiselno jemati kot komplementarna. Npr. simbolični pristop se ukvarja z abstraktno mislijo, planiranjem itn., torej z visoko-nivojski vidiki spoznavanja. Predmet asociativnega pristopa pa sta npr. percepcija in motorika, torej nizko-nivojski vidik. Noben od omenjenih pristopov ne ponuja primerne razlage in modela za oblikovanje semantičnih reprezentacij. Rešitev predstavlja *konceptualni* pristop, ki ga izraža že kognitivna semantika, bolj natančno pa razvije teorija *konceptualnih prostorov*:

pomeni so reprezentirani v preslikavi besed na konceptualne strukture, t.j. *pomene* sestavljajo koncepti. Koncepti reprezentirajo naše znanje oz. vedenje o stvareh v svetu: omogočajo nam identifikacijo novih stvari, kot tudi razumevanje njihovih nevidnih lastnosti (npr. obnašanje, funkcijo ali delovanje), na podlagi že pridobljenega znanja. Koncepti nam nudijo bistvene informacije potrebne v naši interakciji s svetom. Nadalje, obstaja ogromno empiričnih dokazov v prid konceptualni podlagi *pomena* besed (v nadaljevanju je predstavljeno le nekaj tistih primerov, ki se direktno nanašajo na temeljno teorijo disertacije, teorijo *konceptualnih prostorov*). *Učinek kategorizacije* (angl. 'category effect') npr., odkriva kategorijsko strukturo v semantičnem spominu: objekti iste kategorije so bolj sorodni kot objekti v različnih kategorijah (Rosch 1975). Drug primer je *teorija prototipičnosti* (prototype theory) – ta zajema raziskave učinka tipičnosti (angl. 'typicality effect'), ki poudarjajo neenakost med posameznimi člani določene kategorije, z nekaterimi prototipičnimi in drugimi manj tipičnimi (npr. številne raziskave Rosch (1973, 1975, 1987)). Namen omenjenih raziskav je bil pokazati na asimetrije med člani kategorije in asimetrične strukture znotraj samih kategorij, ki bi jih semantična teorija morala ustrezno razložiti. Oba učinka igrata pomembno vlogo v teoriji *konceptualnih prostorov*. Znotraj tradicionalne (objektivistične) interpretacije kategorij in konceptov je namreč učinek prototipičnosti zelo težko razložiti: objekt je ali pa ni član neke kategorije in vsi člani kategorije imajo enak status. Posledično tradicionalni simbolni pristop omenjenih asimetrij niti ne predvideva niti jih ne more uspešno reprezentirati.

3 Metode

Ker *konceptualni prostori* predstavljajo temeljno podlago za izgradnjo računalniškega modela, je pomemben del raziskav namenjen vprašanjem kako *koncepti* reprezentirajo *pomene* in kako te reprezentacije oblikovati na primeren način. Če je cilj izgradnja računalniškega modela za semantično reprezentacijo vsebin, potem je seveda bistveno vprašanje katera struktura je najbolj primerna, tako iz praktičnega kot teoretskega vidika. Za generiranje konceptualnih prostorov so namreč potrebne dodatne metode semantične analize jezika (na podlagi katerih

najprej izluščimo kvalitativne dimenzije), ki pa se med seboj močno razlikujejo. Nadaljni problem je tudi, kako premostiti vrzel med različnimi nivoji oblikovanja reprezentacij. V nadaljevanju je predstavljena teorija *konceptualnih prostorov*. Sledi predstavitev prevladujoče metode za semantično analizo, latentne semantične analize (LSA), kasneje pa še ostale, bolj primerne rešitve.

3.1 Teorija *konceptualnih prostorov*

Za rešitev omenjenih problemov oblikovanja reprezentacij Gärdenfors (2000) predlaga konceptualno reprezentacijo informacij. Konceptualni prostor je sestavljen iz geometričnih struktur na podlagi *kvalitativnih dimenzij*. Temeljna naloga *kvalitativnih dimenzij* je v izgradnji domen oz. področij, potrebnih za reprezentacijo konceptov. Struktura večine kvalitativnih dimenzij je metrična – lahko govorimo o razdaljah vzdolž posameznih dimenzij. Nadalje, obstaja tesna povezanost med razdaljami v konceptualnem prostoru in ocenami *podobnosti*: manjša kot je *razdalja* med reprezentacijama dveh objektov, večja je *podobnost*. Na ta način lahko *podobnost* med dvema objektoma definiramo z *razdaljo* med njunima točkama v prostoru. Posledično nam *konceptualni prostori* omogočajo naravno reprezentacijo *podobnosti*. Iz tega sledi kriterij "naravni koncept je reprezentiran kot niz regij znotraj nekega števila področji, skupaj z dodelitvijo uteži področjem, ter informacijo o tem, kako so regije različnih področji povezane" (Gärdenfors 2000 str. 105). a natural concept is represented as a set of regions in a number of domains together with an assignment of salience weights to the domains and information about how the regions of different domains are correlated

