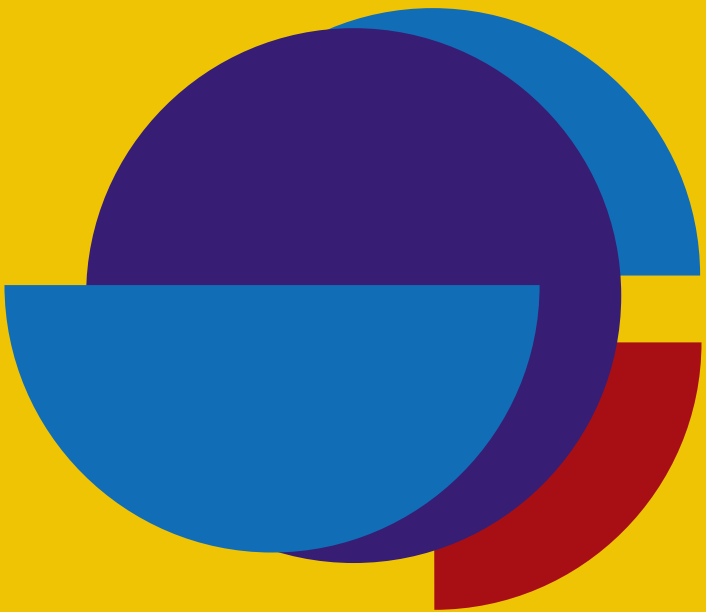
Where Words Get their Meaning

Marianna Bolognesi



CONVERGING EVIDENCE IN LANGUAGE AND COMMUNICATION RESEARCH

23

JOHN BENJAMINS PUBLISHING COMPANY

Where Words Get their Meaning

Converging Evidence in Language and Communication Research (CELCR)

ISSN 1566-7774

Over the past decades, linguists have taken a broader view of language and are borrowing methods and findings from other disciplines such as cognition and computer sciences, neurology, biology, sociology, psychology, and anthropology. This development has enriched our knowledge of language and communication, but at the same time it has made it difficult for researchers in a particular field of language studies to be aware of how their findings might relate to those in other (sub-)disciplines.

CELCR seeks to address this problem by taking a cross-disciplinary approach to the study of language and communication. The books in the series focus on a specific linguistic topic and offer studies pertaining to this topic from different disciplinary angles, thus taking converging evidence in language and communication research as its basic methodology.

For an overview of all books published in this series, please see
benjamins.com/catalog/celcr

Volume 23

Where Words Get their Meaning. Cognitive processing and distributional modelling of word meaning in first and second language
by Marianna Bolognesi

Where Words Get their Meaning

Cognitive processing and distributional modelling of word meaning in first and second language

Marianna Bolognesi
University of Bologna

John Benjamins Publishing Company
Amsterdam / Philadelphia

The paper used in this publication meets the minimum requirements of the American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

DOI 10.1075/celcr.23

Cataloging-in-Publication Data available from Library of Congress:
LCCN 2020040110 (PRINT) / 2020040111 (E-BOOK)

ISBN 978 90 272 0801 9 (HB)
ISBN 978 90 272 6042 0 (E-BOOK)

John Benjamins Publishing Company · https://benjamins.com

To Mike, Sean and Nora

Table of contents

Part 3. Converging evidence in language and communication research

Acknowledgements

I consider myself very lucky. The list of names that appear in this section demonstrates how lucky I am to have such a great network of colleagues, friends and family members who made this editorial project possible. Let me thank all of them.

I am very grateful to several amazing colleagues: Tommaso Caselli, Pia Sommerauer, Phil Wicke, Francesca Strik Lievers, Valentina Cuccio, Francesca Citron and Ludovica Serratrice, for the fruitful discussions, exchanges of ideas and materials, which helped me developing the arguments and ideas in this book.

I am thankful to two anonymous reviewers who provided very helpful and constructive feedback throughout the process, and to the CELCR series editors, Ninke Stukker and Kris Heylen for their collaboration and kindness, as well as to Esther Roth from the Editorial Staff at Benjamins, for her constant support.

I am thankful to my creative and skillful friend Nate Laffan, excellent graphic designer, for realizing the elegant figures in Chapters 2, 4 and 5. And to my friend Eliza Jane Nash, who patiently proofread the whole manuscript.

Moreover, as I am writing I am 8.5 months into my second pregnancy. I am very thankful to Nora Grace, my daughter, soon due. I thank her for her patience in listening to endless tracks of Brian Eno and other instrumental music in the past weeks, while I finalized this manuscript. I look forward to meeting her soon. I am also very grateful to Sean Giordano, my first son, who will be a protective, respectful and fun older brother. His patience (as 5 years old) and understanding are admirable. Finally, I am immensely thankful to Mike, my love.

CHAPTER 1

Word power

1.1 Introduction

Words have an immense power and much of this power lies under the radar of our conscious detection. A single word can deeply influence our behavior, without us being aware of it. And behavior can be fatal. When people (including doctors) are told that a medical treatment has 95% *survival* rate, they are more likely to use it and prescribe it to patients, than when they are told that it has a 5% *death* rate (McNeil, Pauker, Sox and Tversky, 1982). This phenomenon is often referred to as the framing effect (Kahneman and Tversky, 1979; Tversky and Kahneman, 1981).

Similarly, in a recent study it has been shown that hurricanes named with male names are subconsciously taken more seriously, and people are more likely to take greater precautions, compared to hurricanes named with female names. As a result, more people die in hurricanes named with female names (Jung, Shavitt, Viswanathan and Hilbe, 2014). Seana kills more people than Sean, hurricane-wise.

The power of words and the way in which words influence human behavior has been long studied in economics, marketing, psychology, cognitive science, communication science and related disciplines. Politicians and communication strategists are well aware of the power of individual words, and carefully choose how to frame their views and arguments. For example, Americans are not divided between 'pro-abortion' and 'anti-abortion', but between 'pro-life' and 'pro-choice' supporters, where both sides use a positive framing encoded in the prefix *pro-* to name their standpoint in the debate, and none of them uses directly the 'heavy' word *abortion* (Ferree, Gamson, Gerhards and Rucht, 2002).

Words are not just labels for concepts. Words frame situations and construct meaning that goes beyond the objective description of the designated referents in the world. In a well-known task used to investigate decision making strategies, the so-called 'prisoner's dilemma' (e.g., Axelrod and Keohane, 1985), two participants, in the hypothetical scenario of a prison, are told to be prisoners, and are offered a bargain. Each prisoner is given the opportunity either to betray the other by testifying that the he/she committed the crime, or to cooperate by remaining silent. The two participants cannot communicate with one another. However, if they both betray one another, each of them serves two years in prison; if only one of them betrays the other, the first is set free while the latter will serve three years in prison;

if they collaborate and both remain silent, both of them will only serve one year in prison. Empirical evidence shows that if the game is presented to participants as "The Wall-Street game", with participants being two businessmen, then they become less cooperative, while if it is presented as "The Community game", participants behave more cooperatively (Liberman, Samuels and Ross, 2004). Similarly, participants who read words related to pro-sociality before taking part in the game (e.g., *harmony, mutual*) behave more cooperatively than participants who read words related to competitive behaviors (e.g., *rank, power*) (Gerlach, 2018).

But where do words get their meaning from?

This very general question constitutes the starting point of this volume. Although the reader might argue that this question must have been addressed already by several scholars in the past decades (and centuries!), it remains a hotly debated topic. One of the main reasons for this is that scholars from different scientific communities interpret word meaning in different ways. Cognitive scientists and psychologists for example typically align concepts with word meanings. As Vigliocco and Filipovic Kleiner (2004), following Gentner and Goldin-Meadow (2003) and Levinson (2003) describe, the dominant position within cognitive psychology for the last few decades supports the idea that the conceptual structure and semantic structure are closely coupled, and that the architecture of the conceptual system is relatively similar across cultures, because we share similar bodies through which we experience the world. However, as I will argue further in this book, different languages crop and categorize perceptual experiences in different ways, and therefore language-specific properties can play a role in shaping conceptual representation, as many empirical studies have already shown (see Vigliocco and Vinson, 2007 for a review). I will also show, that words can drive the construction of conceptual categories, that language has the power to override conceptual categories formed on the basis of perceptual experience, and finally that the ability to abstract and construct word meaning on the basis of word-to-word associations only develops from the ability to construct meaning from perceptual experiences and uses the same mechanisms. I argue that such mechanisms are summarized by the distributional hypothesis (Harris, 1954).

My approach to the general question of where (and how) words get their meaning is cross-disciplinary: on one hand I describe what we know about how humans construct and represent word meaning, based on empirical evidence, or how word meaning is constructed and represented in the human mind (Part 1); on the other hand, I describe how computational models have traditionally and more recently tackled the construction and representation of word meaning, or how word meaning is constructed in the artificial mind (Part 2). In particular, as anticipated above, I focus on the distributional hypothesis (Harris, 1954), which I claim constitutes the cornerstone principle of how words get their meaning. The

distributional hypothesis, summarized by Firth (1957) as "you shall know a word by the company it keeps", was initially proposed by Zellig Harris, an American linguist concerned with understanding the mathematical and empirical foundations of language. Influenced by Bloomfield's structuralism (Harris, 1973) and by Sapir's theories of linguistic relativity (Harris and Mandelbaum, 1951), Harris realized that the functioning of language could not be easily explained by appeal to *a priori* principles and rules, but rather by appeal to its use, and by how words are used in combination with other words.[1]

When in the early Nineties large amounts of linguistic data became available in digital format, and thus readable by machines, the distributional hypothesis was quickly adopted by computer scientists and computational linguists to implement models of word meaning representation, giving birth to the field of distributional semantics. Pioneering distributional semantic models such as Latent Semantic Analysis (Landauer and Dumais, 1997), exploited the regularities in linguistic co-occurrences to construct word meaning representations based on how words are used in linguistic contexts. However, such models have been heavily criticized by several cognitive scientists, cognitive psychologists and neuroscientists. As a matter of fact, the concurrent rise of the grounded/embodied account of cognition in the Nineties, and the idea that language processing is grounded in perception, action, and emotion (e.g., Pecher and Zwaan, 2005; Barsalou, 2008) was incompatible with the idea that the semantic representation of word meaning could be acquired or constructed by looking simply at word co-occurrences across corpora of text.

As I will explain, the great theoretical debate on the nature of word meaning revolves around the notion of mental symbols and their controversial origin (the *symbol grounding* debate, cf. Harnad, 1990; De Vega, Glenberg and Graesser, 2008; Bolognesi and Steen, 2018). Ironically, such debate in the past few decades has typically involved cognitive scientists, computer scientists, and philosophers (among others), but rarely linguists and experts of language. However, empirical research on first and second language acquisition (which I will review in the next three chapters) shows that word meaning is constructed on the basis of the detection of statistical regularities. Building on such findings I will bring the debate on the nature of word meaning and the mental symbols we use to represent it in the

1. Among the pioneering scholars who advanced the idea that words get their meaning from patterns of use it is worth mentioning Osgood (1952) and Wittgenstein (1953). In particular, the mediated theory of word meaning proposed by Osgood and developed further in Osgood, Suci, and Tannenbaum (1957), presents many ideas that give them historical priority when discussing co-occurrence models. I am thankful to an anonymous reviewer for pointing out the historical perspective on this matter.

field of linguistics and elaborate the implications that such debate has for language studies and communication sciences (Part 3 of this book).

In particular, I argue that the construction, representation and organization of word meaning comes from connections that we establish between words and elements perceived in experience (word-to-world associations) as well as from connections that we establish between words and other words (word-to-word associations). I will show that the distributional hypothesis (which in the Nineties was initially applied to words co-occurrences only, within the community of computational linguists, machine learning and nlp scholars) is equally applicable to both these types of connections. Then, I will argue that the overall process through which we construct and organize semantic representations and their mutual relations, which starts from word-to-world and word-to-word associations, involves two more steps, both typically implemented in distributional semantics, and at the same time both widely supported by cognitive scientific evidence: a pattern detection mechanism, and a mechanism in which (broadly defined) paradigmatic similarity between meanings is constructed, by means of feature matching processes. Finally, I explain that different types of features determine different types of similarity between semantic representations and between word meanings. If the features shared by two word meanings are linguistic (i.e., shared word-to-word associations) we may obtain different types of similarity between the word meanings than if the shared features are experience-based (i.e., shared word-to-world associations).

The more specific questions that I will address in this book can be summarized as follows: How do the word-to-world and word-to-word associations contribute to the construction of word meaning in our mental lexicon? How do children and adult language learners learn new word meanings? And what can the latest endeavors in machine learning and AI contribute to our understanding of the structure and functioning of the human mental lexicon?

As I will explain, the cognitivist turn that characterized the study of language and cognition in the past few decades and that enabled the emergence of the grounded cognition framework focused on understanding how word-to-world associations work in the construction of semantic representations, neglecting the importance of the other side of the coin: the word-to-word associations. This half, however, is crucially important because it enables humans to manipulate and combine symbols to construct abstract concepts, a hallmark of human cognition. I will argue that the exquisitely human ability to establish word-to-word associations and to extrapolate word meaning from other words is based on the same exact principles that allow us to categorize experiences and construct word meaning from them.

1.2 Outline of the book

The book is divided in three parts.

Part 1 deals with word meaning construction and representation from the point of view of language acquisition. In this part I focus on the construction and representation of word meaning in first and second language. In particular, in **Chapter 2** I start off by explaining how word acquisition has been traditionally approached in classic models that can be found in the literature on linguistic development. Such models assume that word meanings are learned on the basis of natural inner constraints and rules that infants and children have and follow. I then continue by reviewing a fairly recent bottom-up approach, namely cross-situational learning, which is supported by an increasingly large body of empirical literature, which shows that infants behave in ways that demonstrate they are sensitive to the statistical structure of the input and can learn word meaning across multiple exposures to perceptual experiences without any predetermined rule, despite exposure-by-exposure uncertainty as to the word's true meaning. This approach, however, does not elucidate straightforwardly how the meaning of abstract words is acquired, a problem that I address later in the book, when I introduce the distinction (and the similarity) between word-to-world and word-to-word association mechanisms. Finally, I explain in a qualitative manner through some examples that words do not only label concepts, but they also drive their construction and force the search for similarities between items that are included within the same conceptual category.

In **Chapter 3**, I focus on how new word meanings can be derived from old ones by maintaining the same word form, thus generating phenomena of polysemy, which can often be motivated by metaphorical or metonymic extensions. In this chapter I also start to introduce the approach that characterizes the whole book, that is, the constant parallel between theoretical models, empirical evidence on the cognitive processing of linguistic phenomena related to word meaning, and methods, challenges and findings emerging from the computational modelling of such phenomena. In relation to word meaning extension based on metaphoric and metonymic shifts, I review some recent computational models that tackle these phenomena, and I anticipate that these appear to be based on the distributional hypothesis.

In **Chapter 4**, I focus on the structure of the bilingual mental lexicon and on how word meaning is constructed and represented therein. As in Chapter 2, I start off by describing a classic (quite static) model of the bilingual mental lexicon based on modules and then proceed to argue in favor of a bottom-up approach that finds evidence in the way word associations are indicated by native speakers and language learners. I then provide a review of empirical evidence supporting

the phenomenon of incidental vocabulary learning, that is, the tendency by which non-native speakers tend to learn word meanings indirectly, mostly during reading activities, by being repeatedly exposed to texts in the target language. I then proceed to introduce the idea that both cross-situational learning in L1 and incidental vocabulary learning in L2 are based on similar mechanisms: the repeated exposure to input, and the detection of patterns which both, children learning their first language and adult foreign language learners, tend to exploit and to use to construct word meaning. This statistical approach to word meaning construction, I argue, differs in the L1 and the L2, in that it seems to be based mostly on word-to-world associations in the L1, and mostly on word-to-word associations in the L2. These hypotheses are then investigated in greater detail in Chapter 6.

Part 2 of this book deals with the computational side of the story: how word meaning construction and representation is approached in what I call 'the artificial mind', as opposed to the human one. I focus in particular on distributional models of word meaning and on the recent developments of word embeddings, which exploit neural networks to construct the vectors (i.e., sequences of numbers) that represent word meaning. In **Chapter 5** I describe the functioning and the basic mechanisms in which the distributional hypothesis was implemented in the first pioneering distributional models, and I focus on the Latent Semantic Analysis model, the most widely used. I describe the mechanisms that underlie the functioning of these models, and illustrate them through explanatory figures. I also provide a general overview of the different types of distributional models that have been implemented in the past two decades, and explain the main differences between them. In particular, I explain how word embeddings (based on neural networks) differ from classic distributional models, describe their basic implementation using the popular word2vec method as an example, and explain that both, classic models and more recent word embeddings are ways to construct word meaning representations by means of vectors of distributed features.

In **Chapter 6** I focus on how distributional models based on corpora of text are typically evaluated against behavioral data. I explain that using psychological data as a baseline to measure and evaluate the performance of these computational models led inevitably to the following inference: if the performance and therefore the output is comparable between the human and the artificial mind, then the processes are also equivalent and comparable, bit to neuron and neuron to bit. I illustrate how this idea was first proposed to support the cognitive plausibility of Latent Semantic Analysis back in 1997, and how it led to a very heated debate on the nature of meaning and of linguistic symbols. Then I focus on the cognitive nature of the associative mechanism that characterizes the implementation of associations in distributional modelling. I describe how the mechanism of conditioning, first observed in animal behavior and then in humans, can be linked to

the (broadly speaking) associative mechanisms through which we connect words and referents as well as words and other words. Within the associative mechanism of conditioning, I highlight the importance of negative feedback (the unobserved occurrence of expected associations) which is highly informative for updating the associations established between two items. Negative feedback (the core feature of discriminative learning as opposed to simple associative learning) is used for example in word embeddings to weigh the associations between words and features. Finally, I report the results of an empirical study (Bolognesi, 2016a) in which I compared the organization of word meaning obtained through a distributional model, to the organization of word meaning emerging from behavioral data collected in L1 and L2 respectively. Discussing the results of this study I finally claim that semantic representations can be both grounded in perceptual experience as well as symbolic, based on word-to-word associations. Adult L2 speakers tend to rely more heavily on this latter type of association to construct word meaning.

In **Chapter 7** I focus on the integration of extra-linguistic information in the implementation of distributional models, showing how such an endeavor has been approached in more recent years in different ways. I explain that a major problem involved in the construction of multimodal representations of word meaning relies in the mechanisms used to combine the information retrieved from the different sources. In particular, while collecting word-to-word associations alone from corpora of texts is a rather straightforward operation, combining perceptual features extracted from extra-linguistic contexts with the linguistic information extracted from texts is a complex task because it remains unclear what would be the balance between the two streams, and how shall they be translated in the same machine readable format, within the same vector. I then provide an overview of the functioning of Flickr Distributional Tagspace (Bolognesi, 2014; Bolognesi, 2017a) a distributional model based on an inherently multimodal stream of semantic information, that is the metadata (tagsets) that users associate to annotate (i.e., to tag) their personal pictures, uploaded on Flickr, the photo hosting service powered by Yahoo!. Finally, I briefly explain how the distributional hypothesis affords the implementation of world-to-world associations, bypassing words altogether. In this brief overview I focus on a pioneering computational model, Perceptron, a neural network capable of classifying and categorizing non-linguistic inputs on the basis of solely perceptual information. The grandchildren of this model are nowadays used to perform, among other tasks, image recognition used in modern AI.

Finally, **Part 3** of this book is dedicated to the elaboration of the converging evidence in language and communication research, obtained from the comparison between the way in which the human and the artificial minds construct and represent word meaning. In **Chapter 8** I elaborate the points anticipated in Chapter 4, related to where words get their meaning. In particular, I explain how

language and experience, respectively, construct categories, and how these two mechanisms function for the construction of different types of word meaning (I focus on meanings denoting concrete and abstract concepts respectively) and how they function for different types of speakers (I focus on native speakers and foreign language learners).

In **Chapter 9** I explain that the distributional hypothesis proposed by Zellig Harris in the Fifties and adopted to construct computational models of word meaning has indeed deep cognitive foundations and it is based on cognitive mechanisms and principles that have been widely acknowledged to be part of the human cognitive system. I explain how this hypothesis has been largely misunderstood and misinterpreted by cognitive scientists in the past decades, and therefore, in my opinion, erroneously rejected. I also explain how, by interpreting the distributional hypothesis in a broader sense, it is possible to fully appreciate its potential and its ability to explain how humans are capable of climbing the ladder of abstraction to construct generic categories and abstract concepts, which are pervasive features of human language. I also explain how the correct interpretation of the distributional hypothesis can explain mechanisms beyond the construction of metaphoric and metonymic extensions of word meaning.

Finally, in **Chapter 10** I elaborate the practical implications that a distributional view of word meaning has, in applied fields of language and communication science, such as in the field of AI research, in the study of human (linguistic) creativity, in the field of first language acquisition, and in the fields of second language acquisition and foreign language teaching.

1.3 What this book is about and what it leaves out

This book is about word meaning. Words are intended as linguistic symbols that we use to label conceptual representations created in our mind on the basis of similar experiences grouped together to form categories. Experiences can be similar to one another in different ways: they can be perceptually similar (e.g., cups and glasses are similar on the basis of the features that we perceive through our sensory modalities and our perceptual experiences with these two objects) or they can be similar on the basis of other, non-perceptual features (e.g., couches and lamps are grouped together in the generic category of furniture based on their shared function which is to make an indoor environment more comfortable or functional). Moreover, language itself has the power to create categories of experiences: labelling a group of items or of experiences with the same word stimulates us to find criteria on which such items can be clustered together.

Because this book is about word meaning, it focuses on those words that are rich in meaning. Most examples reported throughout the book consist of nouns and verbs. Function words such as prepositions, articles and pronouns are not used as examples. However, the same mechanisms of meaning construction and representation are, in principle, applicable to the construction of the (impoverished) meaning of function words too.

Focusing on word meaning construction, this book deals with the semantic aspects of words, more than with their syntactic combination. Nevertheless, the syntagmatic associations that words entertain with other words to form, for example, collocations, are a core aspect of the mechanisms by which word meaning is constructed, as I describe in this book. In this sense, grammar intended as a set of rules that governs how words are used and combined is *not* the focus of this book, but the combinatorial patterns by which words tend to be used together with other words are a cornerstone mechanism through which word meaning is constructed. The syntactic patterns that form what we usually call 'grammar', in this view, emerge in a bottom-up manner from statistical regularities observed in language use, based on the meaning of words and their relations with other words.

1.4 A final remark on the parallel between human and artificial mind

Throughout this book I compare and discuss the similarities (and the differences) existing between the ways in which the human and the artificial mind respectively construct, represent and organize word meanings. These parallels might sound evocative of the computational theory of mind that characterized most of the second half of the Twentieth century (e.g., Fodor, 1983). Computationalism, however, is a specific form of cognitivism in which it is argued that the mind operates by performing purely formal operations on symbols, in a top-down manner, using a 'language of thought' which is made of rules that are applied to symbols like in a classic Turing machine. The approach hereby described, instead, focuses on learning (word meaning) from external stimuli in a bottom-up manner by means of associations, and constructing meaning on the basis of such connections. Such approach could not be more different from the rule-based, top-down, syntaxis-focused, formal approach that characterized the classic cognitivist view and the computational view of mind. If anything, this approach can be rooted in connectionist accounts of cognition, such as those that led to the implementation of the first (artificial) neural networks, which in fact were strongly criticized by classic computationalism. Such a connectionist approach, emerged already in the Fifties, aimed at implementing self-organizing systems based on pattern recognition and parallel distributed processing that proceeded in a purely bottom-up manner,

inspired by principles such as the Hebbian synaptic plasticity (i.e., when two neurons fire together, the synapse between them strengthens), and later adjusted by studies focused on discriminative learning, conducted on animals (e.g., Rescorla and Wagner, 1972; Rescorla, 1988).

The parallel that emerges in this book between the mechanisms employed by the human and the artificial mind to construct and represent word meaning is therefore very different from the metaphor of the brain as a computer suggested by classic computationalism. As I will defend throughout this book, the parallel between the artificial and the human mind, proposed to develop the converging evidence and to support the cognitive foundations of the distributional hypothesis, affords a view of the artificial mind (which I will describe) that can learn and construct meaning with minimal intervention from the programmer. Such view reflects the contemporary achievements obtained in machine learning and (generally speaking) AI.

PART 1

Word meaning construction and representation in the human mind

CHAPTER 2

Word meaning mental representation

2.1 Learning words: A developmental perspective

Most children articulate their first words when they reach the first year of age, but begin to learn their first language much earlier, when they start detecting sound-related features of the speech signal to which they are exposed and start categorizing consonants, vowels, and combinations of these sounds (Polka and Werker, 1994). Learning to *understand* words, as opposed to just *perceiving* their sounds, is a more sophisticated cognitive capacity that children develop when they are capable of interpreting others' goals and intentions. This is commonly acknowledged to start happening only at around 9–10.5 months, although some scholars argue that this happens already around 6–9 months (e.g., Bergelson and Swingley, 2012) and is used to explain the earliest emergence of word learning shortly thereafter (e.g., Carpenter, Nagell, Tomasello, Butterworth and Moore, 1998).

Learning the sound structure of a language implies discerning the elementary sounds (and their combinations) that are used in language, by making discrete a continuous stream of auditory signal. Such segmentation is a form of categorization: sounds that are articulated in similar ways are grouped together under the same category (Kuhl, 2004). Categorization takes place at all levels in language understanding, conceptual processing, and representation: from the detection and classification of sounds at the early stages of cognitive and linguistic development to the categorization of experiences and linguistic information to shape the content of complex abstract concepts such as LEGACY[1] or DEMOCRACY, and the relative meaning of the words denoting such concepts, *legacy* and *democracy*. In other words, categorization is the hallmark of human cognition, and its mechanisms will be widely discussed in this book.

Although learning about perceptual regularities in speech reveals remarkable analytical skill, it is generally accepted that young infants able to discriminate sounds and group them under the same categories do not know yet the meanings of the words they manage to segment (e.g., Thomas, Campos, Shucard, Ramsay and Shucard, 1981; Swingley, 2009). Around their first birthday, children start to

1. Throughout the book I will use italics to indicate words and capital letters to indicate concepts, a standard practice in disciplines such as cognitive linguistics.

figure out what the bundles of sounds that they were able to extract from the continuous stream of sound signal *mean*. But how do they manage to match meaning to sounds, therefore learning word meanings?

The process of word meaning acquisition involves a non-trivial mapping process that starts with the major obstacle of referential ambiguity (Quine, 1960 [2013]): in any naming event, a novel word can in principle refer to any entity present in the given situation, its properties, its position, the speaker's feelings or intentions for it, the actions that can be performed with it and so on. Because in any naming situation there are virtually infinite possible meanings that a child can attribute to an unknown word, the question arises of how do children face such daunting task of solving the immense ambiguity and matching word forms with the correct meaning. The classic literature on first language acquisition has addressed this issue over the past decades, and identified some basic principles that seem to constrain and bias the way in which children attribute meaning to words (e.g., Markman, 1990; Clark, 1995). The main biases can be summarized as follows, for the construction of meaning for concrete nouns:

Whole object bias

> A new word refers to a whole object, not to components or actions involved in its usage, unless specifically indicated. For example, given a teddy bear, the word *teddy bear* uttered by a parent is attributed by the child to the whole toy, not just its ears, its fur, or its color.

Taxonomic bias

> A new word articulated together with old ones denotes a member of the same kind. In particular, while children usually tend to favour thematic relations (e.g., grouping together a monkey with a banana rather than a monkey with a bear, where both are animals) when they are given a new label, they shift their attention to taxonomic relationships. Therefore, instead of grouping together the monkey and the banana, they group together the monkey and the bear, because they denote entities within the same taxonomic category.

Basic level bias

> Basic-level categories (rather than super- or sub-ordinate categories, as defined by (Berlin, Breedlove and Raven, 1973) are chosen by default. In other words, a new label likely refers to the everyday 'level' of naming hierarchy. For example, the word *monkey* is likely to denote the basic level category of monkeys, rather than a specific sub-category such as macaques or a more generic category such as mammals.

Mutual exclusivity

> One label is attributed to one object. Novel labels denote novel objects. For example, given three words and three objects, if a child knows already the meaning of two words and therefore can correctly associate them to two of the three objects, the third word automatically is used to label the third object. Therefore, every two forms contrast in meaning. This bias implies that children initially reject semantic relations such as synonymy, hypernymy and hyponymy, where the same object can be named with different words. As a matter of fact, these tendencies have been empirically observed (e.g., Clark, 1995).

The mutual exclusivity principle (also called principle of contrast) appears to be particularly strong and may lead children to overwrite previous assumptions predicted by the whole object bias. Therefore, given an object for which a child already has an associated word, if a new word is provided for the same object the child will assume that the new word denotes different aspects/parts of the familiar entity (Markman and Wachtel, 1988). As in the original example provided by Quine, if children presented with the label *gavagai* and a picture of a rabbit are already familiar with the entity displayed and its actual name, *rabbit*, then they may attribute the label *gavagai* to a part of the rabbit, such as its ears.