Drug pomemben kriterij teorije *Konceptualni prostorov* se nanaša na *naravne lastnosti* konceptov reprezentiranih kot "… konveksna regija področja v konceptualnem prostoru" (Gärdenfors 2000 str. 71).

V teoriji *Konceptualnih prostorov* razdelitev prostora v konveksne ploskve oz. področja temelji na množici prototipov, t.j. besed, ki so najbolj karakteristične za posamezno domeno. Obstajajo zanimive povezave med reprezentacijo konceptov kot konveksnih področij in *teorijo prototipičnosti* razvito s strani Rosch in sodelavcev (glej npr. Rosch 1975, 1978, Mervis in Rosch 1981, Lakoff 1987). Če koncepte

definiramo kot konveksna področja *konceptualnega prostora*, je učinek prototipičnosti v bistvu za pričakovati. V konveksnem področju je pozicija (npr. objekta) definirana kot bolj ali manj centralna.

Argumentacija je mogoča tudi v nasprotni smeri: če upoštevamo teorijo prototipičnosti, potem je reprezentacija konceptov kot konveksnih področij naravna posledica. Primer: predstavljajmo si, da imamo podanih nekaj kvalitativnih dimenzij konceptualnega prostora, npr. dimenzije prostora barve, in želimo narediti dekompozicijo prostora na nekaj kategorij, npr. konceptov barve. Če začnemo z množico prototipov p_1, ..., p_n obravnavanih konceptov, npr. osrednjih barv, potem naj bi ti bili središčne točke konceptov, ki jih predstavljajo. Informacijo o prototipih lahko nadalje uporabimo za generiranje konceptov s pogojevanjem, da katerakoli točka p pripada istemu konceptu kot najbližji prototip p_i (Gärdenfors 2000). Dokazano je, da to pravilo generira dekompozicijo prostora – tako imenovan Voronoijev diagram (angl. Voronoi tessellation). Osnovna predpostavka je, da najbolj tipični *pomen* besede predstavlja prototip konveksnega področja, dodeljenega tej besedi. Bistvena lastnost Voronoijevega diagrama konceptualnega prostora je, da vedno rezultira v dekompoziciji prostora v konveksna področja. Na ta način Voronoijev diagram priskrbi konstruktiven geometrijski odgovor na to, kako merilo *podobnosti* skupaj z množico prototipov determinira množico kategorij.

Metrična reprezentacija je merilo podobnosti med različnimi objekti, ki so reprezentirani kot *točke* v *konceptualnih prostorih*. Matematično, so te točke v dimenzionalnih prostorih razumljene kot *vektorji*. Posledično izračuni na konceptualnem nivoju v veliki meri vključujejo *vektorske izračune*, na podlagi matric itd. Na podlagi razdalj v in med konceptualnimi prostori lahko določimo različne kriterije klasifikacije (npr. Voronoijevo razdelitev, ki generira razmejitve na podlagi prototipov posameznih domen).

Ena izmed metodoloških značilnosti, ki konceptualni nivo jasno loči od simboličnega je *podobnost* – ta igra centralno vlogo v reprezentacijah na konceptualnem nivoju. Podobnost med objekti in lastnostmi je namreč reprezentirana z razdaljami v prostorih; to definicijo bi težko reprezentirali na naraven način v simboličnih sistemih. V primeru *konceptualnih prostorov*, nam Voronoijev diagram omogoča

konkreten prikaz kako merilo podobnosti, skupaj z množico prototipov in kvalitativnih dimenzij, determinira množico naravnih lastnosti objektov (in konceptov) – v *konceptualnih prostorih* je semantika determinirana prostorsko. To pomeni, da *semantičnost* objektov leži znotraj kvalitativnih dimenzij konceptualnega prostora. Na ta način je teorija prototipičnosti vpeljana v *konceptualne prostore,* na ta način lahko tudi formuliramo nove koncepte in se jih učimo. Struktura *konceptualnih prostorov* je posledično dobra podlaga za oblikovanje semantičnih reprezentacij.

3.2 Latentna semantična analiza

Latentna semantična analiza je v svojem bistvu neke vrste asociativni model. Asociativni modeli, kot ime že samo pove, temeljijo na asociacijah med besedami; *pomen* določene besede je tako množica ostalih besed, na katere je ta beseda navezana. Na ta način, produkcija jezika in razumevanje postaneta preprosto predmet povezovanja niza asociacij. Mentalne reprezentacije in ostali kompleksni mehanizmi naj ne bi bili potrebni pri tolmačenju uporabe jezika in *pomena*. Ti modeli imajo dolgo zgodovino v psihologiji, njihova priljubljenost je bila v veliki meri pogojena z dejstvom, da uporabljajo malo notranjega kognitivnega procesiranja (Murphy 2002). Na področju semantične analize jezika je najbolj razširjena *Latentna semantična analiza* (angl. *Latent Semantic Analysis* oz. LSA; Landauer & Dumais 1997). Ta računalniška metoda temelji na asociativnem pristopu oblikovanja *pomenov* besed, vendar presega klasične asociativne modele, ki omogočajo zgolj reprezentacijo asociativnih povezav med besedami. Kot tehnika strojnega učenja, ki ponazarja kognitivno doumevanje besedila, LSA (Landauer and Dumais 1997, Landauer et al. 2006) povzame *pomen* iz odstavkov na način, da analizira vzorce uporabe besed v več dokumentih in potem reprezentira besede in njihove kontekste kot vektorje v visoko-dimenzionalnem prostoru. Frekventnost pojavljanja besed je definirana v matrici z stolpci, ki jih sestavljajo besede in vrsticami, ki predstavljajo dokumente. Veliko celic stolpcev in vrstic je praznih (oz. vsebujejo 0). Z namenom, da bi ohranili le bistvene značilnosti, je potrebno dimenzionalnost originalne matrice reducirati s pomočjo SVD dekompozicije (oz. *Singular Value Decomposition*) na približno 300 dimenzij. To nam omogoči oblikovanje semantične sorodnosti odstavkov in besed kot vektorjev, z vrednostmi proti 1 (ki izražajo stopnjo sorodnosti med enotami) in

nizkimi oz. negativnimi vrednostmi, tipično okoli 0,02, ki izražajo nepovezanost (Martin in Berry 2006). V tem semantičnem prostoru so odstavki ali besede, ki izražajo isti *pomen,* reprezentirani kot vektorji tesno skupaj, čeprav dejansko ne delijo skupnega izraza. Namesto tega se ti izrazi lahko pojavljajo v drugih dokumentih z isto temo, z redukcijo dimenzionalnosti originalne matrice na podlagi SVD pa so relativne jakosti teh asociacij reprezentirane kot kosinusi ali točke vektorjev v prostoru. Podlaga za kreiranje teh asociacij med besedami temelji na zelo obsežnih zbirkah dokumentov, v primeru LSA je uporabljena TASA zbirka (Touchstone Applied Science Associates Inc.), ki jo sestavlja korpus besedil, knjig, člankov in ostalega splošnega gradiva, kateremu je ameriški študent izpostavljen do 1. letnika univerze. Pri kreiranju matrice je pomembna tudi funkcija uteževanja, ki temelji na frekventnosti pojavljanja besed v odstavkih in je v obratnem sorazmerju do pojavljanja besed preko vseh dokumentov – s tem se izniči pomembnost visoko frekventnih izrazov, ki ne prispevajo bistveno k razlagi pomena (Martin in Berry 2006).

Zakaj bi ta postopek rezultiral v semantični podobnosti? Besede, ki se pojavljajo skupaj, namreč pogosto nimajo semantične podobnosti. Vendar LSA ne uporablja le informacije o tem kako pogosto se *beseda1* in *beseda2* pojavljata skupaj, marveč tudi kako pogosto se pojavljata z vsemi ostalimi besedami v zbirki (Landauer et al. 2006). LSA na ta način analizira celoten vzorec dogodkov in interpretira besede kot podobne če se nahajajo v tematsko podobnih stavkih oz. odstavkih. Poudarjena je vloga konteksta: npr. *pes* in *mačka* sta si podobna, ker se pojavljata v podobnih stavčnih kontekstih.