Whether or not such basic principles are innate, most of the classic scientific literature on first language acquisition assumes these as cornerstones. In the past decade, however, some empirical studies started to question the nature of these principles. For example, Horst and Samuelson (2008) observed that the principle of mutual exclusivity, which children seem to apply in order to infer the correct name of a previously unknown object, does *not* necessarily result in long-term learning: words that are learned during the naming tasks in which children applied these principles do not seem to be retained in the long term memory. A related recent view suggests that the classic constraining principles and biases may be completely unrelated to word learning. Solving the referential ambiguity in specific naming situations may trigger in children the activation of goal-directed strategies, which would be used by the child to solve a *temporary* problem, such as pairing a new label (a word) with an unknown object in the specific situation in which an adult speaker shows a child these two items (the word and the object) and clearly expects something from the child. This, as Horst and Samuelson pointed out, and McMurray, Horst and Samuelson (2012) argued further, does not equal *learning*. The goal of solving the referential ambiguity and therefore solve a temporary problem in a specific situation is a *situation-time* goal, as the authors point out, while word learning relates to the more extensive *developmental-time* goal.

Another critique raised against these principles suggests that word learning is strongly affected by the information that children retrieve from social interactions. Children would therefore solve referential ambiguities by integrating information retrieved from the context and from the social interaction with speakers present on the scene (Baldwin, 1991; Tomasello, Strosberg and Akhtar, 1996; Tomasello, 2003). According to this view, called Social Learning, children learn word-object associations relying on inferences aimed at detecting speakers' intentions. Such inferences are affected, for example, by the speakers' facial expression, gaze direction, and tone of voice (Tomasello and Barton, 1994). The cognitive improvement of social skills seems therefore to foster word learning (Yu and Smith, 2012), allowing children to accumulate new word meanings by exploiting with an increasing pace the subtle clues extracted from their social interactions.

Finally, the classic constraint approach to word learning does not provide much information on how children's vocabulary manages to grow beyond the acquisition of the first words. For example, the taxonomic constraint may explain how children acquire basic-level terms, but it does not explain how children eventually overcome such bias to learn the meaning of super-ordinate terms, which implies that they accept the legitimate attribution of two labels (expressed at different taxonomic levels) to the same referent. Similarly, the constraint-based approach to word meaning acquisition does not explain in detail how children are able to learn synonyms, and therefore accept that two word forms expressed at the same level of abstraction may refer to the same object. The constraint approach does not elaborate in detail all the consequences, interferences and combinations of the identified biases, and does not seem to form a comprehensive framework for how these rules interact with one another, and how conflicts between them can be resolved, to enable children scaling up their vocabulary and learning thousands of new word meanings in the span of a few years.

In more recent years, around a decade ago, a new, compelling paradigm was proposed to address these problems and explain how children learn to associate word forms to their designated referents. Such paradigm is commonly called *cross-situational leaning*.

2.2 Cross-situational learning

If the constraints and biases previously identified in the literature do not properly explain how children learn word meanings and scale up their vocabulary, then how do young speakers actually learn word meanings? Cross-situational learning may provide an answer to this question (e.g., Smith and Yu, 2008; Yu and Smith, 2007).

The basic idea behind cross-situational learning is fairly simple: if multiple entities are under consideration for the possible attribution of a label (that is, a word), then in principle multiple associations can be laid down between the label and each of the candidates. Thanks to regular and repeated exposures over multiple naming events, those linkages that are recurrently stimulated become more consistent and may eventually become established connections between words and referents (Smith and Yu, 2008; Yu and Smith, 2007). This mechanism is based on the assumption that, across situations, there may be only one entity (across many possible candidates) consistently paired with the target word. For example, consider the word *ball*, hypothetically articulated by a parent in different experiential contexts: in one situation *ball* may be uttered in a situation where there is a dog playing with a ball in a park; later on the word may be uttered again in the situation of a playdate with two kids in a bedroom where there are teddy bears, a ball and a few other toys; later on the word ball may appear again, articulated in the context of a soccer game watched on tv, and so on. Across all these situations, connections between ball and each of the salient objects appearing in each of the contexts may be established (e.g., between *ball* and the dog, *ball* and the teddy bear, *ball* and the tv etc.). However, over multiple exposures, the object 'ball' becomes the object that most frequently occurs whenever the word *ball* occurs. The child may therefore only need to accumulate co-occurrence statistics to learn the mappings between words and referents (McMurray, Horst, Toscano and Samuelson, 2009; Yu and Smith, 2007). This bottom-up approach to word learning exploits associative mechanisms[2] that are not constrained by top-down (possibly innate) principles, such as those postulated in the literature described at the beginning of this chapter.

Yu and Smith reported the results of the first empirical analyses aimed at testing the mechanisms of cross-situational learning and its use by both adults and children. In a series of experiments, adults (Yu and Smith, 2007) and infants (Smith and Yu, 2008) were exposed to small artificial lexica, constructed with the purpose of exposing the participants to the situational regularities exemplified above with the example of *ball*. The authors showed that both adults and infants rapidly learn multiple word-referent pairs by accumulating statistical evidence across multiple situations, which, taken individually, featured ambiguous word-object pairings. The authors argued that the indeterminacy problem is solved *not* in a single trial but *across* trials, and *not* for a single word and its referent but for a whole data set of many words and referents processed in parallel. Therefore, word learning may proceed not by solving referential ambiguity with pre-existing

2. Associative is hereby used in a broad sense: as I will explain later, also non-associations can be informative. The relevance of negative feedback in learning new associations is explained in detail in Chapter 6 (in relation to associative and discriminative learning).

biases and constraints, but simply thanks to bottom-up inferred statistical and associative linkages between words and referents that strengthen or weaken with the exposure to multiple situations. Since the first empirical evidence supporting this paradigm, cross-situational learning evolved to explore different aspects and subtypes based on this basic principle (see Cassani, Grimm, Gillis and Daelemans, 2016 for a review). Typically, the different possible mechanisms involved in cross-situational learning are implemented in computational models that run simulations and the performance of the model is measured against the empirical evidence collected from behavioral data. For example, Cassani and colleagues (2016) compared the performance of four different types of computational models based on cross-situational learning, and matched the predictions made by each of them against behavioral evidence. In particular, the authors compared a basic associative model based on simple co-occurrences, a model that learns from co-occurrences and from missed co-occurrences (aka a discriminative learning model, explained in more detail in Chapter 6.3), a probabilistic model that computes the probability over referents for each word, and a model that sets a single hypothesis for each trial and proceeds to test it immediately and then change it if needed (see Chapter 4.4 for a description of this model). The authors conclude that the discriminative learning and the probabilistic models are better predictors for behavioral evidence.

In general, cross-situational learning has been illustrated by Smith and Yu in the way reported in Figure 1.

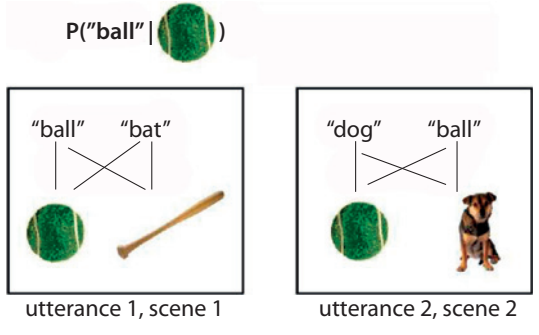


Figure 1. Borrowed with permission from (Smith and Yu, 2008). Associations among words and referents across two individually ambiguous scenes. The probability of the word *ball*, given the object ball is calculate across multiple scenes where *ball* appears (or does not appear) together with multiple possible referents

Here, the probability that the word *ball* refers to the object ball reaches 1 (full certainty) across just two trials. As a matter of fact, looking at trial 1 only, both the tennis ball and the baseball bat have the same probability of being the referents

of *ball*. Looking at trial 2 only, both the tennis ball and the dog have the same probability of being associated with the word *ball.* However, combining the two trials one after the other and thus accumulating experience, only the association between the word *ball* and the object 'ball' is supported, and consequently the association between *bat* and the object 'bat', and *dog* and 'dog' are supported too.

Cross-situational learning can be modelled by means of network-based parallel processing, and therefore relates to classic connectionist models of language learning and processing (e.g., Munakata and McClelland, 2003). In classic connectionist models, language learning and processing can be illustrated by interconnected networks of simple units organized in layers, as in the example displayed in Figure 2, where (by convention) each vertical stack of nodes represents a layer. In principle, units (i.e., the circles in the figure) can be configured into phonemes, whole words, single neurons, and so forth, depending on the aims of the model. In Chapter 5 I will explain how the nodes between the input and the output layer are configured into features that contribute to shape the content of word meaning representations. Within a cross-situational learning paradigm, one can see the nodes as words and referents, as displayed in the example reported in Figure 1 and formalized in Figure 2. In the generic model displayed in Figure 2, the first layer of nodes on the left may represent the three words that appear across the two trials: *ball, bat* and *dog*. The middle layer may represent the three possible referents (the three objects). The arrows between the first and middle layer may represent the possible connections that can be in principle established between the three words and the three objects. The weights associated to each arrow between the first and the second layer are learned through the exposure to multiple contexts in which each word appears with one or more of the objects. In Chapter 6 I will explain in more detail how are these weights can be established and then weakened or strengthened, exposure after exposure. I will explain how the associations between nodes are updated when two entities (or a word and a referent) appear in the same context, as well as when the expected association fails to occur. Finally, the last layer on the right may represent the meaning of each of the three words, which results from the strongest connection between each node in the input layer and a node in the middle layer.

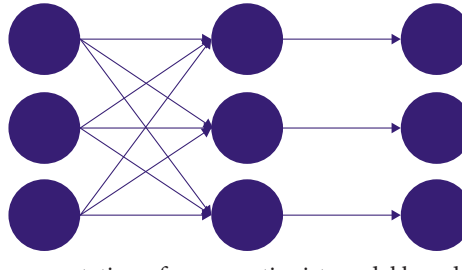


Figure 2. A general representation of a connectionist model based on a network

Connectionist models have been immensely exploited in computer sciences and AI for their regular structures that naturally lend themselves to computational modelling (something that we will describe and discuss extensively in the second and third part of this volume, starting from Chapter 5). An important aspect to take into account, which will also be explained in detail in Part 2 of this book, is the flexibility and dynamicity that these models allow, which reflects the dynamicity and flexibility of the human cognitive architecture for word meaning construction. In particular, in a typical connectionist model, the linkages between nodes are learned through the exposure to external input and are strengthened in a dynamic manner across multiple exposures. Similarly, the way in which word meaning is learned is subject to modifications and updates, due to the ability of children and adults to modify and correct previously acquired knowledge. The mechanisms that allow updating the connections between a word and a referent, or between two words, can be systematically simulated by algorithms that exploit cognitive principles. Moreover, learning involves creating predictions and testing them in communicative settings: children try out the new words, and are ready to modify and recast their acquired knowledge when they are corrected. As we will see in greater detail in the third part of this volume, which collects the converging evidence derived from cognitive science and computer science, the cognitive operations involved in the construction and representation of word meaning are based on mechanisms that have been widely used in computational modelling in the past decades, as well as strongly rejected by cognitive scientists and neuroscientists, because of the erroneously assumed equivalence between the human and the artificial architecture of the mind. Such (metaphorical) equivalence, however, has been formulated on the wrong terms. Once the parallel between the biological and the artificial architectures of word meaning construction is recast and reformulated in different terms, it will be possible to appreciate the similarities between the two systems and the converging evidence that they provide for word meaning construction in language and communication research.

To conclude, humans demonstrate to have from a very young age a remarkable capacity to detect regularities in the environment, starting from regularities in perceptual input. As Saffran and colleagues (Saffran, Newport and Aslin, 1996) pointed out, infants are sensitive to statistical regularities even when they are exposed to a continuous stream of an audible artificial language, and are able to distinguish syllable sequences that are typically used to construct words from improbable syllable sequences. This capacity, which usually goes under the generic name of statistical learning, seems to explain fairly well how children manage to associate together word forms to objects, thus learning word meanings across the exposure to different situations (cross-situational learning). The mechanism described by cross-situational learning, however, may explain how word forms are

associated to objects, thus solving the problem of referential ambiguity, but does not explain how the meaning is acquired of words that denote abstract concepts, which are concepts for which, by definition, there isn't a concrete, tangible referent in the world that can be perceived through our senses and therefore associated to a word form by means of cross-situational learning in a direct and straightforward way. The next section explains in further detail the differences in the constitution of meaning, between words denoting concrete and abstract concepts.

2.3 Words denoting abstract vs. concrete concepts

Learning word meaning does not always occur in transparent situations in which both, a label and its referent are present. Referents, especially, are not always present in the physical environment for different possible reasons. First of all, we can talk about things that are not physically present in the immediate surroundings, because they are in a different spatial location, for example a mother can tell her child "go get the ball!", assuming that the child knows that the ball is in her bedroom and will go there to get it. Second, we can talk about past and future events, and in these cases the mentioned referents are also absent from the immediate surrounding, because they are located in a different temporal dimension (e.g., "tomorrow I'll get you a new ball"). When we talk about concrete entities that are not physically present in the exact moment in which the communication unfolds, because they are located in a different space or in a different time, speakers and listeners can still simulate[3] the referent, relying on previous encounters with such referent. For example, if a child hears the word *ball* but a ball is not present in her visual field in that precise moment, she can still create a mental image of the ball, based on previous encounters with this object.

However, human language allows us to talk also about entities that do not have a tangible referent at all, such as ideas, dreams, and feelings. This is the case for words such as *love*, or *surprise*, or *idea*, which denote abstract concepts and describe intangible referents. How these words are then learned? What sort of referential linkage do children create between these words and something out there, in the world? How can these words be simulated and represented in the mind? These questions do not have a simple and definitive answer, and are currently debated in various disciplines, including linguistics, psychology and neuroscience (Bolognesi and Steen, 2018; Bolognesi and Steen, 2019).

3. On the nature, activation, and necessity of such mental simulations I will talk more extensively in Chapter 7.

From a developmental point of view, a number of studies show that words denoting abstract concepts are acquired later, compared to words denoting concrete concepts, suggesting that the former type of words may be more difficult to learn than the latter (e.g., Barca, Burani and Arduino, 2002; Gleitman, Cassidy, Nappa, Papafragou and Trueswell, 2005; Ponari, Norbury and Vigliocco, 2017; Vigliocco, Ponari and Norbury, 2018). The reason why abstract words would be learned later than concrete words, however, remains hotly debated, and the lack of a tangible referent in the world to be directly associated to abstract words seems to explain only part of the problem.

The different average age of acquisition that characterizes words denoting concrete and abstract concepts relates to the different nature of these two types of conceptual categories. In particular, concrete concepts, such as BANANA or CHAIR, labelled by the corresponding words *banana* and *chair*, categorize referents in the world that share perceptual features. Such shared perceptual features play a prominent role in shaping the cognitive representation of these conceptual categories in the mind of the speaker (e.g., McRae and Jones, 2013; Borghi and Binkofski, 2014). Conversely, abstract concepts, such as FREEDOM or TRUTH, labelled by the corresponding words *freedom* and *truth*, categorize intangible referents that are therefore not characterized by perceptual features. Children from a very young age are capable of detecting perceptual features and recurring patterns of perceptual features across different situations and this skill enables them to learn word meanings. However, such word-to-world associations (i.e., the associations learned through cross-situational learning between a word and the correct referent in the world) are not as easily established when the words denote concepts that lack a referent in the environment. By crossing experiences and situations and detecting statistical regularities in the perceptual input, children may realise that *none* of the tangible referents in the world, to which they are exposed across different situations, is consistently present and can therefore be associated to the new word. For example, consider the word *freedom*, and consider a series of situations to which a child may be exposed in conjunction with the word *freedom*, uttered by a parent, as illustrated in Figure 3.

Figure 3. The hypothetical exposure to four situations in which a child may hear the word *freedom* uttered by a parent, and try to associate it to one of the concrete referents in the input

Given the (limited and hypothetical) situations illustrated in Figure 3, a child may establish preliminary associations between *freedom* and a butterfly, a swing, friends, jumping, a fist pointing up, and so on. However, none of these referents repeatedly occurs across all the situations, thus disambiguating the meaning of *freedom*. Therefore, from a strict cross-situational learning perspective, it is impossible for the child to establish a word-to-world direct connection by crossing perceptual experiences, to learn the meaning of words denoting an abstract concept. If crossing perceptual experiences does not help learning word-referent associations, what type of associations shall be established, to construct the meaning of an abstract word? What are the similarities among the hypothetical situations displayed in Figure 3, to which a child is exposed together with the exposure to the word *freedom*, that allow her to group all those experiences together and extract the meaning of *freedom*?

These questions are hotly debated in cognitive science and cognitive psychology and are embedded in the greater debate that sees supporters of embodied/grounded accounts of cognition vs. supporters of symbolic/amodal accounts of cognition (e.g., De Vega, Glenberg and Graesser, 2008; Bolognesi and Steen, 2018), which is described and discussed in greater detail in Part 2 of this book, in particular in Chapter 6. Moreover, a discussion that can provide an answer to the questions stated above will be elaborated in the third section of this book, which brings together the converging evidence derived from the discussions provided in the previous two parts. For the purpose of this chapter, I will limit the discussion to those theories of meaning suggesting that abstract and concrete concepts may consist of semantic information retrieved from two different streams, in different proportions.

A pioneering model in cognitive psychology that has been largely used to motivate the different empirical results obtained for the processing of concrete and abstract concepts is the Dual Coding Theory (henceforth DCT, Paivio, 1983; Paivio, 2010). In this model it is claimed that there are two main ways to represent meaning in mind: one is verbal and one is imagistic. Concrete and abstract concepts, according to DCT, are represented in different ways in the mental lexicon: while concrete concepts would be encoded in both the representational systems, and therefore by means of imagens (i.e., imagistic representations) as well as logogens (i.e., verbal representations), abstract concepts would be encoded only in the verbal system. The double encoding of concrete concepts, as opposed to abstract ones, is then used to explain various concreteness effects widely reported in the literature, such as the fact that words denoting concrete concepts are easier to remember and recall (Dove, 2016 for a review).

More recently, various scholars have supported the idea that different representational systems co-exist in our mind, and contribute in different proportions

to shape the overall meaning of words denoting abstract and concrete concepts. As Boroditsky and Prinz point out:

> neither perceptual information alone, nor the sets of correspondences between elements in language alone are likely to be able to amount to the sophistication, scale, and flexibility of the human conceptual system. Luckily, humans receive heaping helpings of both of these types of information. Combining information from these two input streams, as well as extracting the wealth of information that exists in the correspondences across input streams can help overcome the shortcomings of relying on any single information stream and can reveal information not available in any one stream. (Boroditsky and Prinz, 2008, p. 112)

Both streams of information take part in the cognitive processing of linguistic input: when we hear or read a word, a combination of information derived from our previous perceptual experiences with such word (and the relative emotional responses) and our previous linguistic encounters with such word allow us to process its meaning. In this sense, multiple systems, and not just one, represent word meaning in our mind: on one hand there are linguistic representations that encode information retrieved from language and linguistic structures, and on the other hand there are (embodied) representations based on perceptual experiences. Empirical evidence (further discussed in Chapter 6) shows that the meaning of both abstract and concrete concepts may consist of information retrieved from language statistics *and* perceptual experiences. In this sense, the clear-cut distinction suggested by Paivio may be too extreme; even abstract concepts may be represented in the brain's modal systems by means of representations derived from perceptual experiences (e.g., Barsalou, 2008; Pulvermüller, 2018). The proportions with which the content of these two types of concepts and their relative word meaning is represented, however, may differ, as illustrated in Figure 4: abstract concepts may be made primarily by information accumulated from linguistic encounters with the words denoting such concepts, while concrete concepts may be made primarily by information retrieved from perceptual experiences with the denoted referents.

Finally, if abstract concepts are more consistently shaped by linguistic rather than perceptual information, then the meaning of words referring to abstract concepts may be subject to a greater cross-linguistic variability, compared to the meaning of concrete words. As a matter of fact, for abstract concepts there would be more room for the influence of language, compared to concrete concepts, and different languages may contribute in different ways to shape the meaning of words denoting abstract concepts. In line with this intuition, Gentner and Boroditsky (2001) argued that the meaning of verbs, which are on average more abstract than nouns because they involve a larger relational structure (e.g., Asmuth and Gentner, 2017) varies more across languages, compared to the meaning of

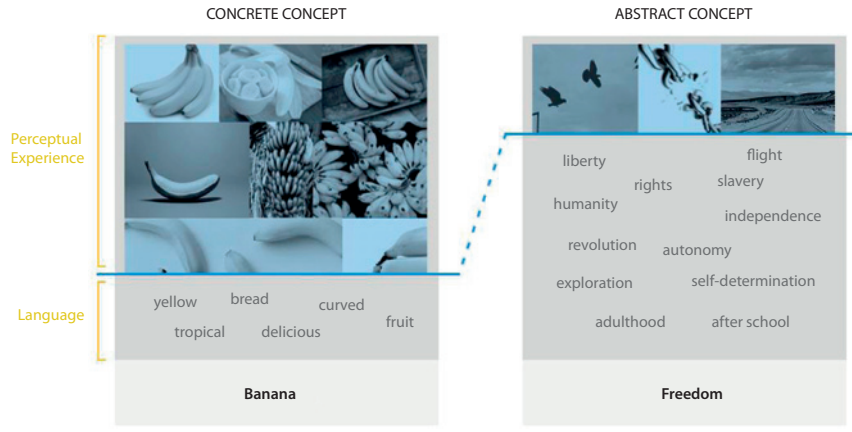


Figure 4. Semantic representations for abstract and concrete concepts, based on two streams of information: The information derived from perceptual experiences and the information derived from language, that is, from the linguistic encounters with the words denoting such concepts

nouns, which is in turn more stable across languages, because it is deeply influenced by perceptual experiences. According to this view, for example, the verb *travelling* is perceived to be on average more abstract than the noun *trip*, because the first relies more deeply on its argument structure, for the determination of its meaning (who is travelling? Where to? How?) while the latter identifies a single entity which, even though does not denote a tangible referent, is mainly defined by its own properties rather than its argument structure. Because *travelling* is more abstract than *trip*,[4] its meaning is more deeply shaped by information encoded in language, while the meaning of *trip* would be more deeply shaped by information encoded in perceptual experience. Another crucial difference between these two words (trip and travelling) is their linguistic structure: the two words belong to different parts of speech and have a different number of morphemes. In a recent investigation (Strik Lievers, Bolognesi and Winter, in prep.) it has been shown that a number of linguistic factors among which "part of speech" and "number of morphemes" are correlated with concreteness: verbs are on average more abstract than nouns, and words with more morphemes tend to be on average more abstract than words with fewer morphemes. These findings suggest that the architecture of language affects word concreteness. Because different languages have different architectures, then different languages might affect word concreteness in different ways.

4. This is confirmed by a search in the database of concreteness ratings provided by Brysbaert and colleagues (Brysbaert, Warriner, & Kuperman, 2014): *travelling* is associated with a concrete score of 3.5 (on a 5 points scale where 5 indicates the maximum degree of concreteness) while *trip* is associated with a concreteness score of 3.71.

And this would be particularly true for words denoting abstract concepts, which are more strongly affected by linguistic information than words denoting concrete concepts. So, words denoting abstract concepts might be subject to a greater cross-linguistic variability, in terms of their semantics, than words denoting concrete concepts. Since different languages construct word meaning is slightly different ways (as further illustrated below), it follows that the meaning of *travelling*, more strongly affected by linguistic information, may be subject to a greater variability among its cross-linguistic equivalents, than the meaning of *trip*. This idea will be further elaborated in Chapter 9.

To support the claim that different languages crop word meaning in different ways, Malt and colleagues (Malt, Gennari, Imai, Ameel, Tsuda and Majid, 2008) compared verbs denoting the actions of running and walking in English, Spanish, Japanese and Dutch. The authors demonstrated that while the two broad types of movement can be roughly distinguished across these languages, the stimulus space is partitioned in very different ways in these languages: for example, the words *jog, run* and *sprint* correspond to a single word in Japanese. Additional evidence on how different languages encode different types of information in word meaning comes from empirical studies in linguistic typology. For example, Talmy (1991; 2003) and Slobin (1996) showed that languages lexicalize different information in their verb roots. For example, in English motion verbs the manner of the movement is typically encoded in the verb (e.g., *stroll, walk, run*), while the path of motion is typically encoded in a preposition or adverb that constitutes the satellite of the verb (e.g., *walk in, walk out*). In Spanish, instead, the path is typically encoded in the verb root (e.g., *entrar, salir*). These studies suggest that word meaning is constructed in different ways across languages and languages crop in different ways and with different levels of granularity the semantic spaces of motion verbs, even when the words seem to refer to the same broad conceptual category. This cross-linguistic variability in word meaning construction is arguably particularly strong for words denoting abstract concepts, because in this case the linguistic information constitutes the main portion of the meaning.

To conclude, both language and perceptual experience contribute to shape word meaning in different proportions for different types of words (notably, for words denoting abstract and word denoting concrete concepts). Moreover, besides partitioning and categorizing perceptual experience language can also *construct* meaning. This point is elaborated further in the next section.

2.4 How words construct meaning

The role of language in conceptual processing goes far beyond the simple labelling function (e.g., Connell, 2018). Words are a driving force in cognition, which allow us to perform cognitive tasks that would otherwise be impossible to perform. Some empirical evidence collected to support this idea relates explicitly to revised versions of the Linguistic Relativity theory, that is, the idea that language shapes thought and that conceptual representations vary across speakers of different languages. Although nowadays the strong Whorfian hypothesis of linguistic determinism has been largely rejected recent research is gradually establishing new connections between language processing, perception, and cognition (Boroditsky, 2001, 2011; Reines and Prinz, 2009; Lupyan, 2008, 2012; Casasanto, 2008; Wolff and Holmes, 2011). This new body of empirical work suggests that language plays an active role in shaping conceptual content. For example, cross-linguistic differences in the lexicalization of the concept TIME correlate with profound differences in the way speakers of different languages mentally represent different aspects of time, such as duration and orientation (Boroditsky, 2001, 2011).

Different languages categorize experience in slightly different ways, and this is particularly visible when we look at the way in which streams of continuous perceptual information are divided into discrete portions across different languages. Consider, for instance, the continuous stream of chromatic information as it appears in the rainbow. Here, hues fade into one another in a continuous stream of color. Different languages partition such continuum in different ways, labelling each portion with a different word. There is now a large and rapidly increasing number of empirical evidence showing that cross-linguistic differences in color vocabularies cause differences in the way speakers categorize, memorize, and perceive the actual colors (Roberson, Hanley and Pak, 2009; Daoutis, Franklin, Riddett, Clifford and Davies, 2006; Winawer et al., 2007; Thierry et al., 2009). For example, Winawer and colleagues presented English and Russian speakers with snips colored in different shades of blue. Russian language distinguishes two basic-level terms (*siniy* and *goluboy*) within the category that English speakers would label with the word *blue*. The participants to the experiment were asked to decide as quickly as possible whether a given shade matched the color displayed on their left or the color displayed on their right. Although all colors were within the category BLUE for English speakers, for Russian speakers the shades could be labelled as *siniy* or as *goluboy*. The authors found that Russian speakers were faster than English speakers to discriminate two colors when they fell into different linguistic categories in Russian (one *siniy* and the other *goluboy*) than when they were from the same linguistic category (both *siniy* or both *goluboy*). This suggests that knowing color words referred to specific shades (e.g., *siniy* vs *goluboy*) improves

one's performance in non-linguistic color discrimination tasks. Word meaning affects perception.

Another domain on which the role of language in meaning construction has been investigated is the domain of emotions. Barrett (2017) argued that our brain constructs emotions by grouping together situations that share very little information and very rudimental elements related to actual bodily sensations. In her view, a word like *sadness*, which denotes the concept of SADNESS, applies to a variety of very different situations in which sad feelings arise in the mind and in the body. While on one hand the word *sadness* helps us to construct and label the concept SADNESS by gluing together instances of situations that share very few common features, on the other hand having a word like *sadness* in our language invites us to find similarities across situations labelled with this word by other people. In this sense, the word *sadness* not only is used to bring and label different experiences together, but it also constructs the meaning of the concept SADNESS, by stimulating us to look for similarities across such experiences. Therefore, on one hand words work as glue, allowing us to group together various experiences that may also share little common features, and on the other hand words drive the search for sameness across such experiences, forcing us to establish a similarity between members of the same category (in this case, the category of situations that represent the concept SADNESS, and can be labelled with the word *sadness*).[5] As Barrett argued, the fact that a word like *sadness* exists in English language invites English speakers not only to group together instances of situations to which this word is applied by other speakers, but also to search for a motivation for this sameness. Words invite us to believe in an essence, in the idea that concepts (labelled with words) must have a core. William James had observed this phenomenon already a century ago, when he wrote:

> whenever we have made a word […] to denote a certain group of phenomena, we are prone to suppose a substantive entity existing beyond the phenomena, of which the word shall be the name.
>
> (cited in Barrett, Mesquita and Smith, 2010, p. 1)

5. This view, applied by Barrett to the construction of emotions, has been inspired by various previous scholars. Notably, John Stuart Mill (1869), in *Notes to Analysis of the Phenomena of the Human Mind* mentions: "the tendency has always been strong to believe that whatever received a name must be an entity or thing, having an independent existence of its own; and if no real entity answering to the name could be found, men did not for that reason suppose that none existed, but imagined that it was something peculiarly abstruse and mysterious, too high to be an object of sense. The meaning of all general, and especially of all abstract terms, became in this way enveloped in a mystical base". I am thankful to a reviewer to point out this quote, to give historical perspective to the discussion.

Words do not just name categories: they encourage a very basic form of essentialism.

Emotions are abstract, intangible concepts, and therefore the meaning of words denoting emotion is subject to great cross-linguistic variability, as I will further explain in Chapter 10. Many words denoting emotions are language-specific and very hard to translate into other languages. For example, Danish speakers use the word *hygge* when acknowledging a feeling or moment, whether alone or with friends, at home or out, ordinary or extraordinary as *cosy, charming* and *special*. *Hygge*, according to Danes, requires consciousness, a certain slowness, and the ability to not just be present – but recognize and enjoy the present. This word does not translate easily into other languages. Similarly, the German *Schadenfreude* is defined as the experience of pleasure, joy, or self-satisfaction that comes from learning of or witnessing the troubles, failures, or humiliation of another. This emotion, constructed and lexicalized by German speakers in the word *Schadenfreude*, does not have a direct equivalent in other languages (unless the word itself is used as a borrowing). This does not mean that non-German speakers cannot understand *Schadenfreude*: they can, and they are likely to have experienced it, even without knowing the word. But because non-German speakers do not have a word to label this phenomenon, their brains would have to work harder to construct such concepts and acknowledge those emotions.

A practical and quite controversial example of how a word may force the identification of a substantive entity existing beyond the phenomena labelled with such word can be observed in the way names for pathologies are created, once a pathology is identified. Such neologisms have incredible consequences also on diagnostics. For example, the contemporary concept of attention deficit hyperactivity disorder (ADHD) which is commonly acknowledged today as a mental disorder in many countries, was recognized by the American Psychiatric Association only when it was published within a revised version of the official Diagnostic and Statistical Manual of Mental Disorders published in 1987. Since that date, ADHD cases began to climb significantly and continue to do so at an increasingly faster rate. While an in-depth socio-cultural analysis of this phenomenon lies beyond the scope of this book, this example supports the idea that words (in this case the acronym *ADHD*) work as labels to help classification and, at least in some cases, they facilitate the identification, labelling and inclusion of new, previously unlabelled instances into the category defined by the word.

Words can construct meaning not only by forcing the inclusion of instances into a category, but also by driving mental simulations and conceptual combinations of existing concepts into new ones. For example, the reader may be unfamiliar with the concept BILES, in English. However, the reader can construct such concept, by mentally combining familiar concepts, as I will now elucidate. BILES is a concept that denotes a relatively recent type of gymnastics move, named after

the first American gymnast – Simone Biles – who performed it in 2013. This jump, done in the context of floor exercise, consists of a double vault up in the air, with a 180-degree turn at the end. That half-turn at the end of the jump means that the gymnast lands blind (she cannot see the ground when she lands), which increases dramatically the skill's difficulty. In this short verbal description, I allowed the reader to construct in her mind the meaning of the concept BILES, using and combining words that, in turn, drove mental simulations that constructed a representation of a previously unknown concept. The reader does not need to physically experience (or perform!) a Biles move, in order to construct the related concept. Such meaning can be constructed by mental simulations triggered by linguistic explanations: words can drive the construction of new conceptual content, even in absence of perceptual experiences.

Finally, a large body of empirical literature on children's cognitive development shows that words stimulate children to identify commonalities between objects (Waxman and Markow, 1995), and perform categorizations (Balaban and Waxman, 1997; Fulkerson and Waxman, 2007; Plunkett, Hu and Cohen, 2008; Robinson and Sloutsky, 2007; Ferry, Hespos and Waxman, 2010). This type of research generated two crucial questions: whether labels enable infants to form categories that they would *not* otherwise form in absence of labels, and whether labels can even *override* non-verbal perceptual information and therefore change the structure of categories previously established on the basis of perceptual similarities. The answer to both questions seems to be *yes*[6] (e.g., Althaus and Westermann, 2016 for a literature review).

2.5 Summary

In this chapter I have introduced some of the core issues related to the development of word meaning and how word meanings are learned by children. I started by providing an outline of the classic top-down, rule-based view of word meaning acquisition, based on principles that would constrain and bias the associations between words and objects in children. Then I explained what type of problems top-down models face, such as the issue of scalability, pointing out that models

6. The power of words in shaping categories already in preverbal infants has been also compared to situations in which infants were presented with non-linguistic sounds, uttered by the parents as a control condition. It has been found that non-words do not have the same facilitatory effect on categorization. Therefore, the invitation to form categories that can even override previously formed categories based on perceptual similarities is a peculiarity of language, and not just of auditory stimuli added to the visual ones (Plunkett, Hu, & Cohen, 2008; Althaus & Westermann, 2016)

based on constraints do not easily explain how such constraints are overcome during development in order to enable children learning new word meanings that clash with the initial constraints (e.g., synonyms and hypernyms). I then described recent models on word meaning acquisition, based on the theoretical paradigm of cross-situational learning, a bottom-up approach that relates to associative and discriminative theories of meaning construction. Such models are central to the implementation of computational models of meaning, and the generation of AI, as I will further illustrate in the next chapters.

In the second part of this chapter I mentioned a methodological problem related to much of the empirical research conducted on word meaning acquisition, namely, the fact that it is almost exclusively focused on the acquisition of words denoting concrete concepts. I explained that cross-situational learning does not explain how the meaning of abstract words is acquired, because for abstract words there is no concrete referent appearing repeatedly across multiple situations. I left this question open and I will return to this in the third part of the book, where I will explain that not only word-to-world associations are key to the construction of word meaning, but also word-to-word associations are equally important. In this chapter I anticipated that the information encoded in language does not simply mirror the information already encoded in perceptual experience (this is demonstrated by means of computational modelling in Chapter 7). I then provided examples showing how language can construct meaning, force classifications, and drive mental simulations of previously unknown concepts. To conclude the chapter, I mentioned recent empirical literature conducted on infants supporting the crucial role that words used as labels play in learning categorizations, showing that they can indeed override the perceptual dissimilarities between objects perceived by infants, and lead them to establish new categories, driven by linguistic labels.

CHAPTER 3

Word meaning extension

Deriving new meanings from old ones

3.1 Word meaning representation and conceptual representation

The literature about knowledge representation, theory of meaning, language processing and comprehension, and semantic memory uses either the notion of word meaning representation or the notion of conceptual representation, without addressing in detail the reasons for such choice. What is the difference, then, between word meaning representation in the mental lexicon, and conceptual representation? Such theoretical and terminological distinction is typically debated in philosophy of language, a discipline and a body of literature that are only marginally touched in this book. From a first oversight, it seems that the different terminology is due mainly to the scientific field from which the topic is investigated: while psychologists and cognitive scientists tend to talk about (and in terms of) concepts and conceptual representations, linguists tend to talk about words and word meaning. Interestingly, however, when talking about conceptual representations, psychologists and cognitive scientists tend to rely on methods for data collection and theoretical paradigms that exploit verbal manifestations of concepts. For example, in the classic featural views of conceptual knowledge (e.g., McRae, Cree, Seidenberg and McNorgan, 2005) conceptual representations are operationalized as bundles of semantic features, which are typically collected in property generation tasks (e.g., given the concept CAR speakers are asked to list the main features of this concept and they typically mention <has 4 wheels> as a core feature). Such features are expressed verbally and therefore are arguably at least in part influenced by the constraints that a linguistic system poses to the expression of conceptual content. Similarly, in rating paradigms, in which numeric judgments are collected from speakers, such as concreteness scores (Brysbaert, Warriner and Kuperman, 2014), modality norms (Lynott and Connell, 2013) and so on, are typically collected by asking speakers to rate on numeric scales sets of *words* in relation to given parameters (e.g., indicating how concrete is the word *banana* on a scale from 1 to 5). The ratings are then taken as indicators about the content of the underlying *concept*.

The terminological choice between conceptual representation and semantic representation of word meaning seems to be partially related not only to the discipline in which the topic is investigated, but also to the school of thought and the theoretical assumptions in which the study is embedded. In particular, supporters of amodal views of cognition tend to see language as a self-contained module in the brain that functions thanks to the manipulation and combination of amodal symbols stored within it (e.g., Fodor, 1983; Pinker, 1995). Such a module is often referred to as mental lexicon, and the representations within it as (lexical) semantic representations.[1] Conversely, supporters of grounded views of cognition (e.g., Pecher and Zwaan, 2005; Barsalou, 2008) argue that cognition and meaning are distributed across brain areas and meaning representation involves the activation and manipulation of symbols that are grounded in the brain modal systems (such as the systems that process perception and action). Thus, meaning, including word meaning, triggers the activation of information that is not simply linguistic, but also extra-linguistic, and the emerging representations are considered to be *conceptual* rather than simply *linguistic*.

In this book, I refer to 'mental lexicon', a notion that is not free from controversy (e.g., Elman, 2009). Nonetheless, when I refer to mental lexicon, I do not implicitly support the idea of a modular architecture of the mind (e.g., Fodor, 1983), and I do not claim that word meanings are stored in a system that is functionally independent, associated with distinct neural structures, computationally autonomous, and possibly genetically determined. Instead, I refer to mental lexicon as a virtual (rather than physical) architecture that collects all the knowledge and information (derived from language and from experience) about a word and allows the different streams of information to interact, combine, and inhibit one another, depending on the context and task conditions in which the speaker is involved. My definition of mental lexicon is therefore not tied to specific configurations of brain activation that would *always* take place whenever a word form is processed; it pertains to a higher theoretical level of semantic representation of word meaning.

Within such architecture, word forms can be associated with various word meanings, each of which consists of different configurations of linguistic and perceptual information combined together, and their modulation within a specific context and in response to a specific goal. For instance, while the meaning of the

1. Note that this terminology is not out-of-date: it is still a well-established terminology used in titles of contemporary scientific journals (e.g., 'The mental lexicon', Benjamin Publishers) as well as in conference names (e.g., the annual International Conference on the Mental Lexicon). Nowadays, however, scholars who adopt this terminology (e.g., Pirrelli, Marzi, Ferro, Cardillo, Baayen and Milin, 2020) seem to agree on rejecting the idea that mental representations of words simply and statically contain lexical knowledge on which grammar operates, as a static corpus of rules.

word form *fork* is prototypically that of a utensil used for eating, in the context of a hike the word takes the meaning of a place where a road, a path, or a river divides into two parts to form a shape like a 'Y'. In the context of software development, *fork* denotes an abstract concept: a new operating system derived from an operating system that is still currently being used (e.g., an *Android fork*). The reader may have been aware of the second meaning of *fork* and may agree that such meaning is derived from the original one thanks to the perceptual resemblance between the pronged shape of the kitchen utensil and the pronged shape of the road or path. This similarity in shape is based on perceptual features shared by the two referents that are arguably learned from perceptual experience. However, it is possible that the reader was not aware of the third meaning of *fork*. Even if this word meaning was unknown, the reader can learn it thanks to the verbal explanation that I provided. The similarity between the utensil *fork* and the software derived from an existing software, also a *fork*, is not based on perceptual features, because the software does not have perceptual features. However, the process of programming a new software that derives and slightly diverges from an existing algorithm can be metaphorically conceptualized as a bifurcation, from which the new meaning of *fork* is acquired. In this case, the new meaning is acquired thanks to information derived from language (i.e., the verbal definition of this word meaning, and exemplification by *Android fork*), rather than direct perceptual experience. In turn, the establishment of this new meaning, achieved through linguistic input, may trigger the mental simulation of the utensil *fork*, and by means of cross-domain comparison may trigger the establishment of cross-domain mappings between the utensil and the new software. In this case, processing the newer meaning of *fork* (i.e., understanding that it can denote a software) is an operation that starts from linguistic information and may then involve embodied simulations (i.e., the activation of previous perceptual experiences related to the basic meaning of *fork*). I specified that the activation of the perceptual information associated to the meaning of the utensil *fork* may take place during the processing of the meaning of the software *fork*, but it does not *necessarily* take place. According to recent literature, in fact, such activation is extremely sensitive to task condition and contextual situations (e.g., Cuccio, 2018). To conclude, the word *fork*, originally denoting a concrete kitchen utensil, is polysemous because it can be used to describe other concrete referents, such as a bifurcated path in the woods, as well as abstract entities such as a software.

Two mechanisms are commonly acknowledged to be often responsible for polysemy and new meaning generations, like those in the example of *fork*: metaphor, and metonymy. The difference between them is non-trivial and hotly debated. In the coming sections I will describe and exemplify prototypical examples of polysemy and the specific cases of word meaning extension by metonymy and by

metaphor for the generation of new word meaning starting from old word forms. I will accompany the description of each of these mechanisms with the description of empirical studies which aim to explain how such meanings are disambiguated by humans, and recent computational models that have been implemented to solve the problems that such mechanisms pose for machine learning and in general for the automatic treatment of natural language. By doing so, I gradually start to outline the idea that parallels can be laid down between the way in which word meaning is structured and processed in the mental lexicon, and the way in which word meaning is modelled computationally. These parallels will be increasingly explored and discussed in the coming chapters, and in the third Part of this book I will finally spell out in detail how the evidence coming from word meaning cognitive processing and word meaning computational modelling converges to inform us on the very nature of word meaning.

3.2 Meaning extension by polysemy

> The fact of polysemy reveals that it is apparently easier for people to take old words and extend them to new meanings than to invent new words … [this] is the preferred route even if it results in very complex word meanings.
>
> (Murphy, 2004, p. 406)

Polysemy is a type of lexical ambiguity, in which a word form can be associated with two or more meanings, which are semantically related to one another. Because they are semantically related and because the form of the word is the same, the meanings are also typically referred to as senses.

In this respect, a word like *chicken* is polysemous because it denotes a type of meat (food) and a type of animal (bird), and the two senses are clearly semantically related. Similarly, the word *keyboard* is polysemous, because it denotes a musical instrument and a computer device. The two senses are semantically related, because they share core semantic features (in this case, the presence of keys that an agent needs to press in order to use these objects). Nonetheless, the two senses of *chicken* are probably perceived to be closer to one another, compared to the two senses of *keyboard*. The level of perceived proximity between two senses of a polysemous word varies, and it is influenced by speakers' knowledge, and notably by their awareness about the word etymology. In fact, distinguishing cases of polysemy from cases of homonymy on the basis of etymology (Lyons, 1977) is extremely tricky. Consider for example the words *bank* and *cardinal.* While for the former word, denoting the side of a river and the place where money is managed, it is fairly easy to see that the two meanings are very distant from one another, not

historically related, and therefore the two forms are homonymous, for *cardinal* the same conclusion would be erroneous. *Cardinal*, as a noun, encodes two senses: a rank within the Roman Catholic Church and the songbird typically found in North America. From a historical perspective these two senses are related: the male cardinals (birds) are mostly red and so this bird was named because of its chromatic resemblance to the red outfit worn by cardinals (humans). Although according to the etymological criterion *cardinal* would be a case of polysemous word, many speakers of English may not be aware of this historical connection. To them the two senses may seem entirely unrelated. The distinction between polysemy and homonymy within the synchronic level of analysis may therefore remain an operation based on speakers' intuitions and the two phenomena may therefore be seen as two extremes on a continuum.

From a psycholinguistic perspective, an interesting question about polysemy and word meaning representation in the mind is whether polysemous words are represented in terms of one meaning or, instead, as two separate entries. Klein and Murphy (2001, 2002) ran a series of experiments designed to investigate precisely whether polysemous words are represented in terms of a common core meaning, from which the various senses may then be modulated by the context, or whether, instead, they are represented as separate entries. In their first experiment, for example, participants were presented with a set of phrases and instructed to memorize them. The phrases included polysemous words in which the activation of one of the two senses was facilitated by the context (e.g., *liberal paper* biased participants toward the newspaper sense of *paper*, while *wrapping paper* biased them toward the sheet sense). In a memory task, participants were asked to recall whether they saw a word or not (reaction times were also measured). The experimenters found that the items that were repeated in the list (i.e., the polysemous words) were the most accurately evaluated, and interpreted these results as to suggest that polysemous senses are stored separately in the lexicon. In a subsequent experiment, Beretta and colleagues (Beretta, Fiorentino and Poeppel, 2005) reported a processing advantage for polysemous words with many senses, compared with polysemous words with only a few senses. This finding relates to an earlier result obtained by Rodd and colleagues (Rodd, Gaskell and Marslen-Wilson, 2002) who tested the so called 'ambiguity advantage' (Borowsky and Masson, 1996) and found a significant advantage for polysemous words in both response time and accuracy compared with monosemous words.

Whether or not the meanings of polysemous words are represented as separate entries in mind, humans starting from a very early age can fairly easily disambiguate which of the senses is intended in any communicative situation, if the context directs the listener toward the intended meaning. However, such disambiguation task poses serious problems to language modelling, and constitutes a major area

of investigation in computational linguistics and machine learning. This critical bottleneck, commonly defined as Word Sense Disambiguation (WSD), was first formulated as a distinct computational task during the early days of machine translation in the 1940s, making it one of the oldest problems in machine learning (Weaver, 1949): how can a machine disambiguate the two senses of a polysemous word, and pick the one that works within a given context? While in the early days WSD was approached by means of manual disambiguation, with humans manually coding polysemous words with the right sense in a given context, when large electronic lexical resources became available in the 1980s hand-coding started to be slowly replaced with knowledge automatically extracted from such resources. However, it was only in the 1990s that WSD became a paradigm problem on which to apply machine learning techniques (see Navigli, 2009 for a detailed review).

In modelling polysemy, recent advances in machine learning and natural language processing have adopted, broadly speaking, two main approaches. On one hand we witnessed the emergence of very successful supervised[2] algorithms (e.g., Raganato, Camacho-Collados and Navigli, 2017) aimed at operationalizing word sense disambiguation as an association task between words in context with their most suitable entry which, in knowledge-based systems is typically extracted from Wordnet (Fellbaum, 1998). On the other hand, recent unsupervised techniques are attracting great attention from the research community. This is because while supervised systems require training data, external resources, and manual effort, unsupervised algorithms do not require any external source of knowledge and use only the structural properties of texts to perform the disambiguation task. To determine the correct sense of a polysemous word in a given context, an unsupervised algorithm typically relies on the assumption that similar word senses occur in similar contexts (this is the core intuition of the distributional hypothesis, widely discussed in this book). Thus, the correct sense of a word in a given context can be inferred by looking at what other possible words can be used in the same contexts to replace the polysemous word. The replacing words will give the analysts an idea of what sense of the polysemous word shall be activated in said context. For example, consider the polysemous word *keyboard* used in the sentence *she played the guitar, he played the keyboard*. In this context, the word *keyboard* appears together with words such as *played* and *guitar*. *Keyboard* here could be replaced with

2. Supervised methods rely on manual labour, such as corpora that are manually tagged for word senses, which are used for training the models. These resources are typically laborious and expensive to create. Unsupervised algorithms, instead, do not require any external knowledge nor repositories of word senses such as Wordnet: they exploit the statistical regularities of word co-occurrences in texts to construct on the fly the most appropriate representation of a word, within a given context.

words denoting other musical instruments, such as *trumpet*, without creating a conceptual clash. Conversely, it cannot be replaced by words denoting other computer devices, such as *printer* or *mouse*. It follows that the correct sense of *keyboard* in this context is that of a musical instrument, rather than a computer device.

Despite the enthusiasm around unsupervised methods for solving WSD problems, their actual implementation is particularly tricky. Such limitations are due to the fact that word meaning in unsupervised methods is typically represented by a vector of contexts in which the word has been observed across corpora (see Chapter 5). In such vector, contextual words belonging to the different senses of a polysemous word are conflated. For example, the words that appear in the same context of *apple* may be *Iphone* and *Steve Jobs* in texts about the brand Apple, of *fruit* and *tree* in texts about fruits. In an unsupervised model both types of contextual words become part of the vector that defines the meaning of *apple* and it is difficult to disentangle the two vectors (for the two senses) within this encompassing vector of *apple*. A related bottleneck of unsupervised methods is the disambiguation of words with opposite meanings, because they typically appear in the same contexts. A dish can be both, *excellent* and *terrible*, and so can be a movie, a behavior, a party, a student or a performance. Thus, the words *excellent* and *terrible* share virtually the same set of contexts in which they can be used, which makes the two words distributionally identical. In an unsupervised model that does not have access to external sources of knowledge or manual intervention the meanings of the two words are hard to distinguish. These problems are still open, and recent approaches to these issues tend to use supervised clustering methods to disambiguate contrastive senses and other types of nuanced uses of word senses, such as sarcasm (e.g., Trask, Michalak and Liu, 2015). Relying on external sources of information, supervised methods easily overcome the problem of knowledge acquisition that hunts down unsupervised methods. For example, when affective information (e.g., positive vs. negative valence) about the meaning of words is taken from an external resource and integrated in a model, it is possible to disambiguate words that otherwise would occur in the same contexts (like *excellent* and *terrible*).

Nonetheless, some recent unsupervised approaches appear very promising in addressing this disambiguation problem. In a very recent study, the unsupervised disambiguation between vectors of words with opposite meanings has been tackled by taking into account a larger portion of context, as opposed to a small context window, to construct the word vectors (Meng, Huang, Wang, Wang, Zhang and Han, 2020). The authors show that while antonyms are very close to one another when only local contexts are taken into account (and thus the specific syntactic patterns of a word occurrence), opposite meanings are more distant from one another when global contexts (e.g., a whole document) are taken into account to construct word vectors. By taking into account a larger context, such as a whole

text, an unsupervised method is capable of distinguishing the vectors of two words with opposite meanings that would have very similar vectors if only their collocational patterns were taken into account.

Unsupervised models typically tackle WSD tasks by renaming them as Word Sense Induction tasks (WSI), precisely because these methods do not rely on human annotations or external knowledge resources: word senses emerge automatically, based on their use, and word senses are therefore induced by the contexts. WSI tasks are typically formulated as follows: given a word (e.g., *keyboard*) and a collection of sentences (e.g., *I love keyboard music, I have an alphanumeric keyboard*, etc.), cluster the sentences in a coherent way, so that each cluster contains sentences where the word is used with the same sense. Notably, in this task the unsupervised algorithm does not need to know which meaning is represented by each cluster, but the sentences within each cluster have to be semantically coherent for one specific sense. Most algorithms for WSI are based on Schütze's early work (1992, 1998) and adopt cluster analyses over word embeddings. In these works, word senses are represented by vectors and they are grouped into coherent clusters by cluster analyses (see Chapter 5 for more details).

The most recent models that tackle WSI provide very interesting results. For example, a recent model based on embeddings and thus implemented by means of neural networks, shows that, given the verb *meet*, the method distinguished between a cluster in which the meeting involved only two persons (*I met my wife*) and a cluster in which the sense of *meet* involved a higher number of participants (*The peasants met the liberators*). This semantic distinction, interestingly, is not typically made in human curated lexical resources such as WordNet (Amrami and Goldberg, 2019).

3.3 Meaning extension by metonymy

Apresjan's classic definition (Apresjan, 1974, p. 18) of polysemy states that "regular polysemy is triggered by metonymy, whereas irregular polysemy is triggered by other metaphorical processes."

Regular polysemy is a specific type of polysemy, in which the two senses of a word are linked by a semantic relation that can be regularly found in language, and has the power to generate new word senses. For example, the relation between *chicken* intended as a type of meat, and *chicken* intended as a type of animal is a case of regular polysemy: we can apply that same semantic relation (i.e., meat/animal) to other cases, such as *lamb, salmon, rabbit, octopus*, and so on.

The literature on regular polysemy provides several lists of cases of regular polysemy (e.g., Pustejovsky, 1995), of which I hereby report some examples, besides the meat/animal one described above:

Container/content

a. *The man broke the bottle.*
b. *The man drank the bottle.*

Producer/product

a. *She works at Ferrari.*
b. *She bought a Ferrari.*

Event/location

a. *Afghanistan is a shame.*
b. *Afghanistan is a country.*

Food/event

a. *The lunch was tastier than usual.*
b. *The lunch was longer than usual.*

Interestingly, as humans we can sometimes combine two senses of a polysemous word within the same sentence and still be able to disentangle the overall meaning of the sentence, such as in: *the lunch was delicious but took forever* (Asher and Pustejovsky, 2013, p. 8). Here, the first clause refers to the 'food' sense of *lunch* while the second clause refers to the 'event' sense of *lunch*. Because of this peculiarity, nouns whose meaning is complex in the way described above are typically classified as instances of *inherent* regular polysemy (Asher, 2011). Inherent polysemy involves senses where there are no substantial reasons for assuming that one of the various senses holds a privileged status in the mind of the speakers: the two senses are so intimately interconnected with each other that they must be viewed as being part of one unitary complex meaning. Conversely, cases in which one of the senses holds a privileged status are typical cases of regular polysemy motivated by metonymy. In these cases of regular polysemy, metonymy determines the semantic shifts that generate new senses derived by metonymic extension from the basic sense.

A specific type of metonymy called logical metonymy can be observed in the combination of an event-subcategorizing verb with an entity-denoting direct object, such as in Example (1).

(1) *The author began the book.*

In this example, the interpretation of the sentence requires the retrieval of an implicit or covert event (i.e., writing). Much research has dealt with this type of metonymy, with the aim of determining whether such metonymic constructions also determine extra processing costs during online sentence comprehension.

The two different types of metonymy (regular and logical) are exemplified in (2), where they are compared to a literal statement. The example is taken from the study performed by McElree and colleagues (McElree, Frisson and Pickering, 2006), described and discussed below.

(2) a. *The gentleman met Dickens.*
b. *The gentleman read Dickens.*
c. *The gentleman began Dickens.*

The sentence in (2a) is literal: Dickens in this example refers to a real person. The sentence in (2b) is a case of standard metonymy based on regular polysemy: Dickens as a book author here stands for the books he wrote. The sentence in (2c) is a case of logical polysemy, in which Dickens stands for the books he wrote, and the action expressed by the verb *began* refers to the sub-event of start reading, which is not expressed, and which is an action that can have *book* (also not expressed) as a typical object.

In psycholinguistics, research on figurative language sees two broad classes of psycholinguistic models proposed to explain how such polysemous words are processed. The first group is usually referred to as *indirect access* models. In these models, the literal sense holds a privileged status over the figurative (in this case metonymic) extension, and therefore it is retrieved prior to the retrieval of the metonymic sense. The indirectness is due to the fact that the metonymic sense would be accessed via the preliminary activation of the literal meaning (Grice, 1975; Searle, 1975). In particular, after the literal sense has proved to be a poor fit with the general context, an alternative meaning is activated to fit in the given context. Conversely, in *direct access* models none of the senses takes priority, but instead, contextual and lexical information determine the intended meaning (e.g., Frisson and Pickering, 1999; Gibbs and Gerrig, 1989; Glucksberg, 2001, 2003).

Specifically on the access of literal and extended meanings in metonymy, Frisson and Pickering (1999) and McElree, Frisson, and Pickering (2006) found evidence that familiar metonymies such as the example in (2b) are processed just as quickly as literal meanings (2a), which suggests that the meaning is accessed directly rather than indirectly, but logical metonymies (2c) require extra cognitive effort, which they measured in an eye-tracking experiment. Nevertheless, Lowder and Gordon (2013) found that familiar metonymies (2b) are processed more slowly than literal meanings (2a), thereby supporting indirect models. After the first study, Frisson and Pickering (Frisson and Pickering, 2007) revisited the processing

of metonymy by investigating the comprehension of novel metonymies in relation to different types of context. In an eye-tracking study, they compared the use of familiar vs. unfamiliar words (in particular, familiar vs. unfamiliar authors' names) used in metonymic constructions, in both, supporting and non-supporting contexts. Their results show that familiar metonymy is processed as fast as literal statements but unfamiliar metonymy is processed more slowly. Moreover, they found that the presence of an appropriate context, in combination with a metonymic rule (based on regular polysemy), can facilitate the processing of an unfamiliar word used in a metonymic construction. For example, in *they read Dickens* the author is a familiar name, and the context constructs a rule-based metonymy 'author-for-book'. Conversely, in *they read Needham* the author is an unfamiliar name, but the context still constructs a rule-based metonymy. In this case, both metonymies, the familiar and the unfamiliar one, are processed alike, thanks to the context, which is based on a regular metonymy. The authors interpret these findings as evidence that speakers are able to process novel senses of a word using context as needed in a "rule-driven fashion" (Frisson and Pickering, 2007, p. 597).

In cognitive linguistics, the types of metonymies that are probably most frequently addressed and used as prototypical examples of metonymy are based on a transfer that is commonly acknowledged to be a *referential* transfer that can be solved at a pragmatic level of analysis, rather than a *lexical* transfer to be solved at a lexical level (e.g., Nunberg, 1995). In this discipline metonymy is defined as a communicative shorthand with strong pragmatic power (Littlemore, 2015) that is used to drive listeners' attention by highlighting relevant aspects of specific concepts, which economically allow for the unique identification of a referent in the given context (Peirsman and Geeraerts, 2006). Examples of this type of metonymy, which can be called circumstantial metonymy (Piñango et al., 2017) to differentiate it from the regular metonymy found in regular polysemy and described above, are found in the examples (3). These examples refer to a typical restaurant environment, in which employees need to exchange brief and effective communications with one another and with the kitchen.

(3) a. *Table 23 asked for the bill.*
b. *The risotto asked for more parmesan*
c. *The cowboy hat ordered the vegan menu*

In these examples, the metonymic uses of the nouns that constitute the subjects of these clauses denote a part or an aspect of the referent that they would normally denote in a literal context, namely, the customer: *table 23* stands for the customer sitting at said table; *the risotto* stands for the customer who ordered it, and *the cowboy hat* stands for the customer that wears it. *Table, risotto* and *cowboy hat* have all been selected by the speaker to indicate to a colleague a specific customer

in the restaurant environment. The difference between referential metonymy and lexical (logical) metonymy is that in the former pragmatic inferencing is key for the interpretation of the intended meaning, while for the latter the interpretation works more arguably on the basis of lexical disambiguation processes.