Landauer in Dumais (1997) sta med drugim preizkusila LSA na testu razpoznavanja sinonimov za Test of English as a Foreign Language (TOEFL), ki je uporabljan kot preizkus znanja angleščine za sprejem tujih študentov na ameriške univerze, in dobila izredne rezultate: sistem je opravil test z 64.4% uspešnostjo, kar je skoraj identično učinku velikega vzorca študentov, ki so test opravili. Landauer in Dumais (1997) zaključujeta, da je ta ocena zadovoljiva za sprejem na večino ameriških univerz.

Uspeh LSA lahko namiguje na to, da konceptualne informacije morda preprosto niso ključne za tolmačenje *pomena*: ker so edini vložek v sistem besede, bi lahko *pomen* reprezentirali kar preko povezav na druge besede, kot pa preko znanja kot osnove tem besedam – t.j. konceptov. Ali to drži?

Glavni problem teh mrež asociacij je v tem, da golo vedenje katere besede tvorijo asociacije med seboj še ne specificira kaj *pomen* posamezne besede res je. Kot poudarja Murphy (2002 str. 429), ne moremo izluščiti *pomena* besed zgolj z referenco na ostale besede: "Če nekdo pozna *psa* zgolj po njegovi podobnosti z *mačko* in *kravo* in *kostjo* ... in *mačko* zgolj po njeni podobnosti z *psom* in *kravo* in *kostjo* ... itd., potem je ujet v krog podobnih besed". Dodaten problem je v tem, da se posamezna razmerja med besedami lahko bistveno razlikujejo in neka celotna podobnost generirana s strani LSA ne specificira podlage kot tudi ne razlik med posameznimi asocijacijami. Reprezentacija podrobnega poznavanja referentov besed uporabljenih v govoru in dojemanju ni zadovoljiva. Besede se morajo povezovati z našim znanjem oz. vedenjem o stvareh v svetu, ne zgolj z ostalimi besedami. To pa so ravno stvari, ki jih konceptualni pristop omogoča: imeti koncept nam med drugim tudi razloži zakaj so določene besede sorodne. Ker so koncepti mentalne entitete, naša ne-lingvistična interpretacija sveta, lahko le z navezovanjem besed na konceptualno strukturo razložimo pomene besed. Tu pride v ospredje teorija *konceptualnih prostorov*.

4 Cilji in rezultati

Predlagana alternativa tradicionalnim modelom semantike temelji na teoriji konceptualnih prostorov (Gärdenfors 2000). Glavni cilj je zgraditi računalniški model za semantično reprezentacijo, ki spaja Gärdenforsovo teorijo konceptualnih prostorov z metodami za semantično analizo naravnega jezika. Slednje so potrebne za ustvarjanje kvalitativnih dimenzij, na podlagi katerih lahko kreiramo konceptualne prostore (glej projekcije v Prilogi, str. 168).

4.1 Struktura aplikacije *SpaceWalk*

Pri izgradnji računalniškega modela smo uporabili in primerjali tri različne metode za semantično analizo jezika: LSA, pLSA (verjetnostni LSA oz. angl. 'probabilistic LSA') in LDA (latentna Dirichletova alokacija oz. angl. 'Latent Dirichlet Allocation'). Omenjene metode so potrebne za generiranje semantičnih relacij, da se izlušči smiselne semantične dimenzije in nato, na podlagi teh dimenzij, generira konceptualne prostore. Razlike med njimi so velike, tako z vidika dobljenih semantičnih reprezentacij kot tudi s teoretskega vidika. Testno okolje je predstavljal korpus člankov in knjig s področja kognitivne znanosti in sorodnih področij.

V primeru LSA, dobimo asociacijsko mrežo besed, ki je generirana na podlagi pojavljivosti posamezne besede znotraj korpusa. Na ta način dobimo globalno oceno podobnosti med *besedo1* in *besedo2*. Ker so prostori generirani z LSA na nek način tudi semantični, se je potrebno vprašati kakšen je ta ‘semantični prostor’? Semantični prostor (generiran z LSA) je distribuirana mreža besed, kjer so besede, sorodne v pomenu, tesno skupaj v prostoru. Na primer beseda ‘ugajati’ bi morala biti blizu besedi ‘ljubezen’ in oddaljena od besede ‘svinčnik’.Vendar so ti semantični prostori kreirani izključno na podlagi statistike pojavljivosti besed v besedilu. Posledično, z LSA generirane dimenzije ponavadi ne zagotovijo zadosten nivo nadrobnosti za semantično interpretacijo.