A decade ago, Gibbs (2007, p. 23) noticed a relative lack of attention to metonymy paid by experimental linguists compared to metaphor. Although in the last decade the attention to metonymy and its processing increased, empirical studies on the processing of metonymy from a developmental perspective are still quite limited. The ability of children to comprehend and produce referential metonymies, in particular, remains largely under-investigated. A notable exception is the study conducted by Falkum, Recanses and Clark (2017), who tested 3, 4 and 5-year-old children in their ability to comprehend and produce referential metonymies based on part-whole relations. Interestingly, the authors found that children as young as three years old were able to understand metonymies when the context made the association transparent, but that, contrary to expectation, they performed less well as they got older, with 4 and 5-year-old finding literal items significantly easier to understand than metonymic ones, to a larger extent than 3-year-old children.

From a computational perspective metonymy, as a form of lexical or referential ambiguity, is particularly challenging. Ideally, a system capable of interpreting the correct metonymic sense of a word in a given context would be an invaluable addition to the real-world natural language processing applications such as, for instance, machine translation. As one would expect to see, computational models aimed at disentangling metonymic from literal senses of a word in context typically address lexical metonymies, rather than referential metonymies, and in particular they aim at modelling logical metonymy by identifying the missing element (recoverable or inferable from the linguistic context) within the elliptical construction. For example, given the sentence *he started Dickens*, a computational model that aims at disentangling the correct sense of the word *Dickens* here would need to first identify the covert event *reading*, to then understand that *Dickens* refers to the typical objects of such event (i.e., books).

One of the first attempts (both theoretical and computational) to model and explain how the covert element of a logical metonymy can be retrieved dates back to the works of Pustejovsky (1995) and Jackendoff (1997), who assume that the covert event is retrieved from complex lexical entries consisting of rich knowledge structures, called qualia roles. According to this theoretical model, the representation of the noun *book*, and therefore its qualia structure, would include telic properties (such as purpose of the entity, e.g., to be *read*), agentive properties (such as the action that enables its existence, e.g., *write*) and so forth. In a metonymic construction such as *start the book*, the mismatch between the predicate and the argument (e.g., between *start* and *book*) triggers the retrieval of a covert event (i.e.,

read) from the qualia roles of the object (*book*), thereby producing a semantic representation equivalent to *begin to read the book*. Recently, however, Zarcone and colleagues (Zarcone, Padó and Lenci, 2014) have shown that qualia roles are not flexible enough to account for the wide variety of interpretations that can be retrieved, because they do not take into account the discourse context, nor the listener's world knowledge.

In a recent attempt to modelling computationally logical metonymy, Shutova and colleagues (Shutova, Kaplan, Teufel and Korhonen, 2013) implemented a system that, given a polysemous word in context, first derives a set of possible metonymic interpretations from a large corpus, and subsequently disambiguates the various word senses using an existing sense inventory. Finally, the system automatically organizes the word senses into a new class-based conceptual model of logical metonymy inspired by linguistic theory. In this way, the model basically provides for each word used metonymically the likelihood of each possible interpretation, based on corpus data. The authors offer an evaluation of their model by comparing the word-sense organization produced by the model with judgments elicited from human participants who were asked to classify metonymic interpretations into groups of similar concepts. Results show that the performance of the model is comparable to that obtained from human participants.

Finally, the most recent computational attempts to solve the problems posed by logical metonymy in natural language processing adopt distributional modelling techniques, which will be extensively discussed in Part 2 of this book. First, Zarcone and colleagues (Zarcone, Lenci, Padó and Utt, 2013) showed that a distributional model of verb-object thematic fit can reproduce the differences in reading times related to metonymy found by McElree and colleagues (2001) and Traxler and colleagues (Traxler, Morris and Seely, 2002). Subsequently, building upon the previous study, Chersoni, Lenci and Blache (2017) proposed a distributional model that simulates the processing costs involved in logical metonymy in terms of costs involved in the necessary disentanglement of the covert event. In this model, the authors simulate the incremental process that leads to the construction of the semantic representation of events such as *start a book*, by unifying distributional information collected for similar events.

3.4 Meaning extension by metaphor

Another mechanism of meaning generation, through which we can derive new meanings from existing ones, is metaphor. Because the literature on metaphor structure, processing, and modelling is extremely vast, for the purpose of this section I will constrain the discussion around metaphor to those aspects that are

functional for understanding: (1) how metaphor relates to word meaning extension; (2) what type of similarity characterizes the figurative and literal meanings of a polysemous word; (3) how are metaphorical words comprehended by humans and finally (4) what sort of problems are involved in the computational modelling of metaphor comprehension.

Metaphors expressed as *x is y* statements, as in *my lawyer is a shark* are not very frequent in language use (Steen, Dorst, Herrmann, Kaal, Krennmayr and Pasma, 2010). Yet, these of metaphors very frequently appear in a specific genre of texts: in scientific articles about metaphor. As a matter of fact, a very large portion of studies about metaphor address their specific *x is y* form, also called direct metaphor. Most metaphors, however are expressed indirectly, through polysemous words that have a literal and a (derived) figurative meaning, which gets activated by the context. While the literal meaning is usually more basic, more concrete, or etymologically older, the figurative extension is typically more abstract, newer or more complex (Steen et al., 2010). For example, in (4) the verb *devoured* is used metaphorically, because this verb has a more basic and concrete meaning which denotes the action of eating quickly or with extreme hunger. The contextual meaning of *devour* in (4)activates a scenario of an action performed by a human subject to a book, which contrasts with the basic meaning of *devour*, and therefore the verb is considered to be used metaphorically, within that context. A dictionary may list both meanings for the word *devour*: eating quickly and reading eagerly. This means that the metaphorical meaning is highly conventionalized and it is lexicalized as a dictionary entry.

(4) *I devoured the last book by David Eagleman*

(5) *I sipped the last book by David Eagleman*

Conversely, the example in (5) may express in a creative way the action of reading a book slowly and in a relaxed manner. Here, the verb *sip* is used metaphorically, because its basic meaning defines the action of drinking something by taking small mouthfuls. Arguably, the metaphorical sense of *sipping* is not lexicalized in dictionaries, and thus the metaphor is more creative than the previous one.

Words that conflate a metaphorical (contextual) and a literal (basic) meaning may be considered cases of lexical ambiguity, and therefore the question arises (as for metonymy) whether their figurative meaning, within a given context, is accessed directly, selected by the context, or indirectly, via the preliminary activation of the literal meaning.

The theoretical debate around the direct or indirect access to the metaphorical meaning, and therefore around the status of the literal meaning in relation to the metaphorical one, is hotly debated. Several studies have suggested in the past 40

years that when metaphors are very conventional, they are processed as polysemous words (e.g, Gibbs, 1984; Glucksberg and Keysar, 1990; Giora, 1997; Glucksberg, 2001; Bowdle and Gentner, 2005). The metaphorical meaning is therefore accessed via semantic categorization (i.e., it is categorized as an element of a higher-level category selected by the literal meaning) or simple lexical disambiguation, much like for other polysemous words. Instead, for words used metaphorically in a creative way, the comprehension may work by means of an active comparison in which the basic meaning needs to be activated in order to be compared to the contextual meaning, and semantic features need to be projected from the basic to the contextual meaning in order to interpret the metaphor (e.g., Bowdle and Gentner, 2005).

In pragmatics, metaphors can be seen as intended violations of the maxim of quality proposed by Grice, according to which speakers typically try to make their contribution to conversation one that is true, in order to cooperate with their interlocutors. Because metaphor would violate the listener's expectation, it is commonly assumed that Gricean views on metaphor would support the indirect access view. Some very early experimental evidence supporting the indirect access view comes from Clark and Lucy (1975). However, shortly after the emergence of the indirect access theory, Gibbs (1984), Gildea and Glucksberg (1983), Harris (1976)[3] and others set out the case against it and proposed a direct access view that suggests that metaphorical and literal meanings are processed alike and hold the same status. As Harris claims:

> [metaphor] is not a highly-specialized form of language that becomes comprehensible only after the use of inferential processes operating on some literal or more basic meaning. (Harris, 1976, p. 314)

Empirical works such as those reported by Inhoff and colleagues (Inhoff, Lima and Carroll, 1984), Ortony and colleagues (Ortony, Schallert, Reynolds and Antos, 1978) and McElree and Nordlie (1999) provided support for the direct access model by showing that there is no difference between processing figurative and literal language. Eventually "a consensus in the field that literal meaning does not have unconditional priority" seemed to emerge (Glucksberg, 2003, p. 92).

However, with the introduction of new experimental techniques such as EEG,[4] new contrasting empirical evidence emerged, showing that at a neural level there

3. Richard Jackson Harris, not to be confused with Zellig Harris, pioneer of the distributional hypothesis.

4. Electroencephalography (EEG) is an electrophysiological monitoring method that allows, by means of a non-invasive technique, to record the electrical activity of the brain by attaching small sensors to the scalp to pick up the electrical signals produced when brain cells send messages to each other.

is in fact a difference between processing literal and metaphorical meaning (Pynte, Besson, Robichon and Poli, 1996; Lai, Curran and Menn, 2009; De Grauwe, Swain, Holcomb, Ditman and Kuperberg, 2010; Bambini, Bertini, Schaeken, Stella and Di Russo, 2016). Within this complex picture some scholars propose accounts that emphasize the importance of the pragmatic context in which the metaphor (or the literal statement) are encountered (Bambini et al., 2016). Others interpret the results of their studies as supporting the primacy of the literal meaning and of the original indirect access view (e.g., Bonnaud, Gil and Ingrand, 2002).

One of the main problems, however, is that empirical studies of this sort are typically based on limited sets of ad-hoc created metaphors, which differ from study to study. This limitation makes the empirical findings on metaphor processing hard to compare across studies, as recently pointed out by Werkmann Horvat, Bolognesi, Kohl and Lahiri (accepted). Moreover, many studies use relatively small samples of direct metaphors (e.g., Glucksberg, Gildea and Bookin, 1982; Gildea and Glucksberg, 1983; Blasko and Connine, 1993; Pynte, Besson, Robichon and Poli, 1996; McElree and Nordlie, 1999; Coulson and van Petten, 2002; De Grauwe, Swain, Holcomb, Ditman and Kuperberg, 2010; Weiland, Bambini and Schumacher, 2014, etc.), or indirect ones (Lai, Curran and Menn, 2009; Lai and Curran, 2013).

From a computational perspective, a metaphor processing system capable of identifying polysemous words, distinguishing between contextual and basic meanings, and selecting the correct word sense in order to understand the meaning of a sentence, is already an extremely hard task (Veale, Shutova and Klebanov, 2016). Even more so, would be the implementation of a system that possesses an embodied understanding of the world and can interpret metaphorical statements by activating deep conceptual representations based on embodied experiences, as the Conceptual Metaphor Theory fathered by Lakoff and Johnson (1980) proposes in relation to metaphor processing in the human mind. Moreover, as Veale and colleagues point out, once we leave the lexical level of metaphor representation to enter the conceptual level, it becomes clear that metaphor conceptualization can be tackled at many different levels of abstraction. A recent theoretical account inspired by linguistic theory addresses such variability among levels of metaphor conceptualization (Kövecses, 2017). In this contribution the author, Kövecses, explains that metaphor affords many levels of conceptual representations, which vary in terms of their semantic richness, ranging from the very schematic *image schemas*, to the less rich *domains*, followed by *frames* and *mental spaces*, which constitute the richest representations. For example, the conceptual domain BUILDING is characterized by several image schematic representations, like the image schemas commonly labelled in the literature as CONTAINER, VERTICALITY, and STRUCTURED OBJECT. As a conceptual domain, BUILDING is used in

conventional conceptual metaphors such as THEORIES ARE BUILDINGS, which emerge from the systematic occurrences of linguistic expressions in which buildings are used to talk about theories and arguments (e.g., "this theory has solid foundations"). At a more fine-grained level of conceptual richness, BUILDING consists of a number of frames: it possesses a CONSTRUCTION frame, a STRUCTURAL ELEMENTS frame and a CONSTITUENT PARTS frame, which encompasses concepts such as walls, rooms, doors, windows, a FUNCTION frame that provides information about who use the building, and so forth. The richest level of conceptual representation, the level of mental spaces, or scenarios, emerges when we use language in real communicative situations and thus, we contextualize, elaborate, and modify frames. At this level, for example, specific types of buildings can be used to talk about attitudes and behaviors toward specific topics, as in the corpus example reported by Kövecses: "public employee unions, in league with compliant state officials, have built a fortress around their pension systems" (2017, p. 338).

Such variability among levels of analysis and theoretical models of metaphor processing at a conceptual level is extremely difficult to tackle in a comprehensive way by means of computational modelling techniques. Veale and colleagues (2017, pp. 33–51) have recently provided a very detailed and exhaustive historical review of the metaphor processing systems implemented in the past 40 years. These models range from computational approaches that aim at understanding the metaphor by detecting its conflictual nature within a given context and correcting it (here metaphor is seen as a divergent semantic entity within a coherent semantic organization) to approaches that comprehend the metaphor on the basis of analogical structures (Gentner, 1983), in which metaphors are seen as cohesive series of systematic mappings across domains, including also approaches based on the identification of a schematic structure from which the actual metaphor can be derived.

In language, and in relation to metaphorical word meanings, the automatic processing of metaphor involves two subtasks: metaphor identification (detecting a word used metaphorically in a given context) and metaphor analysis (identifying the intended meaning of a metaphorical expression). The problem of automatic metaphor identification is very challenging because manually annotated sets of metaphors in language that could be used to train algorithms to learn to identify metaphorical words are very limited. Among these, the most prominent dataset is probably the VU Amsterdam Metaphor Corpus (Krennmayr and Steen, 2017), which covers about 190,000 lexical units from a subset of four broad registers from the BNC-Baby: academic texts, conversation, fiction, and news texts. All lexical units have been manually annotated for metaphoricity. The annotation procedure is based on the Metaphor Identification Procedure (MIP), presented by the Pragglejaz-Group (2007). This procedure introduces a systematic approach based on

dictionaries, with clear decision rules: a word is considered to be metaphorical if it is not used in its most basic meaning (according to dictionaries), and if its contextual meaning can be understood in comparison with the most basic one.

The most popular model for the automated identification and analysis of metaphor is probably the model provided by Shutova, Kiela and Maillard (2016), Shutova, Teufel and Korhonen (2013) on the basis of previous work developed by Shutova, Sun, and Korhonen (2010) and Shutova (2010), who designed a statistical model that captures regular patterns of metaphoricity in a large corpus and is capable of generalizing to unseen examples. This model is based on the identification of selectional violations in texts. Selectional violations are exemplified in seminal works by Wilks (1975, 1978), who explains that the nature of metaphors is to violate selectional preferences of lexical units in context. For example, in the statement *my car drinks gasoline*, the metaphorical use of *drink* violates the literal selectional preferences of this verb (chiefly, the need for an animated agent who performs the action of drinking). An algorithm that aims at comprehending this metaphorical use of the verb *drinks* shall identify this selectional incongruity. The violation can be identified by means of an external source of knowledge that contains information about, for example, the frames in which each word is typically used. A frame is defined as a "structure of expectations" (Tannen 1993, p. 21), or a scenario in which, based on previous knowledge, we expect to find some roles (agents, actions, locations etc) but not others. If a speaker brings frame A into a discourse governed by frame B, then probably words belonging to frame A are used metaphorically in frame B. In the example above, thanks to external knowledge about frame semantics the algorithm shall identify the verb *drinks* as alien to the frame of cars, engines, and gasoline, and eventually replace it with the more literal *consumes*, which is selected from an external resource for its semantics that is similar to the semantics of drinking. In particular, in this case the verb *consumes* is selected from a lexicon, by climbing the taxonomic relations of drinks until a hypernymic verb is found that does not require an animate agent. Alternatively, selectional preferences can be learned in an unsupervised manner using corpus-based approaches, as in some relatively recent studies (Krishnakumaran and Zhu, 2007; Shutova et al., 2010; Huang, 2014). Krishnakumaran and Zhu, for example, acquired selectional preferences from bigram frequencies on the Web. Shutova and colleagues (2010), and Huang (2014) instead focus on the strength of selectional preferences of verbs, assuming that verbs with weak selectional preferences are not likely to be metaphorical. Although these models proved to be quite successful in the detection of metaphor-related words in context, they are limited to the detection of conventionalized metaphors, since frequently used metaphorical word pairs are the only one found in text co-occurrences.

In 2018 a Metaphor Detection Shared Task was launched by Leong, Klebanov and Shutova (2018) within the NAACL Workshop on Figurative Language Processing. The task was formulated to compare models that could automatically identify metaphors across the 4 genres included in the VU Metaphor Corpus, which was used to evaluate the performances of the computational models. Two sub-tasks were formulated, to invite models that could detect metaphors in verbs and models that could detect metaphors across all parts of speech, within the 4 genres of texts included in the BNC dataset. (Leong, Klebanov and Shutova, 2018). The task coordinators reported that all (except one) participants used vectorized word representations based on word embeddings to construct word meanings, rather than using the explicit features provided by the coordinators, which were taken from prior published work on metaphor detection. These explicit features included "unigram features, features based on WordNet, VerbNet, and those derived from a distributional semantic model, POS-based, concreteness and difference in concreteness, as well as topic models" (Leong, Klebanov and Shutova, 2018, p. 59). Among the main results, the task coordinators report that the automatic identification of metaphors in verbs is easier than for other parts of speech, and the genre of academic texts is the easiest among the 4 genres (followed by news, fiction and at last conversation). Overall, three systems (all based on word vectors constructed with word embeddings and thus neural networks) outperformed the stronger of the two baselines provided by the coordinators, which was based on Beigman Klebanov et al. (2016).

Among the computational models of metaphor comprehension that tackle one of the conceptual levels of metaphor outlined by Kövecses, the model proposed by Utsumi (2011) is probably the most popular. This model is partly based on the predicative approach to metaphor analysis proposed a decade earlier by Kintsch (2000). Both these models are corrective, in the sense described above, and use methods based on the distributional hypothesis and represent word meaning by means of vectors of coordinates. Each coordinate indicates the strength of association between said word and a possible context of occurrence, extracted from corpora. In these approaches word meaning is modelled in terms of word distribution across contexts of use, and therefore the similarity between two words (including the similarity between two metaphor terms) is modelled in terms of the linguistic contexts that the two words share.

The theoretical question that remains open is to what extent can these models, which are based on the analysis of word occurrences across large databases of texts, account for the actual human cognitive mechanisms involved in metaphor processing. In the end, machines do not have bodies through which they can perceive the world and store conceptual representations of word meaning constructed from perceptual experiences: they can only gather information encoded in

language. For metaphor, this seems to be particularly problematic, given the large body of scientific literature suggesting that metaphor comprehension is based on embodied cognitive processes. As I will describe in greater detail in Chapter 7 of this book, however, the distributional hypothesis may constitute a core cognitive mechanism through which we extract patterns not only from language, but also from experiences.

3.5 Summary

Metonymy and metaphor are the most productive cognitive strategies for the construction of new meanings based on old word forms. Their cognitive processing is still debated, with empirical evidence showing in some cases that the literal and figurative (metonymic or metaphoric) meanings are processed alike, and in other cases that they are processed differently. The main distinction between theoretical approaches in the study of the cognitive processing of metaphor and metonymy is between supporters of direct and indirect routes to meaning who argue, based on empirical evidence, that the figurative meaning is accessed directly, or alternatively through the preliminary activation of the literal meaning. Recent findings seem to suggest that both direct and indirect access may take place, depending on the familiarity/conventionality of the derived meaning and the context in which the word is presented.

The computational modelling of polysemous words in which new meanings are derived from old ones by means of metonymic or metaphoric extensions is particularly demanding because these cognitive mechanisms operate across multiple semantic dimensions, which cross the boundaries between linguistic expressions and conceptual structures, with the latter dimension being fragmented into multiple layers where the representation of meaning varies in terms of semantic richness (Kövecses, 2017).

Despite the difficulties involved in cracking the cognitive processing of metaphor and metonymy in its variation related to the type of metaphor/metonymy, the context in which it is used, and the level of abstraction at which it is formulated, there have been notable achievements in the way these mechanisms are modelled computationally. The most recent achievements reviewed in this chapter show that there seems to be a convergence toward the methods used to model metaphor and metonymy, which appear to be based on the distributional hypothesis. In Part 2 of this book I will elucidate further how such distributional models are implemented by means of classic frequency-based approaches or word embeddings, and in Part 3 I will finally elaborate in detail how the underlying hypothesis (the distributional hypothesis) provides a flexible and cognitively plausible theory of meaning

that can explain, among other things, how we may move from the processing of metaphor and metonymy via the activation of perceptual experiences by exploiting word-to-world associations, to the processing of metaphor and metonymy via the activation of simple linguistic information by exploiting word-to-word associations. These two processes, as I will argue, are both based on the distributional hypothesis.

CHAPTER 4

The bilingual mind and the bilingual mental lexicon

4.1 Theoretical models of the bilingual mental lexicon

The Revised Hierarchical Model proposed by Kroll and Stewart (1994) to describe the organization of the bilingual mental lexicon is today a cornerstone among scholars working on second language acquisition. This model, displayed in Figure 5 makes a hierarchical distinction between a lexical and a conceptual level of semantic representation. The authors suggest that in early stages of a second/foreign language acquisition words in the L2[1] are connected to the conceptual representations via their L1 translations. This type of connection is called *lexical mediation*: in order to understand a word in a second/foreign language, and therefore access its conceptual representation, a non-fluent language learner will first mentally translate the word in the first language, and then access the related conceptual representation.

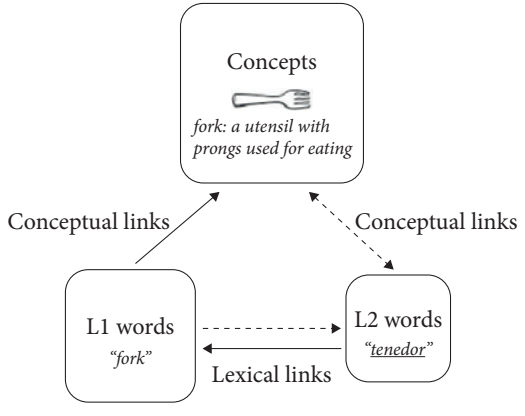


Figure 5. The Revised Hierarchical Model proposed by Kroll and Stewart (1994), displaying the lexical mediation (lexical links between word forms) and the conceptual mediation (conceptual links between word forms and concepts). The continuous line indicates stronger links, whereas the dotted line indicates weaker links

1. L2 is the target language in foreign language learners, and the second language in bilinguals, while L1 is the native language.

When the proficiency in the L2 increases, the links between L2 words and concepts become stronger, and word meanings in the L2 become directly connected to the conceptual representation, without needing lexical mediation.

As described by Brysbaert and Duyck (2010) this model has been largely influential in language acquisition research, because it offered a useful theoretical framework that allowed researchers to explain various behavioral phenomena observed in language learners. Thanks to the distinction between a lexical and a conceptual level, Kroll and Stewart were also able to explain some contrasting empirical findings. For example, in one pioneering study, Glanzer and Duarte (1971) had observed that participants found it easier to memorize words in lists when these words were repeated twice, compared to when they appeared only once in the list. This was also true when the words appeared twice but in different languages: if in the same list there were both *house* and *casa*, English/Spanish bilinguals would remember these words better than words that appeared only once. This study suggested that word representations are shared between the two languages, and stored within one common mental lexicon, and the processing of house and casa involves the activation of the same semantic representation twice, which facilitates its recall during the task. However, a related study had shown that participants who were asked to memorize lists of words, found this task harder if they had to learn lists of words belonging to the same category, compared to lists of words belonging to different categories. For instance, a list containing *apple, grape* and *pineapple* (all fruit) was harder to memorize if they previously memorized a list such as *orange, plum, pear* (also all fruit) than if they had to memorize *arm, head, nose* (body parts). This interference effect was significantly reduced if the lists containing words within the same semantic field were presented in two different languages. This result was taken as evidence to support the idea of two separate storages for words in the two languages (Goggin and Wickens, 1971). This observation suggested a theoretical model in which the semantic representations in the L1 and the L2 were separated, while Glanzer an Duarte suggested a unified mental lexicon for word meanings in the L1 and the L2. Kroll and Stewart reconciled the contrasting findings by arguing that the two tasks tapped into two different types of representations: conceptual representations in the first case (which are shared between languages) and lexical representations of the word forms in the second study (which are language-specific). The separate lexica in the L1 and the L2 proposed by Kroll and Stewart were also used to explain why bilinguals do not code-switch between the two languages continuously and without any control (see Green, 1998; Costa and Santesteban, 2004 for further discussion on this issue).

The model proposed by Kroll and Stewart offered possible explanations to early behavioural evidence reported in second language acquisition research, as well as early neuroscientific evidence. In particular, the emergence and diffusion

of new neuroimaging technologies enabled researchers to investigate the patterns of brain activation during the performance of linguistic tasks in bilinguals. These patterns showed a great overlap between L1 and L2 active brain areas, but revealed also some activation that seemed to be peculiar of the L1 or L2 the only (see Indefrey, 2006, for a meta-analysis). Moreover, some brain areas seemed to be more active in bilinguals, compared to monolinguals. Abutalebi and Green (2007) showed that a large network of neural structures which are known to be related to various functions that fall under the broad category of cognitive control (e.g., attention, decision making, response selection and inhibition, and working memory) tend to be active during language production in bilinguals, especially during the performance of tasks such as translation and switching deliberately from one language to another. In other words, the ability of bilinguals to avoid continuous and uncontrolled code-switching seems to be related to the activation of this frontal control network that enables bilinguals to suppress momentarily the unwanted language. The authors related these findings to previous findings in which it was argued that in bilingual speakers both languages are active all the time, even when only one is in use (see, for reviews, Kroll, Bobb and Wodniecka, 2006; Bialystok, 2011). The precise locus of suppression in bilingual speakers depends on the source of the possible interference with the unwanted language. If the unwanted language may cause interference at a lexical level (e.g., presence of cognate words such as *lamp* in English and *lampara* in Spanish), the suppression takes place in an area of the brain that is different from the area in which the suppression takes place if the unwanted language interferes at a syntactic level (e.g., similar but not equal ways to construct a sentence).

The fact that the control network seems to be playing a complex role in managing the activations and inhibitions of the two languages, which would both be active all the time, does not fit well with a clear-cut separation between the two mental lexica proposed by Kroll and Stewart: a much more dynamic architecture needs to be modelled, on which the control network can operate under different circumstances.

Another debated issue on the structure and functioning of the mental lexicon of bilingual speakers and foreign language learners is when, how, and to what extent are word *meanings* (as opposed to word *forms*) in the L1 and the L2 activated during the performance of various types of tasks (see Plat, Lowie and Bot, 2018 for a review). For example, in simple lexical decision tasks in which speakers are asked to determine whether a word is genuine or invented, word meaning does not need to be fully activated, because the recognition of the word form is sufficient to perform the task. A model of word representation that distinguishes between a lexical level and a conceptual level is not sufficiently granular to be able to successfully explain how various types of semantic information and word representation

may become active in different tasks. Spivey also contested the idea of discrete, static representations, such as those postulated in Kroll and Stewart's model, and stresses the fluid nature of cognitive processing, claiming that mental (whether lexical, phonetic, semantic, or otherwise) representations should be thought of as processes, or as "sparsely distributed patterns of neural activation that change non-linearly over the course of several hundred milliseconds, and then blend right into the next one" (Spivey, 2006, p. 139).

As observed in Chapter 2 in relation to the acquisition of words by children but true also for adult language learners and bilinguals, the top-down approaches proposed in past decades and based on static models governed by rules and constraints, have been slowly replaced by models of word meaning and semantic representation that emphasize the importance of bottom-up approaches that make room for dynamicity and non-linearity. This trend is well explained by de Bot and colleagues (de Bot, Lowie, Thorne and Verspoor, 2013), who observe that Complexity Theories and Dynamic Systems Theory, originating in the physical sciences and mathematics, are becoming increasingly influential among applied linguists working on language acquisition. Such theories enable researchers to replace static models with chaos, which is a necessary initial stage that enables emergent (dynamic) structures to arise naturally.

Despite the fact that in recent years cognitive scientists and neuroscientists are suggesting more dynamic and flexible models to represent the relation between semantic representations in the bilingual mind and in the bilingual brain, some basic agreement seems to remain among scholars, on the fact that word representations in the L1 and the L2 (at least at the first stages of second/foreign language acquisition) are substantially different. Lemhöfer and colleagues (2008), for example, conducted a mega-study on bilinguals with different native languages (French, German, and Dutch) to look into how specifically the L1 influences L2 processing. In their study the authors used a multiple regression model that allowed them to take into account many variables related to the words used as stimuli such as frequency, length, concreteness, meaningfulness and many others. The authors combined all the variables into one model and found that only one factor could systematically predict how the L1 influences L2 processing. This factor was 'cognate words', that is, when two words in different languages share similar forms and have equivalent meanings (e.g., English *lamp* and Spanish *lampara*). However, the authors also found differences between L1 and L2 speakers on word variables related to frequency and ways of occurrence, which led them to conclude that L2 word processing is fundamentally different from L1 word processing (Lemhöfer et al., 2008, p. 27). The authors interpreted their findings suggesting that L2 processing is more *language driven*: within-language factors seem to be highly influential for the determination of the processing strategies in the L2.

4.2 Word associations in native speakers and language learners

Mining free word associations produced by native speakers and language learners in response to a given word prime is a commonly used paradigm in language acquisition research. A well-established phenomenon observed in bilinguals is that while L1 speakers tend to produce quite consistently word associations that are semantically related to a given prime, L2 speakers tend to produce a more varied range of associations, which tend to be only tenuously determined by semantic similarity with the prime. According to Meara (2009) word associations in the L2 seem to be motivated by: (1) existing associations in the first language of the speakers, which are transferred to the L2 equivalent words; (2) Connections based on episodes and collocational patterns found in the L2, and (3) Phonological similarities.

In particular, Meara describes the types of word associations performed by native speakers in the following way:

> Normal adults tend to produce more paradigmatic responses than syntagmatic ones, provided the stimulus words are reasonably common. Less frequent words, which tend to occur in more constrained contexts, are more likely to produce syntagmatic responses. Children under seven years of age have a strong tendency to produce syntagmatic responses as a first preference to any word.
>
> (Meara, 2009, p. 6)

The learners' associations in the L2 (at least in the study performed by Meara on English native speakers, learners of French), are instead summarized as follows:

> In the learners' case, however, this semantic organization seems to be much less established. The learners studied here do show some evidence of semantic organization, but this is mainly dependent on translations between French and English. There also appear to be a conflicting principle of organization, which makes use of the forms of words rather than their meaning. (Meara, 2009, p. 17)

The fact that language learners' word associations are driven by translations, phonological similarities and collocational patterns suggests that linguistic factors (i.e., lexical translations, word phonetic structure, and syntactic patterns) are more prominent in the representations of words in the L2 than in the L1. This is in line with the findings reported by Lemhöfer and colleagues, who suggest that L2 processing is mainly driven by linguistic factors. As a matter of fact, language learners, especially those who acquire a second language in institutional setting, possibly in their home country and therefore without being exposed to natural communication and experiential contexts, derive information mostly from language, and language-related activities such as reading and listening. Conversely, native speakers are exposed to a variety of perceptual experiences, together with a much greater

corpus of communicative experiences and linguistic input. By relying heavily on the information they can retrieve from texts and from language, language learners seem indeed to structure the associations between words in the L2 on the basis of linguistic factors, and on the basis of linguistic similarities between words, rather than on the basis of perceptual similarities between the denoted concepts, or the designated referents in the world. Language learners rely more on language than on experience to construct word representations in the L2.

4.3 Incidental vocabulary leaning

While reading, language learners typically encounter unfamiliar words. If the amount of unknown words is limited this does not necessarily prevent the learners from comprehending the text because they typically infer the meaning of the unknown words from the linguistic contexts, that is, the sentence in which the word occurs. Of course, when readers are not acquainted with a large number of words their reading comprehension may be impaired. But when the amount of unknown words is limited language learners typically engage with the strategy of guessing the meaning from the context (e.g., Harley and Hart, 2000; Nassaji, 2003). Other possible strategies applied when an unknown word is encountered while reading are: (1) Ignoring the unknown word, and (2) Consulting a dictionary. However, research shows that contextual guessing is preferred (Çetinavcı, 2014). This phenomenon reinforces the idea that L2 vocabulary is learned through reading and by guessing the meaning of unknown words by exploiting the linguistic context and the linguistic information provided by the text. This phenomenon is also referred to as *incidental vocabulary learning* (Huckin and Coady, 1999; Webb, 2008).

Guessing word meaning from context is a type of inferencing activity that "entails guessing the meaning of target word based on interpretation of its immediate co-text with or without reference to knowledge of the world" (Haastrup, 1989, cited in Parel, 2004, p. 848). Language learners are arguably better trained at this than native speakers, because they encounter more often unknown words in texts while reading. Therefore, incidental learning is arguably a type of learning with which L2 speakers are more acquainted than native speakers. This may, at least partially, explain why associations based on lexical factors and linguistic information are more salient in the mind of language learners compared to native speakers: language learners are more sensitive to the information provided by the words that co-occur with unknown words in texts.

In extensive literature reviews Huckin and Coady (1999) and Gass (1999) surveyed various empirical studies focused on investigating the mechanisms involved in incidental vocabulary acquisition. The authors, which focused their literature

reviews on slightly different studies, investigated various variables including the type and size of vocabulary needed for a correct guessing, the amount of exposures for a successful retention, the effectiveness of word-guessing strategies, the influence of different reading texts, and the problems involved with incidental learning.

In summary, according to these authors, the factors that enable incidental learning can be summarized as follows:

- Extensive reading (that is, being exposed to a very large corpus of texts).
- Unknown words must occur in texts several times, and be surrounded by different possible contexts.
- Each context in which an unknown word occurs must be encountered more than once. There is no agreement on the exact number of exposures among researchers. Some studies locate this number between 5 and 16 exposures, but much depends on other factors, such as word salience, its recognizability as a cognate, the learners' interests, and the availability of rich informative contexts.
- The amount of unknown words within a text must be around 3 to 5%, to ensure the full comprehension of the text.

Gass indicated that an additional, important aspect involved in incidental learning is the attention to syntactic constructions in which the unknown words occur which stimulates the reader to infer various aspects about the unknown word, such as its part of speech, and subsequently constrain the array of possible candidates for its meaning.

Incidental vocabulary learning is typically opposed to another learning strategy: deliberate vocabulary learning (Nation, 2001; Thornbury, 2013). The phenomenon of incidental vocabulary learning is quite interesting because it opposes the classic deliberate top-down approach to vocabulary learning, with an unstructured bottom-up approach which, as the quick summary above and the more extensive literature review below suggest, is a quite successful and natural way for learners to acquire new word meanings during the common activity of reading. I will therefore review a few more empirical studies in this field, which provide additional evidence for the effectiveness of incidental vocabulary learning during reading, before summarizing the relevance that these findings have for a bottom-up, data-driven account of word meaning representation.

Ponniah (2011) investigated the impact of reading on vocabulary development. In his study he instructed a control group of 23 adult Indian students, learners of English, to use the dictionary (as in deliberate vocabulary learning) to find the meaning of 20 words appearing in an edited passage. The experimental group of 26 participants was instructed to simply read the passage for text comprehension (incidental learning strategy). In a second phase of the experiment he asked all the participants to use the newly acquired words in different sentences. He

found that learners who used dictionaries were unable to use the words learned through dictionaries in sentences. Conversely, learners who acquired word meanings incidentally while reading were able to use them actively in new sentences. Studies like this provide evidence for the effectiveness of incidental vocabulary learning not only for understanding the meaning of new words, but also for using them correctly in new contexts.

Webb (2008) investigated the effect of different types of context on incidental vocabulary learning. The author tested 50 intermediate Japanese university students, learners of English, who were instructed to read three sets of sentences, each one containing 10 target words unknown to the participants. The contexts in which the unknown words appeared were rated by English native speakers on their informativeness. Results showed that the type of context had a significant effect on incidental learning: more informative contexts produced higher retention of word meanings in learners. This finding was also confirmed in a subsequent study by Ahmad (2012), in which different types of context were taken into account.

Finally, Restrepo Ramos (2015) provided an extensive review of empirical studies on incidental vocabulary learning in the L2, and spelled out the pedagogical implications stemming from his literature review. He suggested that language practitioners should consider using authentic texts, adopting the type of text that best suits the interest of learners, and paying particular attention to the quality of contextual hints that enables the students to engage in incidental vocabulary learning.

4.4 Statistical learning based on crossing linguistic contexts and crossing situations

The bottom-up processes involved in L2 incidental vocabulary learning described above are typically related to the activity of reading. Word meanings are acquired incidentally, thanks to the repeated and varied exposure to several *linguistic* contexts in which such unknown words occur. However, the basic mechanism of meaning extraction from multiple and varied exposures may be applied to occurrences of unknown words (and concepts) in extra-linguistic contexts as well. As already observed in Chapter 2, in relation to children's linguistic development, humans have a remarkable capacity to detect regularities to construct meaning in a way that does not seem to require overt effort or even awareness (e.g., Kachergis, Yu and Shiffrin, 2014).

This type of indirect, implicit, non-instructed, bottom-up type learning, of which incidental vocabulary learning in reading is a sub-type, often goes under the name of *statistical learning*, and can be in principle applied to the extraction

of information from linguistic contexts (therefore based on word-to-word associations), as well as from situational contexts (therefore based on word-to-world associations). In this latter case it is referred to as cross-situational learning, as described in Chapter 2, and it appears to be a crucial strategy in first language acquisition. In its broadest sense, statistical learning entails the ability to detect patterns and deviations from patterns in the (linguistic or experiential) input. As we observed in this chapter, in relation to adult second/foreign language learning, as well as in Chapter 2 in relation to children's ability to learn word meanings despite the word-reference ambiguities to which they are exposed, statistical learning is based on the detection of regularities that can be observed in therefore both experiential contexts (typically for children) and language (typically for adult learners).

As briefly mentioned in Chapter 2, pioneering work performed by Yu and Smith suggested that statistical learning proceeds by means of associative learning that can be applied to both linguistic occurrences and regularities as well as extra-linguistic (experiential) contexts. As a matter of fact, preliminary works on statistical learning were based on linguistic input, and for example aimed at showing how 8-month-old infants are able to find word boundaries in an artificial language based only on statistical regularities (Saffran, Aslin and Newport, 1996), or how 12-month-olds could discriminate new strings of letters from artificial grammars to which they are exposed (Gomez and Gerken, 1999). Yu and Smith (2007) first applied the idea of detecting regularities across multiple exposures to extra-linguistic contexts and elaborated a computational associative model that is based on associations between words and situations as well as weights that are established between each word-situation pair, using conditional probabilities, that is, the probability of observing the occurrence of one item, given the presence of the other. We will see in further detail in Chapter 6 how the weights between words and situations are updated, exposure after exposure, taking into account not only their actual co-occurrence, but also the (unexpected) missed co-occurrence.

Although the underlying mechanisms of statistical learning in children and in adult language learners seem to be comparable, a comprehensive model of how statistical learning can take different types of configurations and be applied to different types of contexts and speakers is still missing. From a modelling point of view, what is missing is an identification and formalization of the exact mechanisms that enable the broad concept of statistical learning to take place, when acquiring new word meanings: how do speakers exactly retrieve the meaning of unknown words from the context? What type of operations do they undertake? Towards this broad goal, a number of researchers have recently employed adult learners to explore the underlying learning algorithms that could explain successful statistical word learning (e.g., Kachergis et al., 2014; Blythe, Smith and Smith, 2010; Trueswell, Medina, Hafri and Gleitman, 2013; Yu, Zhong and Fricker, 2012; Yu, 2008).

An open issue is whether statistical learning (e.g., in cross-situational learning) is implemented in one or two steps. The two-step view suggests that speakers would first retrieve from context all the possible meanings of a given word; then, in a second step, they would compare and intersect possible meanings collected across exposures (i.e., cross-tabulate them) via statistical procedures, to determine which one is the correct one. So, for example, when a label like *ball* appears to be pronounced in relation to a situation in which there is a ball, a dog, a leash, a tree, etc., the child would keep in mind *all* these possible associations (i.e.: *ball* might denote a ball, a leash, a dog and a tree). Then, after multiple exposures to various situations in which only one of these referents systematically appears, the child would learn the correct meaning of the word.

Medina and colleagues (2011) and subsequently Trueswell and colleagues (2013) argued, instead, that learners may rely on a single-step cross-situational learning strategy, and use a one-trial "fast-mapping" procedure, even under conditions of referential uncertainty. In other words, probably due to pragmatic needs, speakers would not wait to have all the necessary information retrieved from a vast number of contexts before formulating their hypothesis on the correct meaning of a previously unknown word. Instead, upon hearing an unknown word in context, they would start formulating a single conjecture on its possible meaning extracted from the single occurrence. Consequently, they will seek confirmation, i.e., a new context at least weakly consistent with the newly formed hypothesis. If this step succeeds, the conjecture is further solidified as a confident hypothesis of the word's meaning. Alternatively, if the new context does not confirm the hypothesized meaning, learners would shift to a new conjecture. This shift, however, comes at some cost, as Medina and colleagues explained:

> rather than returning to a state of semantic innocence, learners enter a memorial limbo, which leaves some residue of confusion that interferes with subsequent learning. This confusion is eliminated after a considerable delay, whereupon the machinery returns to its initial state and can again form a first conjecture.
>
> (Medina et al., 2011, p. 9017)

Trueswell and colleagues elaborated this view and suggested that participants seem to provisionally pair novel words with possible referents and then use a statistical-associative learning mechanism to decide whether such guess is reliable (i.e., can be kept) or instead whether it is disconfirmed across situations, and needs to be rejected. By doing so, participants would gradually converge to a single mapping across learning instances, in a learning procedure that the authors named *Propose-but-Verify*. It remains however unclear where the initial guess would come from: would the initial pairing come from innate biases? Would it come from random initial associations, used to 'fast mapping' words and entities, that then undergo the *Propose-but-Verify* procedure?

Fitneva and Christiansen (2011) studied adults' behavior in a cross-situational learning task, in which participants were asked to learn an artificial language, where words were presented together with possible referents over multiple exposures. Participants were first asked to observe the sequences of words and objects, and then in a second phase to name the various objects with the words learned over the multiple exposures. Visual fixation data were used to assess the direction of their visual attention during the first phase. The authors discovered that participants whose longest fixations in the initial trials fell more often on the images that appeared as distractors (i.e., the incorrect referents) performed significantly better in naming the objects with the newly acquired words than participants whose longest fixations fell more often on the 'correct' images. Thus, the authors concluded, inaccurate word-referent mappings seem to actually benefit learning.

Finally, by means of a corpus analysis of adult-child directed speech, Monaghan and Mattock (2012) found that the presence of grammatical cues helps children in cross-situational word learning. By analyzing the corpus of a word learning study focused on adult-child speech, the authors found that when both object-referring and non-referring words occurred in the utterance produced by the adult, referring words were more likely to be preceded by a determiner. This was however not the case when the utterance contained *only* referring words. Adults seemed to use determiners to help children learn the correct word-referent mapping, in presence of possible distractors. This finding suggests that syntactic cues (such as articles) are used by adults to help children disambiguate between possible referents.

4.5 Pattern detection: A hallmark of human cognition

A pattern is generally defined as a repeated or regular way in which something is done, organized, or happens. In cognitive psychology various theories have been proposed to account for how we recognize patterns in the wild. These range from template matching to prototype-matching, and from feature analysis to recognition-by-components. Humans, from a very young age, are extremely good at constructing and detecting patterns and deviations from patterns in many circumstances, including language, music, and visual stimuli (Kurzweil, 2012). Being able to detect the repetition of items that characterizes a pattern enables us to perform classifications of items to which we are exposed. These can be sounds, letters, words, objects, experiences, and so on.

But how are such classifications performed, starting from the detection of recurring elements in the input? What is the relation between the ability to detect patterns and the ability of categorizing, a hallmark of human cognition?

The core thesis of this book is hereby anticipated at the end of this first Part (which encompasses Chapters 1–4), motivated at the end of Part 2 (which encompasses Chapters 5–7) and then fully elaborated in the third and last Part of the book (Chapters 8–10), which provides the converging evidence in language and communication research.

The basic mechanism that underlies the human ability to perform categorizations and construct word meaning (and conceptual representations) by grouping together similar experiences and similar linguistic structures consists of three steps. The first step consists of a broadly defined associative process, thanks to which we connect together objects, sounds, words, experiences, based on the amount of times they appear (and they do *not* appear) in the same contexts (Figure 6a).[2] The second step consists of a pattern detection mechanism, in which we detect similar configurations of context-based representations (Figure 6b), and the third step consists of a feature-matching process,[3] thanks to which we tend to see as similar to one another items that share many common features (Figure 6c).[4]

The most important conceptual operation that links the three steps is a *switch* from the syntagmatic[5] to the paradigmatic level of analysis, which enables humans to move from the direct observation and detection of the repetition of items that

2. The establishment of associations is considered to be the basis of learning in cognitive psychology, cognitive science and neuroscience. Its cognitive reality is seen in classical conditioning (Pavlovian effect) discussed in more detail in Chapter 6.

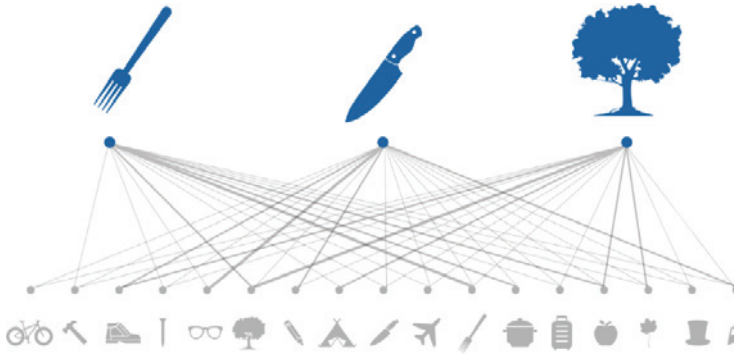
3. Feature matching processes are widely used and psychologically supported in cognitive science, neuroscience and cognitive psychology. The classic feature matching model was proposed by Tversky (1977) and used to explain conceptual categorizations and relations of prototypicality and family resemblance.

4. As I will explain in detail in Part 3, such features can be intrinsic features of the compared items, or extrinsic features, such as contextual features. For example, for the object *mug*, the features *handle* and *porcelain* are intrinsic, because they belong to the entity *mug*. Conversely, the features *milk* and *bottle* are extrinsic, because they do not belong to the entity *mug* but to the typical contexts in which cups can be found.

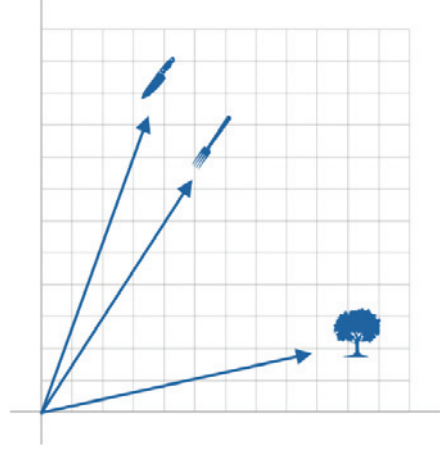
5. In linguistics, a syntagmatic relation is defined as the relation between two linguistic entities (phonemes, morphemes, words or utterances) that occur in the same text. Syntagmatic analyses are therefore focused on the rules of combination between linguistic entities, such as the combinatorial properties of nouns and verbs (e.g, *child-runs*). Paradigmatic relations, instead, hold between words that occur in similar contexts, but do not occur (necessarily) together. This is typically the case of synonyms and antonyms (e.g, *child-kid*, or *young-old*). In this book I refer to syntagmatic relation as to associations established between entities that tend to occur in similar contexts. As I will explain in Chapter 6, the strength of such associations is informed not only by the actual co-occurrence but also by the missed co-occurrence between two entities that are expected to co-occur, based on previous experience.

occur (or do not occur) in the same contexts, to the categorization of such items, based on the similarity between their patterns of occurrence. This is, I argue, the core mechanism that explains how we are capable of abstracting and constructing meanings, categories and relations between them, moving beyond the simple observation of elements that co-occur or fail to co-occur in language and in perceptual experience.

a.



b.



c.

Figure 6. The construction of categories based on co-occurrences of objects in contexts

The syntagmatic level of analysis is a level at which we can observe stimuli and establish associations between items that appear (and fail to appear) together in context. Figure 6a illustrates how this process is setup, starting from the exposure to entities in context. In this figure, assuming that each circle represents a context, or an experience to which a speaker is exposed, the fork initially entertains (broadly speaking) syntagmatic relations with all the items that co-occur with it. Later, the strength of these associations is updated on the basis of further co-occurrences as well as missed (but expected co-occurrences). Similarly, the knife and the tree entertain relations with the items that co-occur with them.

A number of empirical studies have shown that such syntagmatic (or thematic) connections between items that tend to appear in the same contexts are indeed part of the repertoire of connections that we have in mind. Although typically based on words only, these studies tend to use methods based on priming paradigms, according to which if two words A and B are connected, then seeing or hearing word A will facilitate the processing of word B. For example, Moss, Ostrin, Tyler, and Marslen-Wilson (1995) showed priming effects based on various types of event knowledge, involving tools (such as *broom-floor*) and participants (*hospital-doctor*). Moreover, Ferretti, McRae, and Hatherell (2001) have shown that verbs prime their typical agents (*cooking-chef*), their typical patients (*serving-customer*), and typical instruments used to perform actions (*stirred-spoon*), but not the typical locations (*skated-arena*). One possible explanation for this lack of priming is that locations tend to be background information for most situations and thus may not be as salient in the causal structure of events as are agents, patients, and instruments. In a subsequent study it was found that locations (e.g., *arena*) are primed by prototypical actions expressed by verbs with an imperfective aspect (e.g., *was skating*) but not verbs with a perfective aspect (e.g., *had skated*) (Ferretti, Kutas and McRae, 2007).

Once the syntagmatic relations are established between items that co-occur, these links are strengthened thanks to the exposure to new experiences in which the pairs co-occur or fail to co-occur (in Figure 6B this is graphically visualized by the thickness of the lines that connect each of the three objects to the contextual entities: the thicker the line, the stronger is the connection).

The switch to the paradigmatic level of analysis occurs after the patterns of co-occurrences shared by pairs of items are detected. For instance, Figure 6b shows the pattern detection stage for the three objects: fork, knife and tree. Each co-occurrence between one of these three objects and its contextual entities is weighted and strengthened thanks to multiple exposures to contexts and situations. The configuration of each object's pattern of co-occurrences is compared to the patterns of the other objects, and (paradigmatic) similarity emerges between objects as a result of a feature-matching process, where the features that are matched consist of

the co-occurring entities. The similarity constructed between the three semantic representations of the three items (fork, knife and tree) is displayed in Figure 6c in terms of proximity. In this hypothetical situation, the semantic representations constructed for the fork and the knife are closer to one another than the representations of fork and tree, and knife and tree, because the pattern of associations with other objects (6a) of the fork and the knife are distributionally more similar to one another than to the pattern of associations constructed for tree.

While Figures 6a, 6b and 6c use objects to illustrate the relations that each item entertains with other items in a small set of 3 hypothetical contexts, it shall be clarified that in principle each object could be replaced with a phoneme, which co-occurs with other phonemes to form words, or it could be replaced with a whole word, which co-occur with other words in linguistic contexts, and so on. Moreover, relations like those illustrated in Figure 6a can also be established between an object and its components or its features. For example, an association may be established between a mug and its handle, a teddy bear and its softness, or a ball and its roundness because these pairs are repeatedly observed occurring together, and they are typically experienced together, and the occurrence of a teddy bear predicts the presence of softness. These associations established between objects or components, experienced through our bodies and through our senses are broadly based on **world-to-world** associations, i.e., connections between objects, objects and their parts, or objects and their perceptual properties, which all belong to the experiential domain. Syntagmatic associations, however, may be established as well between objects and the words that are repeatedly used to name them. Being exposed to a ball and hearing at the same time the word *ball* enables children to connect the object ball with the word *ball* syntagmatically, in a **word-to-world** type of syntagmatic reference, that is repeatedly experienced in various contexts. Finally, hearing repeatedly the word *ball* together with the word *play*, will enable children the establishment of a syntagmatic relation of the **word-to-word** type. These different types of syntagmatic relations are visualized in Figure 7.

Note that the type of relations called word-to-world are the associations observed in relation to word learning, tested in cross-situational experiments (described in Chapter 2 and in this chapter). Children (and adults) learn to associate a word with the correct referent, by cross-mapping the situations in which it occurs, in order to disambiguate between concurrent references. Similarly, the type of relations called word-to-word are those that characterize the incidental vocabulary learning (typical of second/foreign language learners).

The three different types of associations illustrated in Figure 7, which can in principle be used to construct semantic representations in the way illustrated by Figure 6, are hereby described.

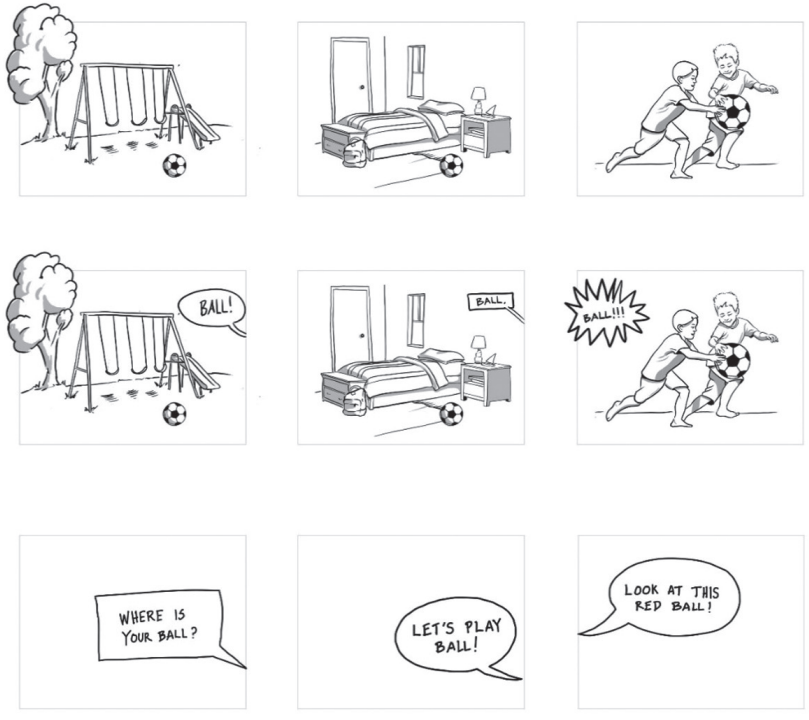


Figure 7. Syntagmatic relations of the type world-to-world (top row of contexts), word-to-world (middle row of contexts), and word-to-word (bottom row of contexts), for the meaning of *ball*

World-to-world associations

Thanks to multiple exposures to experiential contexts, and to the establishment of syntagmatic relations between items that co-occur (e.g., a cup and a liquid substance, a cup and a hand that holds the cup, a cup and a table; a mug and a liquid substance, a mug and a hand that holds the mug, a mug and a table) a paradigmatic relation of distributional similarity is established between items that tend to occur with the same items in experiential contexts (i.e.: the cup and the mug, which both tend to appear with liquid substances, with hands, and with tables). Similarly, syntagmatic associations may be established between objects and their salient features (components, perceptual properties such as shape, texture, material). This is a crucial point: items perceived through perceptual experiences that can be associated to a word can be both objects that appear in the same experiential contexts

or entity-related features such as components and qualities.[6] Consequently, objects that share similar features are perceived to be similar to one another (e.g., a cup and a mug have similar shapes and and appear together with similar objects, although they do not share the handle component, and are thus perceived to be similar). The detection of similarities based on experiential features allows us to group together items that are distributionally similar, to form basic experience-based categorizations. Cups and mugs become members of the same category (containers used for drinking) because they share similar world-to-world patterns of associations. Objects that share similar patterns of features are similar and thus are categorized together.

Word-to-world associations

Thanks to multiple simultaneous exposures to objects and words, syntagmatic relations are established between referents and their linguistic labels, once the correct word-referent attribution is established by means of cross-situational learning (see Chapter 2). Note that words are used to name various instances of referents that typically appear in very similar contexts and share many features, such as a white mug and a red mug. These two items are not identical, and yet are both associated with the word *mug*. This phenomenon enables a type of categorization that differs from the purely experience-driven categorization described above. As illustrated in Chapter 2, imposing a label (a word) on two objects that are non-identical (a red mug and a white mug, maybe having slightly different shapes and dimensions) stimulates the search for commonalities between the two objects. This type of categorization is therefore language-driven. The similarity between two objects (red mug and white mug) is driven by the fact that both objects are named with the same word (*mug*) and thus both share the association with the word *mug*. Objects named with the same linguistic label are similar and thus are categorized together.

Word-to-word associations

Thanks to multiple exposures to words in linguistic contexts (such as during reading, but also in younger children during listening and dialogues) syntagmatic relations are established between words that tend to occur repeatedly together (e.g.,

6. In knowledge-based taxonomies of feature types such as the taxonomy proposed by Wu and Barsalou (2009) but already reported in Appendix F in McRae et al. (2005) a distinction is made between entity-related features (which include components, quantities, perceptual properties etc.) and situation-based features (such as associated entities, locations, participants etc.). Both these types of features are often linked to words by means of word-to-world associations.

play and *ball, your* and *ball, red* and *ball*). Paradigmatic relations are then established between words that tend to occur within the same linguistic contexts, that is, with the same other words (e.g., *teddy bear* is paradigmatically associated to *ball*, because both may tend to appear together with words like *play*, and *your*). The detection of similarities based on linguistic co-occurrences allows us to group together words that are distributionally similar, to form categories of word meanings, based on their distribution across texts. Words that appear in the same linguistic contexts are similar, and thus are categorized together.

It is important to notice that the three types of categorizations emerging from the three types of syntagmatic relations (world-to-world, word-to-world, and word-to-word) do not produce necessarily different categorizations. On the opposite, they arguably produce similar categorizations. The streams of information used to perform the switch from syntagmatic to paradigmatic levels are qualitatively and theoretically distinct from one another, but they tend generate comparable results. For example, a mug and a cup are distributionally similar to one another, and such similarity is arguably retrievable from more than one type of syntagmatic relation, among the three types described above. Cups and mugs have similar distributional properties across experiential contexts (e.g., they both tend to appear together with liquid substances, and have similar shapes), are used to name similar objects, and the related words *cup* and *mug* are used in similar linguistic contexts (e.g., with words such as *drink, sip, coffee*, etc.). Therefore, the three types of similarity, although constructed through paradigmatic shifts that evolve from theoretically distinct syntagmatic relations, tend to converge toward similar classifications.

4.6 Summary

The Revised Hierarchical Model is a classic model for the bilingual mental lexicon. This model, a cornerstone in language acquisition research, has been recently challenged by empirical findings that showed that word meanings and their representations in the L1 and the L2 are not static and embedded in separate, language-specific lexica, but are dynamic, flexible, and emerge in a bottom-up manner from experiences and linguistic encounters. Word meaning representations and models of the mental lexicon (including the bilingual mental lexicon) may emerge without any top-down constraint or rule from simple repeated occurrences, observed across various types of context.