Kot je bilo omenjeno, problem je kako premostiti vrzel med različnimi nivoji reprezentacij. Obstajajo številne variante asociacijskih modelov na podlagi LSA. Za vse pa je značilno, da so njihove omejitve v asociativnem pristopu in posledično, v nezmožnosti reprezentacije resničnega pomena posameznih besed, t.j. znotraj konteksta v katerih so nastale asociacije med konkretnim besedami. V konkretnem primeru imamo na eni strani visoko-dimenzionalen vektorski prostor, generiran s statistično metodo semantične analize jezika (LSA), na drugi strani pa konceptualne prostore, znotraj katerih naj bi bil poleg semantične podobnosti besed reprezentiran tudi vpliv konteksta. LSA tu ne ponuja zadovoljive rešitve, saj ne zmore tematske analize korpusa, marveč generira semantična razmerja le znotraj besed na podlagi statistike njihovega pojavljanja, brez dodatne strukture in navezave na tematsko raznolikost dokumentov. Kar, vsaj s teoretskega vidika razlage semantike, t.j. v

smislu vpliva konteksa, sinonimije in polisemije, ni zadosten pogoj za kreiranje relevantnega konceptualnega prostora.

Alternativa statističnemu pristopu so verjetnostni modeli, ki v zadnjem času v literaturi kognitivne semantike dobivajo vse večji pomen (Griffiths et al. 2008, 2010, Clark (v tisku)). V disertaciji sta predstavljena dva: pLSA (Hofmann 1999, 2001) in LDA (Blei, Ng in Jordan 2003). pLSA se razlikuje od LSA v tem, da v semantično analizo uvaja verjetnostno metodo in s tem omogoča analizo posameznih tem znotraj korpusa. Vendar ima pLSA, poleg nekaterih tehničnih omejitev (npr. kot pri LSA, število parametrov raste linearno z velikostjo korpusa), resno pomankljivost: ni ekspliciten verjetnostni model in porazdelitev tem lahko izvede le na dokumentih znotraj učne množice, ne pa za dokumente izven le te (Blei et al. 2003). LDA (Blei et al. 2003) je verjetnostni tematski model, ki na podlagi verjetnostne porazdelitve prikaže mešanico tem znotraj dokumenta. Glede na temo, dobi beseda različno verjetnostno porazdelitev; npr. za besedo "um" je velika verjetnost, da pripada temi "filozofija", in majhna verjetnost pripadnosti temi "zelenjava". Na ta način LDA zagotovi reprezentacijo večih pomenov neke besede, kot tudi tematsko strukturo znotraj nekega dokumenta. Ta princip da zadostno strukturo za zajem nekaterih kvalitativnih vidikov semantike naravnega jezika, npr. sinonimijo, polisemijo in kontekst. To, in pa zmožnost generaliziranja na nove dokumente izven učne množice, je ena izmed bistvenih prednosti LDA pred ostalima metodama. Na podlagi LDA generiranih kvalitativnih dimenzij pridobimo neko osnovno semantično strukturo in lahko posledično izpeljemo bolj smiselne reprezentacije pomena na konceptualnem nivoju, znotraj *konceptualnih prostorov*.

5 Zaključek

Dva različna pristopa k obdelavi naravnega jezika, statistični in verjetnostni pristop, ki odražata logiko tradicionalnega simboličnega in asociativnega pristopa do kognicije, sta bila implementirana v računalniški model. O učinkovitosti obeh se je na široko razpravljalo v znanstveni literaturi (npr. Landauer in Dumais 1997, Seidenberg in MacDonald 1999, Landauer et. Al 2006, Blei et al. 2003, Hofmann 1999, 2001, Steyvers in Griffiths 2007 , Griffiths et al. 2010, Blei 2011). Na podlagi