Native speakers and foreign language learners display some differences in the way they construct word associations. While the associations indicated by native speakers tend to be coherent and systematic between participants, based

on semantic (paradigmatic) relations, the associations indicated by language learners tend to be less systematic between participants, and based on different criteria, among which are featured a syntagmatic, episode-based criterion (thematic relations) as well as linguistic criteria such as phonological similarity and lexical translations.

Statistical learning is an influential theoretical paradigm (that encompasses cross-situational learning as well as incidental vocabulary learning) that summarizes the approaches taken by native speakers and language learners in learning word meaning from multiple exposures to situations (cross-situational learning) and from linguistic contexts (incidental vocabulary learning). The latter type of learning characterizes, typically, the way in which vocabulary in a target language is acquired by language learners.

In the last part of this chapter, based on the empirical findings reviewed in the previous sections and partially in Chapter 2, I claimed that a common underlying principle explains how world-to-world associations, word-to-world associations, and word-to-word associations (all based on syntagmatic relations) enable us to establish paradigmatic relations of similarity between objects, objects and words, and words. This mechanism encompasses a switch from syntagmatic to paradigmatic associations that allows us to perform categorizations, and classify concepts on the basis of their distributions across different types of contexts.

Overall, in this first part of the volume, I focused on reviewing psychological evidence that shows how word meanings are learned, and in particular how they are learned from perceptual experiences as well as from other words, in both, children and adults (both, native speakers and language learners). I have also explained some basic cognitive mechanisms through which new word meanings are derived from old ones (based on metonymy and metaphor, which are specific types of polysemy), and gave quick overviews of computational models implemented with the purpose of modelling such mechanisms. I briefly introduced the structure of classic top-down, rule-based hierarchical models that accounts for the construction of the mental lexicon in children (the constraints described by Clark) and the bilingual mental lexicon (the Revised Hierarchical Model), and explained the limitations of such models. I then explained how bottom-up, non-rule-based models can overcome such limitation and explain how word meaning is learned by children and adults alike (by means of statistical learning such as cross-situational learning and incidental vocabulary learning). All these elements will allow me to introduce and motivate, in Part 2, the variety of bottom-up computational models proposed in the past thirty years to account for the nature of word meaning and semantic representation. These models are all based on the distributional hypothesis (Harris, 1954; Firth, 1957), which I will clarify and exemplify further in Part 2. In Part 3, finally, I will bring together insights derived from psychological evidence

described in Part 1, and insights derived from computational modelling described in Part 2, to show how this converging evidence coming from (broadly speaking) the cognitive sciences and the computer sciences can inform us on the nature of word meaning. Specifically, in Part 3 I will spell out my own proposal (which I briefly anticipated at the end of this chapter) suggesting that the distributional hypothesis has deep cognitive foundations, and when applied adequately to various types of contexts (i.e., experiential and linguistic ones) can explain where and how word meaning is constructed by, respectively, children, adult native speakers, and adult language learners.

PART 2

Word meaning construction and representation in the artificial mind

CHAPTER 5

Distributional models and word embeddings

5.1 You shall know a word by the company it keeps

Back in the early Nineties, the idea that word meaning could be acquired by humans simply by comparing distributional patterns of words collected over text data mainly during the activity of reading gained popularity among computer scientists who aimed at simulating vocabulary learning, modelling the relations among words in the mental lexicon, and attempting to crack the nature and origin of word meaning. The need to access and represent semantic information about lexical items was a key step also for the implementation of machines that could automatically recognize speech, one of the first applications of neural networks. As Schütze (1993) mentions in a classic example, the pioneering machine for speech recognition Vocoder, implemented at Bell Labs in 1928, allegedly once mis-rendered the statement *recognize speech* into *wreck a nice beach*, because of the homophony of the two phrases. Such mistake shows that a system that tries to recognize speech by simply relying on the stream of sound waves can make mistakes that, given appropriate context, a human being would not do.[1]

In the classic connectionist literature, the techniques used to represent word meaning were not capable to scale up and construct semantic representations for a large number of words in parallel (see Schütze, 1993 for a review). The introduction of vector spaces, or Word Spaces, as first labelled by Schütze (1993) made this operation possible. This chapter describes what vector spaces (or distributional models) are, by focusing on the description of the most popular exemplar of this category of models: Latent Semantic Analysis (henceforth LSA, Landauer and Dumais, 1997).

LSA is a model of word meaning construction and representation based on the distributional hypothesis (Harris, 1954), which is motivated by the observation that words with similar meanings tend to appear in similar contexts and suggests that word meanings can be learned by looking at how words behave in context. LSA is a mathematical model of word meaning based on a conceptual metaphor

1. Happens, however, to misinterpret song lyrics. For example, instead of hearing "and it seems to me you lived your life *like a candle* in the wind" (by Elton John) might happen to hear "and it seems to me you lived your life *like a Ken doll* in the wind".

(Lakoff and Johnson, 1980), which can be summarized as follows: SIMILARITY-IS-PROXIMITY. In LSA, in fact, the semantic similarity between two words is operationalized in terms of the geometrical proximity between the vectors that represent each of the two words in a *n*-dimensional space. As Sahlgren (2006) summarizes: "meanings are locations in a semantic space, and semantic similarity is proximity between the locations" (Sahlgren 2006, p. 19).

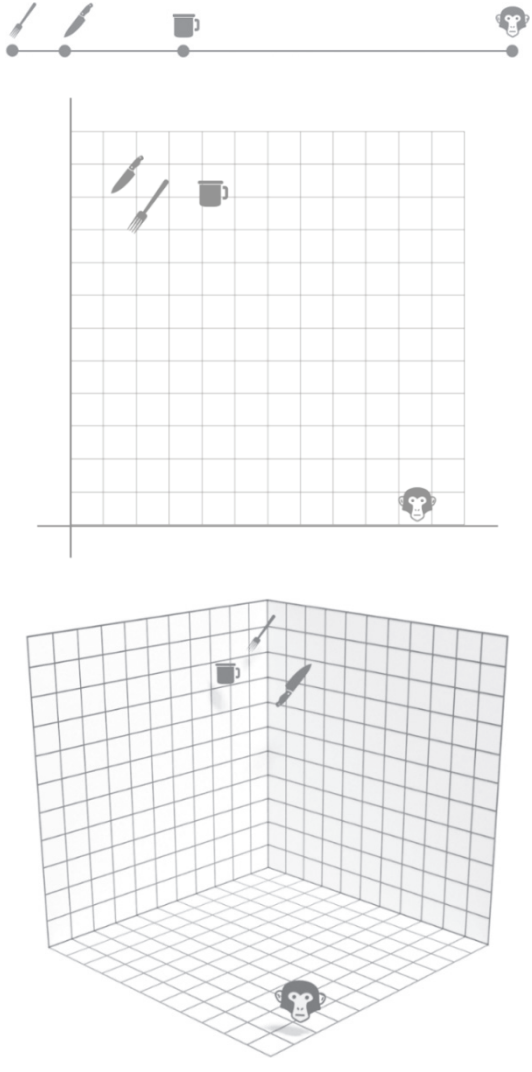


Figure 8. The geometrical proximity between *fork, knife, cup*, and *monkey* in a one-dimensional, a two-dimensional, and a three-dimensional space

Making use of vectors to construct word meaning, distributional models represent word meanings in a highly multidimensional space, which is quite hard to imagine and graphically visualize. To grasp this idea, the difference between proximities represented in spaces with varying dimensions is displayed in Figure 8. Here, I exemplified the proximity obtained from a hypothetical distributional analysis between the following words: *fork, knife, cup, monkey*, in a one-dimensional (A), a two-dimensional (B), and a three-dimensional (C) space, respectively. Word meanings are represented by means of visual symbols, for clarity. Note that the proximities displayed between semantic representations in Figure 8 are only hypothetical, not based on real corpus data. For distributional similarities based on actual corpus data, see Section 5.2.

In each of the spaces illustrated in Figure 8, from a mathematical point of view, the position of each semantic representation is determined by numerical coordinates, each telling us where the word is, in relation to one of the dimensions represented by an axis. In the mono-dimensional space (Figure 8A), for example, the position of each word is determined by its distance from the origin of the horizontal line, x, which is the only dimension the points relate to. This single coordinate is sufficient, in a mono-dimensional space like that in Figure 8A, to display the four meanings. Once each word is represented by a numerical value that shows its position in the mono-dimensional space (i.e., on the axis x), it is possible to calculate the distances between word pairs, by comparing the coordinates that represent each of the words. This will result in *fork* and *knife*, for example, being closer to one another (i.e., shorter distance measured on the axis x) compared to *fork* and *cup*, or *fork* and *monkey*, or *knife* and *monkey*, etc. The geometrical proximity reflects the (distributional) semantic similarity between the two meanings. In mathematical terms, the distance *d* between two words *A* and *B* in a mono-dimensional space on the axis *x*, is formalized by the following mathematical formula:

$$d_{(A,B)} = |X_A - X_B|$$

where the coordinate of each of the two words is represented as x_A (for word A) and x_B (for word B), and the vertical bars stand for the absolute value of the difference between the two coordinates.

When each of the words is represented in a two-dimensional space, xy, then each point has not one but two coordinates, each determining the position of the point in relation to one of the two axes, x and y. Therefore, in Figure 8B, for example, each of the words is represented numerically by a pair of coordinates. While the actual distance between each two points is intuitively easy to understand (this is the length of the shortest line that can be drawn between point A and point B), the formula for calculating such distance mathematically becomes slightly more complex than in the mono-dimensional space. In Euclidean geometry, the

distance between two points A (x_A, y_A) and B (x_B, y_B) can be derived from the Pythagorean theorem as the length of the hypothenuse of a rectangular triangle, where the hypothenuse is the shortest distance between the two points, and the two sides are each parallel to one of the two axes, x and y.

$$d_{(A,B)} = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$$

With this formula it is possible to calculate the hypothenuse of the rectangular triangle (that is, the distance between two points in a bidimensional space) by comparing the coordinates obtained by projecting the triangle's sides on each of the two axes, x and y. Similarly, in Figure 8C each word is a point in a three-dimensional space, and its position can be formalized as a series of three numbers, each giving the position of the word in relation to one of the three axes, x, y, and z. In mathematical terms, the formula to calculate the geometrical distance between two points in a three-dimensional space does not differ much from the geometrical distance calculated in a two-dimensional space.

$$d_{(A,B)} = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2 + (z_A - z_B)^2}$$

In principle, two words (or two points) may be represented in any multidimensional space, with hundreds or thousands of dimensions, given that the equivalent number of coordinates is provided, to assess their relative position in relation to each dimension. This is however quite hard to imagine (let aside to illustrate it on a sheet of paper).

The sequence of coordinates that represents numerically the position of a point A in a multidimensional space of n dimensions is called a *vector*. The geometrical proximity between two vectors, which accounts for the similarity between the words for which each vector stands, in distributional semantics is typically calculated as the cosine measure between the two vectors. The cosine similarity (CosSim) is computed with the following formula:

$$\text{CosSim} = \frac{\sum_{i=1}^{i=n} a_i \times b_i}{\sqrt{\sum_{i=1}^{i=n} a^2} \times \sqrt{\sum_{i=1}^{i=n} b^2}}$$

where a and b are two vectors, each with n dimensions. To grasp the meaning of this formula, it might be helpful to realize that it does not differ very much from the formula used to calculate the Pearson's correlation coefficients between two variables.[2]

2. The Pearson's correlation coefficient indicates to what extent two variables are related to one another. Mathematically, this is defined as the relation between the co-variances of the two variables and the product of their individual standard deviations. Basically, the only difference between these two formulas (i.e., the CosSim and the Pearson's coefficient) is that while the cosine

As explained above, distributional models use these mathematical notions (i.e., vectors, coordinates, cosine similarity etc.) to model word meaning. A word is mathematically represented by a vector, whose dimensions are the linguistic contexts in which the word appears. The exact coordinates of the word are the measures of association between the word and each of the considered contexts. It follows that the word meaning is formalized as a list of numerical coordinates, each determining the relation between the word and a linguistic context. In LSA the linguistic contexts consist of whole documents: each context is a text in which a target word appears, and the strength of the association between a word and a document is typically calculated with a measurement called tf-idf (term frequency-inverse document frequency). As the name of this measure suggests, the weight of the association between a term and a document, and therefore the relevance of a term within a document, is calculated by a formula that takes into account both, the frequency of the term in the document as well as the number of documents in the corpus that contain the word. This helps to adjust for the fact that some words appear more frequently than others, in general.

What makes this model revolutionary compared to other models of meaning representation is that the semantic space is built automatically, in a bottom-up manner or, in technical terms, in an unsupervised manner, from the automatic acquisition of contextual information from corpora. In this model, and in the distributional models derived from it and inspired by it, words that tend to appear in similar contexts cluster together in the multidimensional space. This model, therefore, is not simply corpus-*based*, such as models where lexicographers set rules in a top-down manner to determine how word meanings shall be organized, and then they find examples in corpora to corroborate their claims. LSA is corpus-*driven*, which means that semantic representations and relations between them *emerge* from the distribution of corpus occurrences.

In this sense, the vectors constructed by distributional semantic spaces reflect a prototype-based approach to the construction of meaning (Erk and Padò 2010; Sikos and Padò 2019). In these models, a word meaning is represented by a single vector of features/contexts. Such vector represents a prototype for the

formula deploys the coordinates within the two vectors as they are, Pearson's formula does not operate directly on the coordinates, but on their standard deviations. The two values therefore capture different ways in which two vectors are similar to one another. If the cosine and the Pearson coefficient values are calculated between vectors a and b, and then a constant value k is added to each of the coordinates of vector a, the cosine similarity between a and b will change, because vector a changes its proximity to vector b, while the Pearson coefficient between the two vectors will remain the same. Usually, in distributional modelling, cosine similarities are preferred to calculate the exact proximity between two vectors in a multidimensional space, for their higher sensitivity compared to Pearson correlation coefficients.

category that it refers to, which can be seen as an abstraction over individual instances (e.g., Rosch 1975). A highly competitive theory of meaning representation argues that conceptual categories are formed and represented via the specific instances of their experiences (Nosofsky 1986; Daelemans and van den Bosch, 2005; Chandler 2017). According to this theory, constructing a conceptual category from experience does not require any form of abstraction (such as the construction of a prototype): all specific experiences are directly stored and retrieved when needed, as exemplars that instantiate the category. New exemplars are classified by similarity in relation to the nearby exemplars, using a mechanism that resembles Saussure's analogical reasoning (De Saussure, 1916). In computational modelling, in exemplar-based models each instance of a category is represented by a vector. These models, therefore, require the management of large amounts of vectors, each for an instance or an episodic experience of a conceptual category. Despite this, exemplar-based models have proven to be highly competitive for the representation of word meaning, for various reason. Notably, they are capable of handling polysemy very well. In fact, while a prototype-based model (like LSA) constructs a single vector for a word meaning, in which typically the contexts of different senses of a polysemous word are conflated, an exemplar-based model like TiMBL (Daelemans and van den Bosch, 2005) constructs several vectors for the same word meaning, and can then remove all the vectors for the exemplars that are not relevant to represent meaning in a given context, by analogy with similar occurrences. Moreover, by keeping a record of each and every event, exemplar-based models keep track of low-frequency events as well, which in some cases tend to be filtered out in prototype-based methods. By keeping track of rare events, exemplar-based models allow nlp scholars to tackle exceptions and subregularities in language, as indicated by Daelemans and van den Bosch (2005). Recently, a direct comparison between a prototype model that produces a summary representation of its categories, and an exemplar model that represents individual instances (both implemented using the same embedding model) has shown that the first outperforms the latter in a frame identification task (Sikos and Padò 2019). Overall, the jury is still out about the final judgment between these two theory-driven models of meaning representation.

5.2 Constructing distributional models

While we will not go too much into the details of the mathematical formulas that characterize the implementation of various types of distributional models, in this section I will outline their general mechanisms. In particular, I will focus on explaining how the switch from syntagmatic to paradigmatic relations between

words takes place in these models, and I will elaborate the cognitive underpinnings of this its functioning.

Consider, for example, the words used in Figure 8 to construct a distributional semantic space. In order to model the (distributional) similarity between these words, as a first step the contexts of occurrence for each of the words shall be collected from corpora. This, by itself, is a non-trivial operation, because a linguistic context extracted from a corpus can take different configurations: it can be a single word that appears with one of the target words within a window of text of a predetermined size; it can be a syntactic pattern in which the word is embedded; or it can be a whole document in which the target word appears (as in LSA). Determining the type of context that is used to implement the distributional semantic space is already an operation that affects the type of syntagmatic relation that the target word entertains with said context, and therefore the type of paradigmatic similarity that the target word has with other words. For example, the words *boy* and *girl* may be distributionally similar if we look at texts in which we talk about *people, crowds, gender, communities*, etc. However, if we look at the syntactic and semantic collocations used respectively with these two words (which are tighter contexts, more heavily constrained) then the distributional similarity between the words *boy* and *girl* may decrease, and we may find that we tend to use each of these two words in different specific collocations, with different verbs.

Let us now see a concrete example in which we show linguistic contexts for the four words used in Figure 8. Typically, corpus data are annotated (i.e., pos-tagged, parsed, etc). Here, for the sake of simplicity and clarity, I indicated a variety of raw contexts in which the target words can be used, extracted from the word-sketches listed on Sketch Engine.[3] I selected 6 contexts for each of the 4 target words.

(6) *fork*
stick a fork
stab with a fork
handmade fork
light fork
clean fork
pick up a fork

(7) *knife*
sharp knife
stick a knife
Swiss knife
handmade knife
stab with a knife
pick up a knife

3. https://www.sketchengine.eu/

(8) *cup* *coffee cup*
plastic cup
drink with a cup
pick up a cup
round cup
clean cup

(9) *monkey* *jumping monkey*
monkey cage
rescue a monkey
intelligent monkey
macaque monkey
behave like a monkey

Once the contexts of occurrence are extracted from corpora for each of the target words, a contingency matrix needs to be created, where all the unique contexts (of all the target words) are collected together with the target words and the weighted co-occurrences of each word within each context. An example of such matrix is displayed in Table 1.

Table 1. Contingency matrix displaying the frequencies of occurrence of each of the four target words with each of the contexts extracted for the four words

	Fork	**Knife**	**Cup**	**Monkey**
stick a	532	706	0	0
stab with a	40	256	0	0
handmade	10	18	0	0
light	26	9	0	0
clean	10	38	0	0
pick up a	300	761	698	0
sharp	0	6011	0	0
Swiss	0	2842	0	0
coffee	0	0	7845	0
round	0	0	43	0
plastic	0	0	4624	0
drink with a	0	0	5988	0
jumping	0	0	0	312
cage	0	0	0	512
rescue a	0	0	0	267
intelligent	0	0	0	30
macaque	0	0	0	1178
behave like a	0	0	0	113

In the implementation of a distributional semantic space, the cells that correspond to each word-context intersection are automatically filled with the word/context co-occurrences extracted from corpora. In this case, I simply indicated the number of hits that are reported by Sketch Engine for each word-context pair, extracted from the internet-based corpus EnTenTen 2015. The screenshot in Figure 9 shows some of these co-occurrences, for the word *fork* with the context *stick a*.

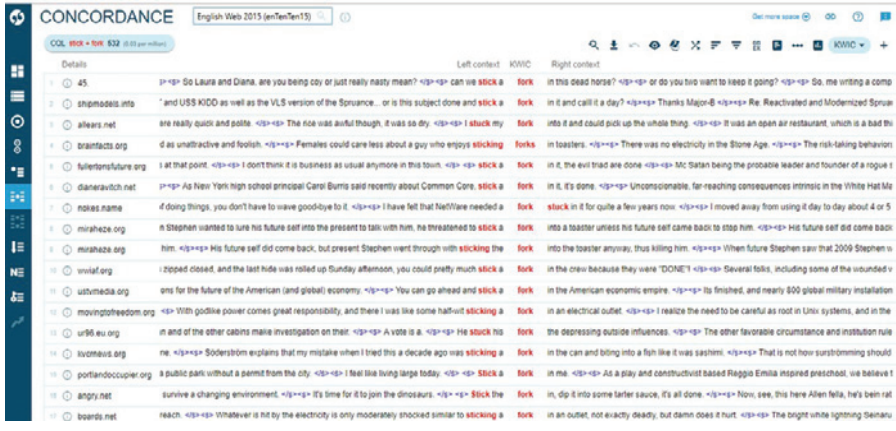


Figure 9. Screenshot showing the concordances for the word *fork* used within the context *stick a*, extracted from SketchEngine, using the corpus EnTenTen 2015

The raw frequencies of occurrence between a word and a context are typically then transformed into more informative measures of association, which give an approximation of how strong the syntagmatic relation between a word and a given context is, despite the frequency of the individual entities taken alone. This is achieved, for example, by dividing the raw co-occurrence by the overall individual occurrences of the word and of the context alone. One of the most commonly used measures of association is called Mutual Information (Church and Hanks, 1989), and it basically measures the logarithmic relationship between the observed frequency and the expected frequency of a word within a context.[4] The reason why raw frequencies of co-occurrence are turned into measures of association can be found in the well-known phenomenon that goes under the name of Zipf's law

4. The formula to calculate the Mutual Information is commonly used to implement distributional models, although it has been shown that it tends to favour the most idiosyncratic contexts of each word, which are also general and less frequent, but various adaptations have been proposed in the scientific literature, to adjust for this peculiarity (e.g., Baroni & Lenci, 2010).

(1949):[5] if we look at how words appear in corpora, a few words occur very frequently, while the vast majority of words are used very seldom. Therefore, looking at words in corpora, it is very likely to observe high co-occurrence scores between those few highly frequent words, and low co-occurrence scores between all those words that are not as frequently used. However, a high level of co-occurrence between highly frequent words is not as informative as a high level of co-occurrence between rare words. For this reason, it is crucial to take into account not only the frequency of the co-occurrence, but also the overall frequency of the individual words in the corpus, to effectively weight the strength of their association. A measure of association provides this type of information.

At this point, each of the target words, *fork, knife, cup* and *monkey*, is represented numerically by a list of coordinates, each indicating the relation between the word and one of its dimensions of meaning (aka a context), in a multidimensional space that in this case consists of 18 dimensions (the 18 unique contexts). For example, looking at Table 1 where the target words are displayed on the columns (note that usually the matrix is transposed and the target words are displayed on the rows), the meaning of *fork* is represented as follows:

fork = (532, 40, 10, 26, 10, 300, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)

whereas the meaning of *knife* is represented as follows:

knife = (706, 256, 18, 9, 38, 761, 6011, 2842, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0).

The representation of each word is therefore distributed across a series of coordinates. The geometrical proximity between these two vectors can be calculated using the cosine formula illustrated in Section 5.1. Cosines values range from −1 to 1, where the maximum value, 1, represents the complete overlap between the two vectors, which are therefore identical on each and every dimension. A cosine equal to zero represents two orthogonal vectors. In terms of meaning similarities, the higher the value of the cosine between two words vectors (i.e., the closer to 1), the higher the distributional similarity between the two words. Based on the contingency matrix displayed in Table 1, the table of cosine similarities (i.e., geometrical proximities) between each pair of word vectors is reported in Table 2.

As Table 2 shows, the proximities between each pair of words are symmetrical, and the values on the diagonal of this table represent the proximity between a word and itself, which is 1, the maximum value.

5. George Kingsley Zipf (1902-1950) reported the observation that the frequency of a word seems to be a power law function of its frequency rank. The law is commonly referred to as Zipf law.

Table 2. Cosine similarities between the four vectors representing the four target words

	Fork	Knife	Cup	Monkey
fork	1	0.149	0	0
knife	0.149	1	0.007	0
cup	0.031	0.007	1	0
monkey	0	0	0	1

Table 2 displays therefore a specific type of paradigmatic relation between words: the distributional similarity obtained by looking at word statistical patterns of occurrence across a corpus of text. The resulting proximities (i.e., cosine values displayed in Table 2), as previously explained, indicate how close two word vectors are in a multidimensional space which, in this case, encompasses 18 dimensions (see the 18 coordinates that compose each word vector, displayed in Table 1). It is impossible to imagine or visualize the proximities between word meanings represented by vectors in an 18-dimensional space.[6] Fortunately, there are mathematical techniques that allow us to condense and visualize the word representations into a space with less dimensions, typically a two-dimensional semantic space, by using a variety of algorithms. One of the most commonly used algorithms is Multidimensional Scaling (MDS, see for example Cox and Cox, 2001), which reduces the dimensionality of a series of vectors by means of non-linear mathematical transformations. Figure 10 shows a plot obtained with MDS techniques in which the 18 dimensions were reduced to just 2, using a function that aims at preserving most of the variance in the data and minimize the loss of information. Once the words are represented by two (condensed) coordinates, it is easy to visualize them in a bidimensional Cartesian space. Another technique that is often used to visualize the relations between word vectors is by means of cluster analyses then visualized in dendrograms (tree graphs) in which the shorter the arch that connects two words, the higher the similarity between them (for examples, see Chapter 7, Figures 19–20).

In this figure, words that are distributionally similar are close to one another. As I will explain in the next chapter, such similarities, which emerge automatically from corpus occurrences, usually achieve incredibly high levels of correlation with the similarity judgments provided by humans.

6. Distributional semantic models are typically applied to hundreds of words and can use several thousands of contexts. Therefore, the similarities between word vectors are expressed in relation to a multidimensional space with thousands of dimensions.

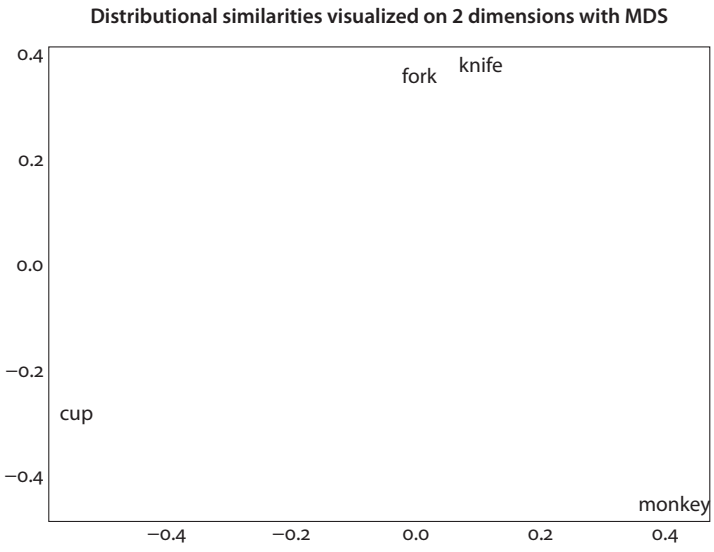


Figure 10. Bidimensional plot showing the cosines for *fork, knife, cup, monkey*. The plot was obtained using the functions *cmdscale* and *plot* (R-core R-core@R-project.org) in the software for statistical analyses R (version 3.5.1). This method returns the best-fitting bidimensional representation for the given dataset, so the configuration returned is given in principal-component axes. For more information, refer to the *cmdscale* documentation in R

5.3 Macro types of distributional models

5.3.1 Structured and unstructured models

A macro distinction between types of distributional models can be made, depending on the type of context that they take into account to construct the word vectors: unstructured and structured models. The first types of models are also called bags-of-words models and are typically based on a context window that extends for some units to the left and/or to the right of the target word. For example, consider the list of concordances displayed in Figure 9, where *fork* appears as the target word. Here, a bag-of-words model with a context window of 3+3 (that is, three words to the left and three to the right of the word *fork*) will take into account as contexts for *fork* all the words (lexemes) that fall into this window, such as *we, enjoys, core, electrical, depressing*, which are all considered to be contexts for the word *fork*. Typically, it is customary to filter out words within the context window that are not semantically informative, such as function words and possibly the most frequent lexemes that appear in the corpus. These words are usually

called stop-words. Needless to say, the size of the context window significantly affects the type of information that can be extracted for a given target word, and therefore the similarity that the word has with other words. Nonetheless, it is hard to establish what would be the optimal window size to capture the collocations of a target word without leaving out important parts of the collocations (i.e., too narrow window), as well as without including noise (i.e., too wide window): the window size is a parameter to be fixed empirically. Despite these shortcomings, bags-of-words models based on context windows are the most popular kind of models used in distributional semantics (especially by non-linguists) because they are extremely simple and fast, and do not require preliminary processing of the corpus. However, such models do not take into account syntactic information and the linguistic structure of specific collocations, which might change in length. For example, a collocation may be more developed to the left side of a target word, or to the right side, it might include one, or two, or three function words, and so on. As Harris's (1954, p. 156) indicated in a pioneering work, "language is not merely a bag of words."

From a theoretical perspective, structured models, which unlike unstructured ones take into account the structures of the specific collocations and allow to harvest linguistic contexts that reflect such structures, are much more precise and theoretically accurate than bags-of-words approaches. Considering the screenshot in Figure 9, for example, a structured model would retrieve collocations such as *stick a*, or *in a toaster*, as collocational contexts for the target word *fork*, and it would know the part of speech of each word involved in the collocations. Despite their appeal, structured models are much more complex to implement than bags-of-words models, because they require the corpus to be lemmatized, parsed and tagged, in order to be able to retrieve the exact collocations, while the bags-of-words approaches simply retrieve the lexemes that fall inside the window size. Moreover, it remains an open and debated question whether structured models provide indeed an advantage in terms of the psychological plausibility of the distributional representations that they construct. So far, it seems that such advantage is highly dependent on the semantic task (see Kiela and Clark, 2014; Lapesa and Evert, 2014; Lenci, 2018).

5.3.2 Explicit and implicit vectors

The terminological differentiation between explicit and implicit models (or better, explicit and implicit vectors) has been recently introduced by Levy and Goldberg (2014) and elaborated by Lenci (2018). With this terminology it is possible to bring onto the table of discussion two problems involved with distributional

models: the high dimensionality and the high sparsity (i.e., many zeroes) of the word-context matrices. Let us explain these two problems and their implications.

As previously mentioned, word frequencies follow the Zipfian law, according to which only a few words are highly frequent, while the vast majority appear only sporadically. This phenomenon has implications for the construction of co-occurrence matrices: for low-frequency words the number of contexts that can be extracted is limited. Conversely, for a few highly frequent words it is possible to extract a high number of contexts. Contexts extracted for each target word in both, structured and unstructured models, are subject to great variability, and therefore scarcely shared by other words. It follows that the matrices in which the co-occurrences are collected are very large and have many cells in which the observed word-context occurrence is actually zero, because the corpus from which the vectors were extracted did not document most of word-context co-occurrences (see for example our hypothetical matrix displayed in Table 1). It remains however unknown in this matrix whether the associations marked as zeroes are due to methodological issues (i.e., the corpus accidentally did not document such co-occurrence) or instead they are *informative* zeroes (i.e., the co-occurrence is semantically and/or syntactically impossible and that is why it is undocumented in the corpus). The difference in the *information* encoded within an associative score that links two words is a crucial aspect that characterizes different theoretical models of learning. In particular, when the simple co-occurrences are taken into account to weight the association between two words, we talk about *associative* models, and their overall functioning can be summarized by the Hebbian principle paraphrased as "what fires together, wires together". Conversely, when we take into account not only the co-occurrence between two items, but also their missed co-occurrence, or the negative feedback, we focus our attention to the actual *information* encoded in the tokens. This is the field of *discriminative* learning. In Section 6.3 I will explain in further detail this difference, which derived directly from the observation of animal behavior and their ability to learn from negative feedback, which led to a revision of the classic notion of Pavlovian conditioning. The difference between simple associative learning and discriminative learning is key to understand the difference between classic distributional models based on co-occurrences and word embeddings based on probabilistic measures of associations (the weights) that are updated with each exposure to a new context, depending on what is expected (based on previous knowledge) and what is observed (based on corpus occurrences).

In distributional models, the high dimensionality and high sparsity of these contingency matrices suggest that many linguistic contexts are not very informative, on average, for the set of target words to be analyzed: a context used zero times with all but one the words does not carry, overall, much information. This

phenomenon has been tackled already in pioneering distributional models such as LSA, by means of algorithms used to reduce the dimensionality of the matrices and keep only highly informative dimensions. In LSA an algorithm called Singular Value Decomposition (SVD) is used, and it consists of a linear algebra technique that divides a matrix into a product of submatrices, allowing one to extract and retain only those that account for the highest amount of variance in the original data. This operation makes the word vectors that represent word meaning significantly shorter and more efficient, even though the algorithm performs the reduction only after having compiled the whole matrix (post-hoc). For LSA, usually the number of dimensions retained is 300, although this number is decided based on intuition and empirical tests, rather than being theoretically motivated.

In recent years, alternative ways have emerged, to construct short and more manageable word vectors that contain only highly informative coordinates from the very beginning of the word meaning construction. The new generation of models goes under the name of **word embeddings**, and they are used to construct low-dimensional vectors to represent word meaning starting from the information derived from the co-occurrence of two words (or a word and a context) as well as the information derived from their missed (but expected) co-occurrence. In these new models, the weights in a word vector are learned in a supervised task, that is, a task where both, the input and the output are known, and the model has to learn the steps in between, which are the weights that construct the word vector. A word embedding algorithm, for example, will be trained to learn the probability for each context to be observed occurring as a context of a given word. By learning these probability measures, the model actually 'embeds' the meaning of a given word in a vector of weights, as displayed in Figure 11 (the hidden layer containing latent features). Word embeddings, typically built with neural networks, are therefore based on probability measures rather than on observed frequencies (e.g., Bengio et al., 2006; Collobert and Weston, 2008; Collobert et al., 2011; Mikolov et al., 2013).

The length of the embedded vector (that is, the number of weights, represented by the number of yellow circles in Figure 11) is set a priori by the analyst. Once this number is set, for example at 300 units, the algorithm will learn their weight using a neural network that will be briefly described in the next section.

5.4 From frequency-based models to word embeddings

Frequency-based models construct the vectors that represent word meanings by *counting* and weighting the linguistic contexts in which words tend to be found.

Conversely, word embeddings construct vector representations by *predicting* which contexts they may be found around these words.[7]

The major pragmatic advantage that word embeddings have over classic frequency-based distributional models is that of relying on a reduced and dense matrix that has only a limited number of dimensions, set by the analyst from the beginning. Conversely, in classic frequency-based models the algorithm needs first to compile a whole large and sparse matrix that takes into account all contexts, and then the matrix can be filtered with some additional methods (e.g., SVD). As such, word embeddings are more easily scalable and can be efficiently used for analyses based on large sample of words: the computations are more efficient, thanks to their construction, which is typically based on neural networks.

Whether word embeddings provide a genuine advantage over classic distributional models is still a hotly debated topic, and the related research is quite inconclusive, with evidence supporting one of the two sides and discarding the opposite view (e.g., Baroni, Dinu and Kruszewski, 2014; Levy, Goldberg and Dagan, 2015; Sahlgren and Lenci, 2016; Mandera, Keuleers and Brysbaert, 2017).

One of the most popular word embeddings used today, word2vec, has been implemented by Mikolov and colleagues (2013) at Google. As soon as it was released, word2vec quickly became the dominant approach for vectorizing linguistic data and it still is, today, widely used. This model, similarly to other word embeddings, is based on a simple neural network.

Word2vec (in its SkipGram version) works as follows. Starting from a set of words, like *fork, spoon, cup* and *monkey*, used as input of the neural network, the algorithm gives as output the probability for each item in the list to be found together with each other word. To do so, the algorithm first transforms each of the 4 words into numbers, to be read by a machine (because computers cannot read words *per se*, but they can read numbers). The way in which these 4 words are translated into numbers is by creating 'one-hot' vectors of 4 coordinates each for each of the 4 words, like a word identifier in which we display a 1 if the word is present, and 0 if it is absent. Considering the order in which the 4 words were presented above, we will have *fork* (1,0,0,0), *spoon* (0,1,0,0), *cup* (0,0,1,0), and *monkey* (0,0,0,1). Comparing these vectors does not make much sense because they do not share any information, besides the fact that none of them is identical to another. Therefore, the vectors are all equally distant from one another. In order to make the vectors comparable, each of them must contain information about the

7. I am here referring to some types of word embeddings, while others do the opposite: given a series of contexts with a missing word (a blank) they try to infer the missing word. In particular, continuous bag of words or CBOW, predict word tokens from its contexts, while in skipgram a given word token is used to predict words in its context.

distributional properties of each word. To do this, the analyst needs to establish an arbitrary number of dimensions on which the distributional meaning of each word will be constructed. This is the number of latent features included in the hidden layer of the (shallow) neural network. The hidden layer is constructed as a matrix in which each of the 4 words is displayed on a row and the latent features will be on the columns, precisely as the contingency matrices are built in classic distributional models. Each row of the matrix, then, will be the vector that represents each word. These vectors are updated, starting from the initial uninformative one-hot vectors, every time the algorithm encounters a co-occurrence of two words (of the 4 words in our small lexicon).[8]

The most important aspect to understand is that the word vectors are constructed in the embedding (the hidden layer) and their length can be set by the analyst. Moreover, the vector coordinates are measures of probability, rather than co-occurrence counts like in classic distributional models. These probabilities, which express the likelihood of observing each two words in the lexicon appearing in the same context, are learned by taking into account their observed co-occurrences (extracted from corpora) as well as their missed co-occurrences. Relying on the information extracted from positive feedback (co-occurrence) as well as on negative feedback (expected but missed co-occurrence) word embeddings, as well as classic distributional models, learn word meanings on the basis of mechanisms that are directly analogous to well-known animal learning models, as I will describe in the next chapter.

Mikolov shows that these vectors can model semantic analogies such as "*woman* is to *x* as *man* is to *king*" and can therefore be used to model some aspects of reasoning. In this example, by subtracting[9] the vector (*man*) from the vector (*king*), and adding to the result the vector (*woman*), we obtain a vector that is very close to the vector of *queen*. That is, by taking off masculinity and adding femininity to the meaning of the word *king*, we obtain the meaning of the word *queen*. This type of operation, as well as the nature of the semantic representations that are used to embed the word vectors do not differ much from the lexical representations and the related operations between word meanings that have been proposed in classic lexical semantic theories, according to which *man* is [+human], [+male], *woman* is [+human], [−male], and so on (e.g., Leech, 1974).

8. A clear and dynamic visualization of the implementation of word embeddings is provided by Xin Rong, and can be found online at the following url: https://ronxin.github.io/wevi/

9. This is done by literally subtracting the value of each specific coordinate in a vector to the corresponding coordinate another vector.

The difference in which the word vector is constructed in a classic frequency-based model and in a word embedding is summarized in Figure 11.[10]

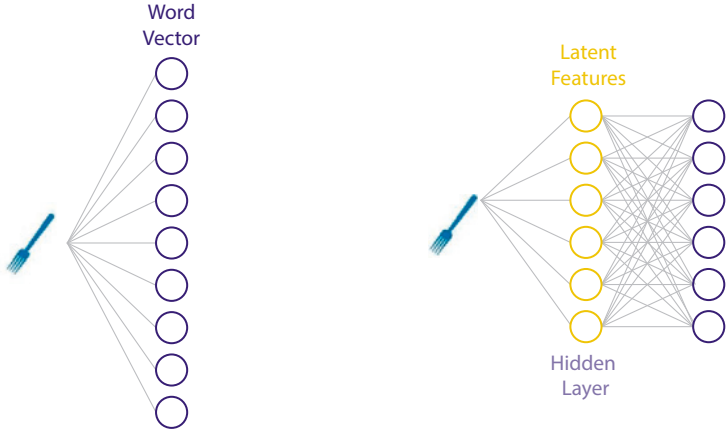


Figure 11. The way in which a distributional model and an embedding model construct word meaning. For a more detailed and interactive visualization of how word2vec construct word vectors, see https://ronxin.github.io/wevi/

The word vector obtained with a classic count-based model (on the left), before the dimensionality reduction techniques are applied, consists of a list of association scores between *fork* and each of the contexts (the blue circles) with which *fork* appears. The strength of the connection between a word and each of the contexts can be weighted by a measure of association, which takes into account both, the co-occurrence of word and context, as well as the occurrence of word alone and context alone (e.g., mutual information metrics). The word vector obtained with an algorithm like SkipGram (on the right) is obtained by embedding the word representation in a hidden layer (using a shallow neural network), and consists of a list of probability measures between *fork* and the predicted contexts. The interpretation of the information encoded in the hidden layer, the embedding, can be postprocessed for example by applying geometrical transformations (rotations) to the matrices (e.g., Rothe, Ebert and Schütze, 2016; Park, Bak and Oh, 2017; Dufter and Schütze, 2019).

The literature on word embeddings is piling up at a very fast pace, with seminal works about word2vec (Mikolov, Chen, Corrado and Dean, 2013) reaching almost 20K citations after just 7 years from publication. A thorough and up-to-date review of the most recent models proposed thereafter lies beyond the scope of this book. Nevertheless, it shall be mentioned that the year 2018 is considered

10. This illustration is conceptual, and therefore simplified. For example, it does not display the one-hot vector used as input for *fork*, as described in the next paragraphs.

to be particularly important for the progress in nlp and in the implementation of word embeddings. This is because in 2018 Google released BERT,[11] a particularly versatile model that can be used, among other tasks, to construct vectors to represent word meaning and sentence meaning. The advantage offered by BERT over models like word2vec is that while in word2vec each word meaning is represented by a fixed vector regardless of the instance of context within which the word appears, in BERT word representations are dynamically constructed in each context and informed by the words around them (McCormick and Ryan, 2019). Therefore, BERT can for example disambiguate polysemous words, retrieving information from the specific context in which a word is analyzed for a specific task. For other tasks, however, this dynamicity of BERT over word2vec becomes problematic. For example, when comparing word vectors to model the similarity between word meanings (the switch between the syntagmatic and the paradigmatic level implemented through cosine measures) BERT embeddings make comparisons between word meanings less stable and thus less valuable (McCormick and Ryan, 2019). This is precisely due to the fact that a word meaning represented by a BERT embedding is fully contextually dependent and therefore the word meaning always changes, depending on the context in which the word is used.

To conclude, both, classic vector spaces constructed on the basis of word occurrences across corpora of texts, as well as the more recent word embeddings constructed using neural networks, represent word meaning by looking at how words appear in context with other words, and construct vectors on the basis of such observations. The two types of models simply differ in the way these vectors are constructed. Once word vectors are constructed, the paradigmatic similarity between words (in both types of models) can be computed in the same way for the two types of models. Such paradigmatic relation of similarity between words, which is a building block of human mental lexicon, as well as the output of distributional models and word embeddings, is based on a switch between the syntagmatic and the paradigmatic level of analysis, which constitutes the core tenet of the distributional hypothesis.

5.5 Summary

In this chapter classic and more recent ways in which 'the artificial mind' constructs and represents word meaning are explained. The symbolic (vectorial) format is particularly useful in computer sciences, because it can be read and processed automatically by machines. Computers read and process numbers, and if

11. The acronym stands for Bidirectional Encoder Representations from Transformers.

natural language is represented by way of numbers, a computer can understand, learn and simulate word meaning structure and processing, possibly by mirroring the functioning of the human mind and the human brain.

The models that I described in this chapter are all based on the assumption that the meaning of a word can be represented in terms of the relations that it entertains with other words. In this sense, words acquire meaning via other words.

In computer sciences and computational linguistics this intuition was implemented by transforming word meaning into numerical vectors, that is: lists of numbers, each standing for a weighted association between a word and one of its contexts of use. Such vectorial representations are based on the so-called distributional hypothesis, according to which word meaning is distributed across syntagmatic relations that words entertain with other words (but, as we will see in Chapter 7, the same principle can be applied to extra-linguistic contexts). Based on such syntagmatic relations, it can be inferred that words that tend to entertain similar patterns of co-occurrences (i.e., words that tend to appear in similar contexts) have similar meanings, paradigmatically speaking.

The best way to formalize word meaning by means of distributed representations encoded in vectors is a hotly debated issue and in the past two decades has generated a wide variety of distributional models that function in slightly different ways. The most recent ones, called word embeddings, are neural networks and are becoming extremely influential and widespread inside and outside academia, for their extreme flexibility, scalability and efficiency. These models are based on the same theoretical assumptions that characterize the classic distributional models which are more extensively described here, because of the higher transparency of their functioning.

In the next chapter, I will explain how the outputs produced by these models are typically evaluated against psychological data elicited from humans. I will also explain what are the theoretical implications involved in trying to compare the outputs of such models to human behaviour, and the challenges that these computational models have encountered (and overcome) in the past decades, when their cognitive plausibility was proposed within the cognitive science community.

CHAPTER 6

Evaluating distributional models

6.1 Evaluating distributional models against psychological data

Distributional models are typically evaluated on their ability to produce outputs that mirror human performance. Because these models aim at representing word meaning and the relations between word meanings, the typical tasks in which their performance is evaluated against human performance are semantic tasks, such as synonym detection and semantic categorization.

One of the most widely used benchmarks of human data against which the performance of distributional models is typically evaluated is the TOEFL synonym detection task. This task, used to test language proficiency in learners of English as a second language, was first used as an evaluation benchmark of LSA by Landauer and Dumais (1997). The TOEFL synonym detection task consists of 80 multiple-choice questions, in which an English target word (which can be a noun, a verb, an adjective or an adverb) is followed by four possible options, among which only one is a synonym of the target word. English language learners have to identify the correct synonym, thus proving their semantic knowledge of English words. For example, given the target word *prominent*, the task asks: 'Which of the following is closest in meaning to *prominent*: *battered, ancient, mysterious* or *conspicuous*?'. The task is quite hard for language learners, because the relation of synonymy is fairly intuitive rather than fully rational, and its definition is quite debated also among linguists, up to the point that some lexical semanticists have concluded that true synonymy does not actually exist (e.g., Cruse, 1986).

Synonymy is typically operationalized in terms of substitutivity: if two words can replace one another in various contexts, then the two words are synonyms. The problem is that synonymy is not a categorical property (two words are synonyms or not). Synonymy is rather a gradual property of words: *the more* two words can replace one another across contexts, *the more* two words are synonyms. In fact, there isn't a threshold number of contexts in which two words can be replaced, that grants the two words the status of synonyms. For language learners taking the TOEFL exam, being able to determine which word among four possible candidates can replace a target word in more contexts implies that the learners must know virtually *all* the contexts of use of *each* of the words used in the test in order

to make a correct guess. In other words, it implies that the language learner knows the distributional properties of each of the words involved in the test. Moreover, the criteria that have been chosen in the TOEFL test to select the non-synonyms vary across trials: in some cases, the non-synonyms are semantically related to the target word, while in other cases they are completely unrelated words.

Landauer and Dumais (1997) reported that language learners on average provide the correct answers to the synonym detection task in 64.50% of the trials. In a subsequent study, Rapp (2004) administered the test to a group of language learners and a group of native speakers at Macquarie University, and reported that learners provided correct answers in 86.85% of the cases, while native speakers reached the 97.75% of accuracy. Various distributional models have been evaluated against these percentages, starting from Landauer and Dumais (LSA accuracy: 64.38%). Other authors who tested the performance of their distributional models reached even higher percentages of accuracy: Kalgren and Sahlgren (2001) reported a 72.50% accuracy for the Random Indexing model; Padó and Lapata, (2007) reported a 73.00% accuracy for the Dependency Space model; Turney (2001) reported a 73.75% accuracy for the PMI-IR (Pointwise Mutual Information and Information Retrieval) model; Baroni and Lenci (2010) reported a 76.9% and a 75% accuracy measures for the two variants based on the Distributional Memory model; Bullinaria and Levy (2007) reported a 85% accuracy for a HAL-type of model (Hyperspace Analogue to Language) based on the Positive Pointwise Mutual Information measure of association, and extended their reported levels of accuracy to 100% when various parameters such as the application of stop-lists, word stemming, and dimensionality reduction using Singular Value Decomposition (SVD), were finely tuned (Bullinaria and Levy, 2012). Overall, distributional models perform so well in synonym detection tasks, that they often outperform (or provide comparable outputs as) intermediate and fluent non-native human speakers.

An alternative way in which distributional models are typically evaluated against human judgment is by measuring the correlation coefficients between similarity judgments provided by humans and similarity scores constructed by the model. Similarity judgments provided by humans are typically collected by asking participants to rate the similarity between two words on a numeric scale (e.g., Rubenstein and Goodenough, 1965; Finkelstein et al., 2001; Hill, Reichart and Korhonen, 2015). Different datasets of human ratings (such as those indicated above), however, used slightly different instructions for the participants to the task. As a result, participants tended to favor slightly different types of similarity between words. Moreover, even within the same dataset, various semantic relations can be identified for those word pairs that have been rated as highly similar by the participants. Baroni and Lenci (2011), for example, point out that in some cases high similarity scores are attributed to synonyms (e.g., *journey*/*voyage*) and

in other cases to co-hyponyms (e.g., *king*/*queen*). If we look at word pairs with similarity scores around the median value of the scale (e.g., on a 5-point scale, word pairs with similarity scores around 3) the variety of semantic relations between the words involved increases dramatically. Nonetheless, correlation coefficients between distributional models and human ratings can reach high values, for some cases even around 0.80.

Given the theoretical and methodological challenges involved in these evaluation techniques, as Baroni and Lenci (2011, p. 2) pointed out, "perhaps a more principled way to evaluate DSMs[1] that has recently gained some popularity is the concept categorization task, where a DSM has to cluster a set of nouns expressing basic-level concepts into gold standard categories". In 2008 a competition was launched involving a shared categorization task in which various distributional models were invited to test their ability to perform a categorization task on the basis of a shared resource (Baroni, Evert and Lenci, 2008): a dataset of 44 concrete nouns belonging to 6 semantic classes, extracted from the dataset of concrete nouns used by McRae and colleagues (2005). The nouns included a variety of types, including natural categories such as birds (*chicken, eagle, duck, swan, owl* etc.) and artifacts such as vehicles (*boat, car, ship, truck, rocket* etc.). The categorization task performed by the distributional models was evaluated on a cluster analyses, in which the purity of the clusters automatically detected by various distributional models was compared. In this task, various models performed very well, thus showing that the semantic information that allows humans to categorize similar concepts together into semantically coherent classes, is captured by these computational models which are based exclusively on the information encoded in linguistic co-occurrences.

Finally, another way used to evaluate the performance of distributional models against human behavior is to investigate whether distributional models can predict the semantic priming effect observed in human participants. This type of evaluation technique is particularly interesting because it taps into a semantic effect observed in human behavior by means of direct measurements of online cognitive processes, such as the measurement of reaction times in lexical decision tasks, rather than on the measurement of offline cognitive processes, such as categorization, similarity judgment and synonym detection. In addressing this type of evaluation, some studies have looked at the words that are used in psycholinguistic experiments testing the priming effect, and investigated whether there was a significant difference in the similarity scores obtained by the distributional model between the word pairs that generate a priming effect in human participants and those that do not, finding significant differences (e.g., McDonald and Brew, 2004;

1. This abbreviation stands for Distributional Semantic Models.

Padó and Lapata, 2007). Other studies used distributional models to model the semantic priming effect by means of regression analyses (see Mandera, Keuleers and Brysbaert, 2017 for a review of these studies).

To conclude, the fact that the evaluation of distributional models is typically done against psychological data retrieved from human participants in different types of linguistic and cognitive tasks suggests that the outputs of such computational models are comparable to those resulting from human behavior. But is the cognitive plausibility of their functioning legitimate? In other words: to what extent it is valid to infer that because the outputs generated by these computational models correlate with human performance, then also the computational operations implemented by the models can be equaled to the cognitive processes that allow humans to deliver such performances? This big question has been discussed among scientific communities, as described in the coming sections.

6.2 Learning associations by conditioning

LSA, the pioneering and most widely used distributional model of word meaning, has been introduced by its authors as "a new general theory of acquired similarity and knowledge representation" (Landauer and Dumais, 1997, p. 211). In their paper, the authors explain that their model starts with the observation of words and contexts in which words are used. The authors compare the linguistic contexts to episodes which, in principle, mirror actual human experiences. The associative mechanism between words and (linguistic) contexts in which the words are used, or words and episodes as Landauer and Dumais argue, is not alien to the cognitive science community. In fact, this idea is based on a cornerstone principle of learning called Classical Conditioning effect, or Pavlovian Conditioning, observed in both, animal and human behavior. The history of its discovery begins with the Russian scientist Ivan Pavlov and his dogs, in the first years of the XX century.

Pavlov observed that when dogs were repeatedly exposed to a situation in which the presentation of food was preceded by a ringing bell, after a number of exposures they would start salivating just by hearing the bell, thus showing that they constructed a strong association between the two events that occurred repeatedly together: the ringing bell and the presentation of food. This phenomenon, labelled as Classical Conditioning can be formalized as follows: when a strong stimulus (e.g., food) is paired with a previously neutral stimulus (e.g., a ringing bell), the repeated association results in a pairing of the two stimuli. As a consequence of the pairing, the neutral stimulus (e.g., the ringing bell) generates a behavioral response (e.g., salivation) that was previously contingent only on the strong stimulus (e.g., food).

In the distributional model LSA, the repeated co-occurrence between a word and a text (which in the authors' explanation stands for an episode) results in an association between them. Properties of the text (or of the episode) become part of the meaning of the word. The word therefore acquires meaning from the context in which it is used, thanks to its repeated co-occurrence with contextual elements. To make an analogy with the example described above (Pavlov's dogs): a word from (i.e., a bell) that repeatedly co-occurs with a context (i.e., food) acquires meaning (i.e., the salivation response) from it.

The basic associative principle according to which two entities that repeatedly appear together become associated is also a fundamental principle of the so-called Hebbian learning, in which the simultaneous activation of cells leads to pronounced increases in synaptic strength between those cells. Moreover, this principle is to some extent similar to the principle that characterizes the learning process in neural (both biological as well as artificial) networks. In fact, Landauer and Dumais suggested that, conceptually, the LSA model can be viewed as a simple neural network although they do not elaborate in detail how the implementation of a neural network could evolve from LSA.

The story about Pavlovian conditioning described above, however, is far from being complete. In order to fully understand and appreciate the power and the limitations of Hebbian learning, and of the Conditioning phenomenon described in the way in which Pavlov initially reported it, let us elaborate in further detail the subsequent discoveries on animal and human behavior which led to a revision and extension of the notion of conditioning. This further in-depth explanation is key to understand how the associations between two words or a word and a referent are weighted (that is, strengthened or weakened), episode after episode, in the human and in the artificial mind.

6.3 Associative and discriminative learning

Various psychological experiments conducted after Pavlov's seminal experiments demonstrated that we do not learn associations from the simple co-occurrence of two items (as Pavlovian conditioning would predict) but we learn associations thanks to complex dynamics in which the co-occurrence of two items is balanced with their missed co-occurrence (e.g., Rescorla, 1988).

To grasp the importance of the negative feedback in learning associations, consider the following contexts, in which you can comfortably wear sneakers: informal dinner with friends, shopping for groceries, jogging at the park. Now consider a different type of footwear: flip flops. As for sneakers, flip flops can be worn at an informal dinner with friends and when shopping for groceries. However,

while sneakers are also associated with jogging at the park, flip flops are arguably not. The missed association between flip flops and jogging at the park, however, is an informative zero: it *has* meaning. This missed co-occurrence between flip flops and jogging at the park informs us about the meaning of flip flops and about their degree of similarity/difference with sneakers.

Let us know turn back to the empirical studies on conditioning. The seminal works by Pavlov do not take into account negative feedback (the missed co-occurrence), but only the simple co-occurrence of two events: a neutral and an unconditioned stimulus (such as the ringing bell and the presentation of food). From the repeated co-occurrence, and therefore from the *contiguity* of these two stimuli, the neutral one (the bell) becomes conditioned, and produces the same response (e.g., salivation) of the unconditioned stimulus (food). Thanks to the repeated co-occurrence of the two stimuli (the bell and the food) the first starts to cause the typical response given by the latter: dogs start salivating when they simply hear the bell ringing. In this perspective, *contiguity* (between the two stimuli, bell and food) is a key factor for learning the association between them.

However, in the Eighties, Rescorla (1988) argued, based on empirical evidence, that the mere repeated co-occurrence of two events does not entail the establishment of an association between them: contiguity alone is not enough. In one experiment Rescorla exposed rats to two different scenarios. In the first, a mild electric shock was delivered right after the sound of a bell. In the second scenario, the mild electric shocks and the bell sounds occurred in an uncorrelated manner, sometimes together and sometimes alone. While in the first condition rats learned the association between the two stimuli, as in the seminal Pavlovian experiment, in the second condition they did not, because the two stimuli (the bell and the shock) could occur together but also alone, with no predictable pattern. Based on this result, Rescorla explained that conditioning is not a raw reflex learned automatically by means of contiguity alone, but is instead an operation that derives from the evaluation of the information contained in the occurrences. The information learned by the rats comes from the co-occurrence of the two stimuli as well as from their individual occurrence alone, and therefore from the missed co-occurrences. In this updated view, conditioning can be defined as learning relations across events, based on the information carried by the stimuli rather than simply by their co-occurrence. Moreover, as further argued by Rescorla on the basis of previous studies, not all stimuli are equally associable: some types of stimuli tend to be more easily associated to one another than others (e.g., Garcia and Koelling 1966). Finally, the blocking effect demonstrated by Kamin (1968) a couple of decades earlier provided additional elements for a thorough revision of the notion of conditioning. In Kamin's experiments two groups of animals received a compound stimulus (a light + a bell) followed by an unconditioned

stimulus (food). Both groups were tested for their conditioning of one of the two stimuli in the compound, for example, whether the bell alone produces salivation. The difference between the two groups, however, was that the first group had a history of the light alone signaling the arrival of the food, whereas the second group lacked that history. Therefore, while both groups shared the experience of the co-occurrence between the compound light + bell and the presentation of food, one group had the additional information of the light alone being associated with the presentation of food. For this group, therefore, the bell was actually redundant. Results showed that the bell became strongly associated with food in the group with no previous information about light, but only weakly associated with food in the group that had previously established an association between light and food. This effect is referred to as blocking effect, because the previous association between light and food blocks the association between the bell and the food. Rescorla used these results to argue, once again, that the associative learning mechanism is not governed by simple contiguity but rather by the informational relation on which two paired stimuli differ, and by the previous knowledge associated with them.

Taken all together, these results show that conditioning is a rich mechanism based on the information carried by the stimuli, the circumstances in which they can be found, together or alone, and previous knowledge about them. Conditioning is influenced by the contingency between two entities (their causal relation), rather than by their simple contiguity (their correlation). From the observation of this phenomena in animal behavior, Rescorla developed a theory and a model of learning, based on previous studies (Rescorla and Wagner, 1972) capable of taking into account positive and negative feedback (what is observed) and previous knowledge, to update associations between two entities. The theoretical account proposed by Rescorla emphasizes the importance of the discrepancy between the actual state of the world (what is observed) and the organism's representation of that state (what is expected). On the basis of the 'surprise' derived from the mismatch between observed and expected outcomes, organisms adjust their associations and therefore learn new information. In this sense, learning is a process informed by the mismatch between expectations based on previous experience and new incoming information.