rezultatov pridobljenih iz teh študij, so verjetnostni modeli na splošno bolj primerni za semantično analizo kot statistični (Blei et al. 2003, Hofmann 1999). Kot se je izkazalo, z vidika kognitivne znanosti statistični modeli ne zagotavljajo dobre teoretične ali praktične podlage za semantično analizo naravnega jezika. Teoretične predpostavke, na katerih temelje statistični izračuni semantične podobnosti, ne podpirajo osnovnih rezultatov raziskav kognitivne psihologije, npr. raziskav o kategorizaciji (Rosch et al.) ali raziskav o podobnostnih sodbah (Tversky et al.), ki so temelj človekove interpretacije pomena. Res je, da omenjeni statistični pristop lahko generira semantično podobnost z analizo nekega korpusa besedil, in morda celo simulira nekatere učinke uporabe jezika, kar dokazuje že omenjeni test razpoznavanja sinonimov. Vendar so ti učinki ustvarjeni zgolj na podlagi pojavnosti besed in razen asociativne mreže besed ne generirajo nobene dodatne semantične strukture. Človekova konceptualna struktura pa ni zgolj rezultat statističnega sklepanja na podlagi besednih zvez, marveč nanjo vpliva konceptualni in kategorični vidik, predhodno znanje in kontekst, ter kulturni in družbeni vplivi (Jäger in van Rooij 2007). V tem pogledu je verjetnostni pristop drugačen in bolj primeren za oblikovanje semantičnih reprezentacij. Kot generativni tematski model, je LDA sicer konceptualno bližje simboličnemu pristopu k oblikovanju reprezentacij, vendar odpravlja večino njegovih slabosti. Ker generira osnovno hierarhično strukturo, s tem omogoči povezovanje različnih nivojev reprezentacij in je prvi korak k hibridnemu pristopu. LDA na eni strani izkorišča asociativni pristop in tako omogoča, da 'okolje' vpliva na semantično strukturo. Po drugi strani pa verjetnosti pristop predstavlja sklop 'top-down' omejitev, ki jih lahko interpretiramo kot efekt indukcije, oz. učinek pristranskosti (v obliki predsodkov, mnenj ali predhodnega znanja) na človekovo sklepanje (Griffiths et al. 2008, 2010, Clark (v tisku)). Skupaj s teorijo konceptualnih prostorov, nam verjetnostni pristop omogoča bolj prožen okvir za ustvarjanje in raziskovanje semantičnih reprezentacij.

Vloga konceptualnih prostorov je ključna za razumevanje in reprezentacijo pomena in semantike naravnega jezika. Konceptualni prostori obstoječim kvalitativnim dimenzijam generiranim z LDA, LSA, ali pLSA, dodajo dodaten, konceptualni nivo. Kar dobimo, v strojno berljivi obliki, ni zgolj reprezentacija lastnosti, konceptov in podobnostnih relacij, marveč formalni okvir, ki izkorišča kvalitativne dimenzije na

način, ki ustreza izsledkom raziskav kognitivne psihologije (npr. teoriji prototipičnosti).

Obstajajo številne možnosti za aplikacijo predstavljenega modela. Poleg podpore raziskavam kognitivne znanosti, tako pojasnevalnim kot konstrukcijskim (kot so npr. problematika oblikovanja reprezentacij, formacija konceptov, semantična reprezentacija vsebin, itd.), je prihodnost semantičnih modelov predvsem na področju strojnega učenja (in umetne inteligence na splošno), reprezentacije znanja, digitalnih vsebin in semantičnega spleta. V tem kontekstu je namen predlaganega modela lahko dvojen: na eni strani bi služil kot sistem strukturiranja informacij v semantične oz. konceptualne geometrijske strukture, kot to predvideva teorija *konceptualnih prostorov*, po drugi strani pa takšna reprezentacija sama po sebi omogoča uporabniški vmesnik za dostop do teh vsebin. Uporabnik bi tako lahko pregledoval *konceptualne prostore*, posamezna konveksna področja konceptov, različne *pomene* besed, semantične relacije itn.

V prihodnosti je predviden bolj celosten in dinamičen pristop in možnost pregledovanja konceptualnega prostora z manipulacijo kvalitativnih dimenzij, npr. s spreminjanjem števila tem ali s spremembo teže posameznih tem, in s tem odkrivanje alternativnih semantičnih povezav. V sedanji različici računalniškega modela so namreč mogoči le delni posnetki konceptualnega prostora in morebitne projekcije alternativnih skupin kvalitativnih dimenzij potrebujejo nadaljnje izračune. Teorija *konceptualnih prostorov* predstavlja temelj za nadaljne, bolj metodološke raziskave bistvenih kognitivnih vprašanj o oblikovanju semantičnih reprezentacij, učenju in formaciji konceptov, kategorizaciji itn. V tem kontekstu lahko predstavljeni računalniški model služi kot eksperimentalno orodje in pomoč omenjenim raziskavam. To ostaja motivacija za nadaljnje delo.