The change in associative strength between two stimuli, due to a new exposure to an episode, is formalized algorithmically in the Rescorla-Wagner model by means of a mathematical equation, which is the same equation used in many neural networks (e.g., Baayen, 2010; Hollis, 2019) to adjust the weights (the associations) between nodes. In neural networks this is called Delta rule (Rosenblatt, 1957), and it is identified by the Greek letter Δ. The Delta rule suggests that the association between two items is determined by the existing associative strength between them and the associative strength of all the stimuli present. The change in

association strength observed at a subsequent moment, called Delta value, can be zero (if in the previous moment the association was zero as well), can be a positive value that increases the association strength between the two items if an occurrence between two stimuli is observed, or it can be a negative number that decreases the new association strength if an expected occurrence between the two stimuli is not observed. Therefore, experience after experience, and exposure after exposure, the strength of the association between two stimuli can change dynamically, increasing or decreasing, depending on the observed or non-observed co-occurrence of two items that were expected to occur together, based on previous knowledge.

Discriminative learning, or learning associations from negative feedback, as well as from positive feedback, is a phenomenon that characterizes human learning processes as well as animal ones, but it is also a phenomenon that puzzled cognitive and computational scholars alike. From a cognitive perspective, it is clear that not *all* the missing associations between two items are equally informative, and it is quite likely that most of them, which are irrelevant for the stimulus, are not taken into account to learn and update associations. For example, while the missing association between flip flops and jogging at the park is relevant for the meaning of flip flops, the missing association between flip flops and baking a cake, or burning your hand, or going to the barber shop, are arguably uninformative about the meaning of flip flops, because there isn't any causal connection that can be established between the entity flip flops and the missed co-occurrence with the events described. Similarly, from a computational perspective, taking into account each and every missed association between two entities (or two words) is a cumbersome and unneeded operation that slows down significantly the performance of a computational model. Within the Rescorla-Wagner model, however, the learner (human, animal or machine) is supposed to update all the associations in their lexicon, every time a new input is taken in. There is no way to distinguish between knowledge that is relevant and knowledge that is not relevant to a given meaning, and therefore between associations that need to be updated and associations that do not need to be updated, after a specific episode. As Hollis (2019, p. 1418) recently pointed out:

> Learning to delimit between what is relevant and what is irrelevant is a nontrivial problem and bears resemblance to the philosophical frame problem: In a sufficiently rich environment, there is no tractably identifiable boundary between (1) knowledge that is relevant to a particular context, and thus needs to be updated through learning, and (2) knowledge that is irrelevant to a particular context, and thus can be left alone.

Various models of the human and the artificial minds have been proposed to account for negative feedback, suggesting different ways in which missed associations

shall be considered (e.g., Baayen, Hendrix and Ramscar, 2013; Hollis, 2019). A thorough review of these models lies beyond the scope of this book, but in Part 3, when we discuss in further detail the associative mechanism in relation to cross-situational learning, we will see how the results of empirical studies on human and artificial behavior progress in parallel, shedding light onto one another to achieve the common goal of learning how associations are learned. What is important to underline here, is that there are various ways to engineer the associations between words or between words and referents (the first of 3 steps described in Chapter 2) by taking into account (or not) negative feedback (see Cassani, Grimm, Gillis and Daelemans, 2016 for a review).

In their seminal study, Landauer and Dumais (1997) realized that the bottleneck of the argument that sees LSA as a psychologically plausible theory of meaning is the cognitive plausibility of the *transformations* that follow the creation of the word-context matrix. Notably: the backbone SVD algorithm that condenses the dimensions of the matrix (see Section 5.3.2), and tf-idf (term frequency-inverse document frequency), the equation used to weight the strength of association between a word and a document in which it occurs. These transformations consist of mathematical operations that can be hardly mapped in a one-to-one manner to specific cognitive processes, even though the authors insist on arguing that some correspondences can be established (Landauer and Dumais 1997, p. 219).

6.4 Grounded and ungrounded symbols

The heated debate on whether distributional models such as LSA may constitute a psychologically plausible theory of meaning is well described in an edited collection extracted from an academic event in which supporters of opposing views took part (De Vega, Glenberg and Graesser, 2008). The editors of this project framed the topic within the search for the nature of the symbols on which cognitive processing is based. Among such symbols, a specific type consists of words, which stand for the concepts that we construct in our minds and denote the designated referents in the world. The debate therefore pertains to the definition of the nature of word meaning too. The main question can be asked in the following terms: to what extent are words, and more specifically the processing of word meaning, grounded in the human sensorial and motoric cognitive systems?

This very general question sees supporters of the grounded and embodied accounts of cognition opposing supporters of the amodal accounts of cognition. As De Vega, Glenberg, and Graesser pointed out:

> linguistic symbols are embodied to the extent that: (a) the meaning of the symbol […] depends on activity in systems also used for perception, action, and emotion and (b) reasoning about meaning […] requires use of those systems.
>
> (De Vega, Glenberg, and Graesser, 2008, p. 4)

Within this framework LSA is a theory of meaning that relies on *non-embodied* (also called *amodal)* symbols because words in LSA acquire their meaning from their relations with other words, and not from modality-specific (i.e., mainly sensorimotor) neural configurations. However, Landauer and Dumais predicted this possible critique and suggest that perceptual information extracted from experiential contexts and actual perceptual experiences may be integrated in the language-based representations of word meaning delivered by LSA, specifically in the matrix from which the words' vectors are constructed (1997, p. 227). In Chapter 7 I will explain how this point has been implemented in more recent years, in multimodal distributional models.

Supporters of the embodied theories of meaning and of the grounded nature of word processing pointed out that LSA simply does not satisfactorily explain how words get their meanings, invoking the so-called *symbol grounding problem,* summarized by Harnad in the following terms:

> How can the semantic interpretation of a formal symbol system be made intrinsic to the system, rather than just parasitic on the meanings in our heads? How can the meanings of the meaningless symbol tokens, manipulated solely on the basis of their (arbitrary) shapes, be grounded in anything but other meaningless symbols?. (Harnad 1990, p. 335)

These questions evoke an argument that was raised in 1980 by American philosopher John Searle by means of an analogy generally known as the Chinese Room Argument (Searle, 1980). Searle imagines himself alone in a room, receiving messages written with Chinese characters slipped under the door. Searle understands nothing of Chinese, and yet, by manipulating the Chinese symbols on the basis of their syntax alone he is able to generate replies to these messages, in the shape of appropriate strings of Chinese characters that fool those outside the room into thinking there is a Chinese speaker inside. Searle uses this analogy to make an argument against the Strong AI view, according to which a computer program able to manipulate symbols on the basis of their simple syntagmatic relations is also capable of *understanding* the meaning of such symbols. According to Searle, in fact, the accuracy of the output (the Chinese strings in the hypothetical scenario of the Chinese room, or the outputs generated by a computer that processes and combines words) may fool the analysts into thinking that the agent (the human in the room or a computer program) understands and masters the language it produces while this is not the case. Similarly, LSA or any distributional model, according to this view, cannot be considered

a psychologically plausible theory of meaning because they represent word meaning (only) on the basis of other words, which in turn gain their meaning through other words; such a circular process does not allow conceptual grounding to take place. Conceptual grounding determines, according to Searle, how words mean.

While the Chinese room argument highlights the weaknesses of a Strong AI approach in which word meanings are constructed solely on the basis of word statistics and word co-occurrences with other words, supporters of embodied theories who use this argument against distributional models of meaning seem to imply that the truth lies in a strong embodiment view, according to which word meaning and its related conceptual content is represented *entirely* in terms of sensorimotor information and computations over sensorimotor content and information that we retrieve from perceptual experiences. This strong embodiment view, however, has been heavily criticized as well (e.g., Mahon and Caramazza, 2008; Mahon, 2015) and is not always backed up by empirical evidence: in many situations and for many different types of speakers the processing of word meaning does not or cannot rely on perceptual information (see Mahon 2015 for a review). It follows that when language processing does not involve the activation of information retrieved from perceptual, motoric and emotional experiences, it must rely on the activation of information retrieved from other words, from language itself. It remains an open question that of describing and testing the factors that determine in which circumstances and for which speakers is word meaning processed by means of the activation of sensorimotor simulations (and therefore grounded in perception, action and emotion by means of word-to-world linkages) and in which cases is word meaning processed symbolically, without relying on such simulations but simply on information retrieved from language and from word co-occurrences with other words (word-to-word linkages). An interesting case that shows differences in the way participants rely on word-to-word relations, is the comparison between native speakers and foreign language learners.

6.5 Word meaning in native speakers, language learners, and distributional models

Native speakers and language learners seem to construct and process word meaning (respectively in the L1 and the L2) in slightly different ways, with language learners relying more than native speakers on linguistic structures, lexical links rather than conceptual ones, and syntagmatic relations between words. This observation, together with the arguments raised above suggesting that word meaning (at least in native speakers) consists of information that we retrieve from embodied experiences, leads to the formulation of the following research question: Is it

possible that distributional models based on simple word co-occurrences capture linguistic information only, and therefore mirror with more accuracy the mental lexicon of language learners, than the mental lexicon of native speakers?

This question was addressed in a couple of empirical investigations (Bolognesi, 2011; Bolognesi, 2016a) in which word similarities obtained from a structured distributional model (Distributional Memory, Baroni and Lenci, 2010) were compared with similarity judgments elicited from English native speakers and English foreign learners, and replicated for Italian native speakers and Italian foreign learners. Let us summarize and discuss the findings obtained on the English data.

The studies were conducted on a sample of 48 verbs divided into two classes: 24 motion verbs (e.g., *run, jump, skip*) and 24 verbs designating mental operations (e.g., *judge, appreciate, decide*). Verbs were preferred to nouns because of their relational nature (see Gentner, 1978), proven also by the fact that in free association tasks verbs tend to generate associations that denote the verb arguments (Guida and Lenci, 2007, p. 18). This makes verbs good candidates to be represented in a distributional model. Motion verbs in particular tend to be used in figurative constructions, and are therefore semantically richer (i.e., more polysemous) than verbs denoting mental operations. Consider, for example, the different types of meaning conflated in the meaning of the verb *follow* reported in the examples in (10)–(13), based on Talmy (2000).

(10) *The policeman follows the thief.*

(11) *I follow my instinct.*

(12) *He follows the footsteps of his father.*

(13) *The railway follows the stream of the river.*

In (10), the movement is literal. In (11), it is metaphorical: there is no actual movement involved. In (12) the movement is based on an idiomatic expression that also modulates a figurative meaning of the verb (although different from the previous one), and in (13) it modulates a fictive type of motion that is not metaphorical in the way described by sentences (11) and (12), but also not literal as in (10). The polysemous nature of many motion verbs makes the semantic representation of these verbs quite interesting, because within the same form multiple meanings are possibly conflated. This is for sure the case of the distributional representations that emerge from the computational model, but it could also be the case for the representations that emerge from speakers' judgments of semantic similarity between word pairs: one could argue that if speakers indicate high similarity scores between *follow* and other motion verbs like *run, walk, stroll*, they are arguably thinking of the literal meaning of *follow* (the meaning exemplified in the first sentence), while if they indicate high similarity scores between follow and verbs

denoting mental operations like *doubt, believe* and *understand*, they are probably thinking about figurative meanings of *follow* like those exemplified by sentences (11) and (12). Moreover, research shows that metaphorical meanings may not hold the same status as literal meanings in the mental lexicon (see Werkmann Horvat, Bolognesi and Lahiri, forth., for a literature review), and that metaphors are difficult to understand for non-native speakers, especially in those cases in which there isn't an equivalent expression in the speakers' native language (Littlemore and Low, 2006; Nacey, 2013; Werkmann Horvat, Bolognesi and Kohl, forth.).

Similarity judgments between verb pairs were elicited from American English native speakers ($N = 40$) and Italian native speakers, learners of English ($N = 40$). Notably, many English motion verbs can be used in figurative constructions that do not have direct equivalents in Italian. For example, the verb *run* is often used in collocations such as *run a business, run a school*, etc. (i.e., similar to the meaning of the verb *manage*). In Italian the equivalent verb (*correre*) cannot be used in similar constructions (**correre una scuola*). Similarly, the verb *fall* is often used within the idiomatic expression *fall in love*, while in Italian there isn't such expression (**cadere in amore*). Finally, *drive* is often used within the figurative expression *driving someone crazy*, which does not translate into an equivalent Italian expression (**guidare qualcuno pazzo*).

The structured distributional model Distributional Memory (DM, Baroni and Lenci, 2010) was adopted to create the mental lexicon of the artificial mind, to which those of English native speakers and foreign learners were compared. In typeDM for each verb the context was retrieved as a compound (syntactic link plus semantic collocate) followed by a measure of association (LMI, a derived version of the mutual information).

Two distributional semantic spaces were constructed with the similarity ratings between verb pairs elicited from native speakers and from language learners respectively (Figure 12). The two tables display very similar patterns, with the judgments provided by language learners looking slightly weaker than those provided by native speakers (in the graph, there seem to be on average brighter shades of grey in L2 than in L1). The difference between the average similarity scores provided by language learners and by native speakers is not significant. The two tables of cosine similarities were then reduced to two-dimensional spaces by Multidimensional scaling and plotted (Figure 13, Figure 14). As these two figures show, verbs denoting mental operations tend to cluster automatically on the bottom of the graph, while motion verbs tend to cluster on the top, in both datasets: the scores provided by native speakers and those provided by language learners. For language learners this division between these two verb classes appears to be even more marked than with native speakers (there seems to be more space between the upper and the lower groups of verbs in the plot constructed on L2 data).

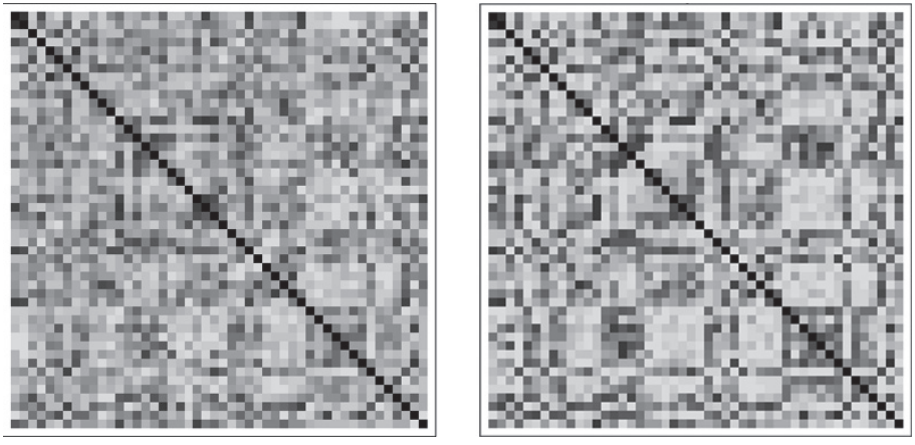


Figure 12. Tables of similarity scores between verb pairs (native speakers on the left, language learners on the right). The 48 verbs are displayed on both, columns and rows, in the same order. The higher the similarity between each two verbs, the darker the little square in their intersection. On the diagonal are displayed the highest values (Cos-Sim = 1), which represent the similarity between a verb and itself, which are therefore in the darkest shade

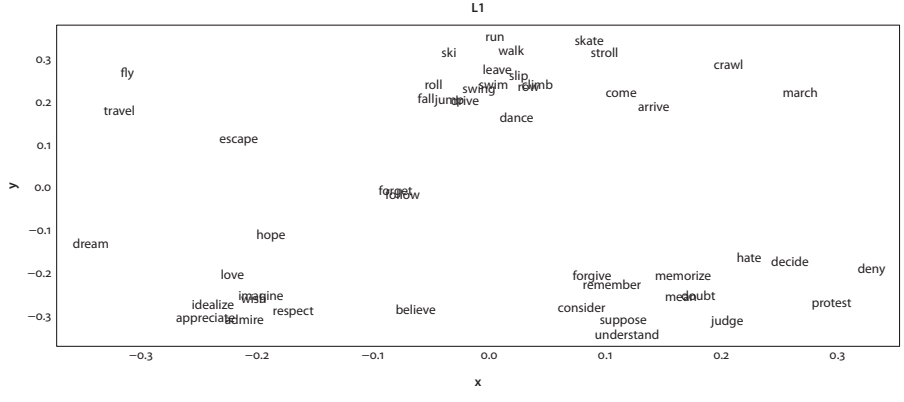


Figure 13. Word map generated on the basis of the similarity scores between word pairs provided by native speakers of English (R version 3.5.1)

The 920,710 triplets retrieved from DM (verb plus syntactic link plus semantic collocate) were very varied: some verbs displayed very high weights for some very frequent collocations (e.g., *fall in love*), others had only very low weights (e.g., *admire beyond measure*). Looking at the 100 triplets with higher association scores for each verb, it became clear that these were following a Zipfian distribution, with a few triplets displaying a very high association score between the verb and the context, and most of them displaying very low weights. Figure 15 displays the triplets with the highest LMI values of the first 10 verbs of the sample.

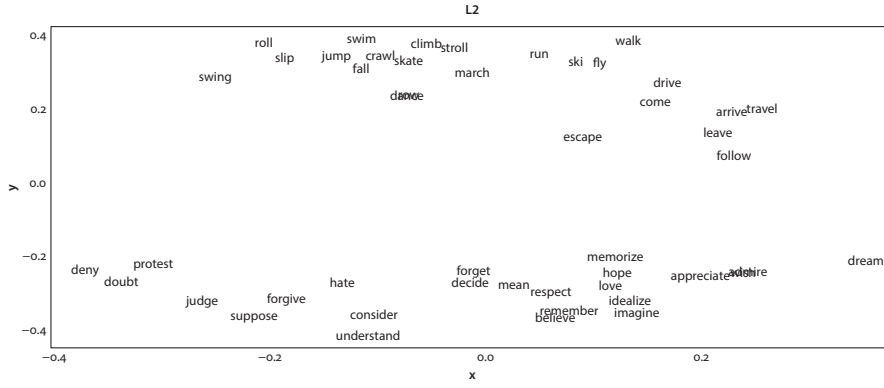


Figure 14. Word map generated on the basis of the similarity scores between word pairs provided by English language learners (Italian native speakers). (R version 3.5.1)

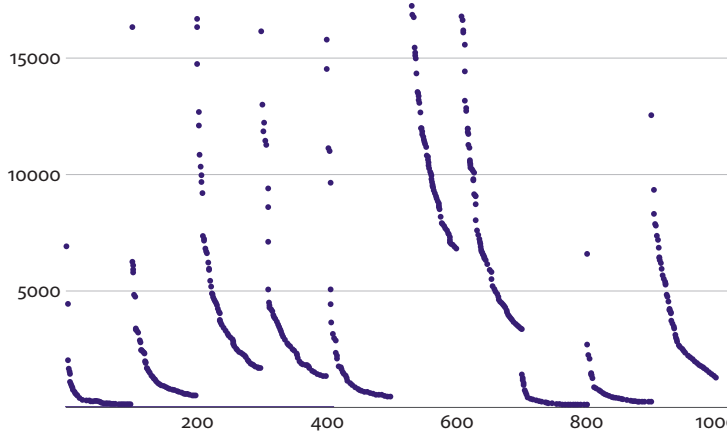


Figure 15. Zipfian distribution of the LMI values for the 100 triplets with higher weights for the first 10 verbs of the list. On the horizontal axis the triplets are represented, on the vertical axis the weights (LMI values)

To construct the contingency matrix 'verbs by contexts', and to avoid having a very sparse matrix with several non-informative weights, only the top 100 contexts for each verb were considered, plus the top 100 triples with highest LMI value overall. On the basis of this matrix the table of cosine similarities between the 48 verbs was constructed (Figure 16) and then displayed in a word map obtained with multidimensional scaling techniques (Figure 17).

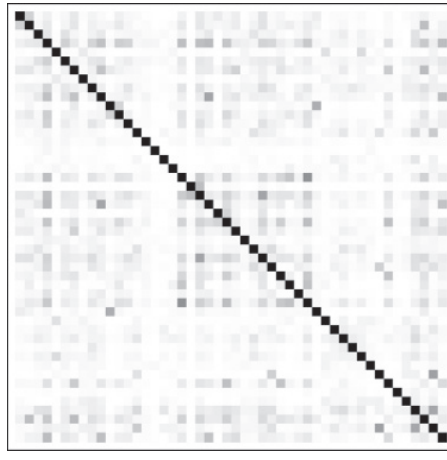


Figure 16. Table of cosine similarities between verb pairs, obtained with Distributional Memory, using the link+word2 as context in the co-occurrence matrix. The 48 verbs are displayed on both, columns and rows. The higher the similarity between each two verbs, the darker the corresponding cell. On the diagonal are displayed the highest values (=1), which represent the similarity between a verb and itself

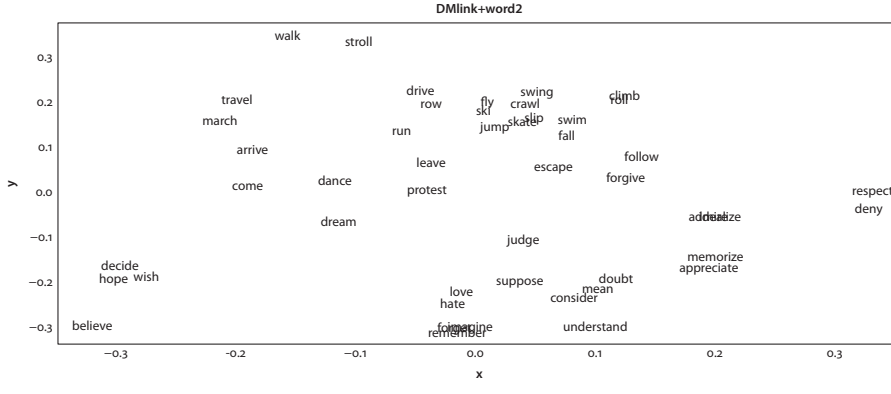


Figure 17. Word map generated on the basis of the similarity scores between word pairs obtained from DM

Finally, by means of a correlation study, the vectors of cosines were compared with one another across the three semantic spaces. In other words, the relative position of each verb was compared across the spaces, to see whether the semantic representations that emerged from corpus data were more correlated to those emerged from native speakers or language learners' judgments. The data was also disaggregated, so that the group of motion verbs and the group of verbs denoting mental operations could be analyzed also independently, to compare the behavior of native speakers and language learners toward the construction of mental

representations of these two verb classes. The results showed the following trends. First, the similarity judgments provided by native speakers and language learners are highly correlated (coefficients above 0.7 for the whole sample of verbs, as well as for the individual subsamples of motion verbs and verbs of thought). This was already observed in a qualitative manner in Figure 12. Second, speakers' judgments (both, L1and L2) are positively correlated with the similarity scores obtained from corpus data (medium/high correlation scores). Third, when the two groups of verbs are considered individually, the data show that on average for native speakers the correlation with distributional data is higher for motion verbs than for verbs denoting mental operations, while this trend is reversed for language learners. This curious trend has been interpreted in relation to the fact that motion verbs are often used in figurative constructions, and the meaning modulated by such figurative constructions is captured by the distributional models very well (the DM triplets with the highest association scores construct figurative meanings of the motion verbs). Conversely, such figurative meanings are not well managed by language learners, who may not take such information into account when providing similarity scores between verb pairs.

Finally, in relation to the general research question that motivated this study, it can be concluded that indeed, verbs denoting mental operations, which are typically used in their literal sense, are represented in the mental lexicon of language learners by semantic representations that correlate better with those emerging from distributional modelling. Conversely, for the motion verbs, which tend to be used metaphorically, the correlations between verb representations in the human and in the artificial minds are on average lower, and in particular they are very low when the distributional semantic space is compared to the similarity judgements provided by foreign language learners.

Besides the specific case of motion verbs which are typically involved in figurative constructions, the semantic representations obtained from a structured distributional model correlate better with the semantic representations in the mind of language learners than with those in the mind of native speakers. This is quite interesting, considering that language learners have arguably *less* information about word meanings, compared to native speakers. The way in which this trend has been interpreted is the following. Language learners, especially those who learned the foreign language in an institutional setting (like the participants to these studies), have less *experiential* information about word meaning, compared to native speakers. The semantic representations of word meaning constructed by language learners, compared to those constructed by native speakers, are more strongly affected by information about word uses in linguistic contexts, and encompass much less information derived from perceptual experiences. Language learners who learn a foreign language in institutional setting typically construct

the semantic representations of word meanings in the target language by means of extensive reading activities, and (as indicated also by Meara, 2009) by mentally translating the foreign words into their L1. These operations enable them to construct semantic representations that are strongly influenced by *linguistic* information, by means of strategies typically based on incidental vocabulary learning (see Chapter 4, Section 4.3). Conversely, native speakers' semantic representations are strongly affected by information retrieved from *perceptual* experiences. The information retrieved from perceptual experiences is particularly important to make sense of the figurative uses of motion verbs, and this can explain the weak correlations between learners and distributional data for motion verbs: the distributional model captures figurative uses easily, while language learners do not, due to their lack of exposure to enough extra-linguistic experiences in the target language. Relying on the linguistic information about word meaning in the foreign language, language learners generate semantic representations that better resemble those emerging from word co-occurrences only for those verbs that can be successfully modelled based on this information, such as verbs denoting mental operations. Conversely, verbs that are typically used in figurative ways are semantically represented by a combination of information retrieved from both linguistic as well as extra-linguistic contexts.

The differentiation between the two streams of information, that is, information retrieved from *language* use and information retrieved from *perceptual experiences*, and their combination in richer and multimodal semantic representations, are described in Chapter 7.

6.6 Summary

Distributional models of word meaning are typically evaluated against data collected from speakers. In this way, such models hint to the idea that, if the outputs that they produce are comparable to human judgments, then also the underlying functioning may be equaled. It follows that distributional models implicitly suggest that word meaning is constructed and represented thanks to referential associations of words with other words (word co-occurrences).

This assumption raised strong critiques from the community of supporters of embodied and grounded theories of cognition, who claim that word meaning cannot rely on symbols (words) that are constructed on the basis of their relation with other symbols of the same type (i.e., other words). This mechanism based on word-to-word reference only, does not account for the origin of word meaning, that is, the word-to-world reference.

The organization of word meanings in the artificial mind, structured on the basis of word co-occurrences, seem to reflect the way in which foreign language learners organize word meanings in L2 in their mental lexicon, relying on (mainly) linguistic occurrences and retrieving meaning from linguistic contexts.

In the next chapter, I will illustrate how the criticism about the lack of perceptual information encoded in distributional models has been taken onboard by various scholars, who worked on the integration of perceptual information within the linguistic representations of word meaning in the artificial mind, to overcome the symbol-grounding problem raised against language-based distributional models.

CHAPTER 7

Distributional models beyond language

7.1 Word meaning is both, embodied and symbolic

Within the debate on the nature of word meaning outlined in Section 6.1, an interesting position was originally advanced by Louwerse (2007, 2008, 2011, 2018). The author proposes an encompassing and hybrid theoretical framework of cognition, in which embodied and symbolic processing strategies co-exist. The core hypothesis proposed Louwerse is that language already encodes perceptual information, and therefore, language users may rely on linguistic processing and language statistics only, to perform cognitive tasks, rather than relying on energy-consuming deep mental simulations of embodied experiences. This hypothesis goes under the name of Symbol Interdependency Hypothesis and suggests that language bootstraps perceptual experiences, allowing speakers (and listeners) to be faster and more efficient in processing linguistic input than if they would have to generate and rely on fully-fleshed deep embodied simulations for each word, in order to understand one another in natural communication settings. The claim is supported by empirical data in which the author shows that distributional analyses based on simple word co-occurrences (implemented through LSA) generate word maps in which words denoting concepts that are perceptually similar cluster together, on the basis of their use in language. For example, in a popular study Louwerse and Zwaan (2009) showed that language encodes geographical information: applying LSA to newspaper texts the authors obtained similarity ratings between 50 big cities in the USA that allowed for a multidimensional scaling (MDS) of these cities. They then showed that the MDS coordinates (and the resulting map) correlated with the actual longitude and latitude of these cities, showing that cities that are geographically close to one another have names that share similar semantic contexts of use.

Together with the symbolic and amodal symbols that bootstrap perceptual information co-exist, embodied symbols, which are grounded in perception, action and emotion, and typically activate mental simulations during language processing. Relying on one or the other type of symbols is an operation that depends on the task. If the task is more language-oriented, such as in lexical decision tasks or some types of translations, amodal symbols may be sufficient. If the task is more perception-oriented, such as answering to questions that require the activation of referents and perceptual experiences, embodied symbols may be required. In

this perspective, embodied simulations are *not* always necessarily activated during language processing. As Louwerse explains in an evolutionary perspective, language has evolved such that it maps onto the perceptual system, and it therefore bootstraps meaning also when grounding is limited (Louwerse, 2018). Relying on "good enough" representations retrieved from language statistics and indexical relationships that words entertain with other words, speakers can be more efficient and faster in processing language and word meaning.

A similar account has been proposed, within the same original debate on the nature of symbols and word meaning, by Barsalou and colleagues (Barsalou, Santos, Simmons and Wilson, 2008). Their account goes under the name Language Activation and Situated Simulation (LASS) theory. According to Barsalou and colleagues, language comprehension relies on two types of processing: a 'shallow' processing based on linguistic representations obtained from language statistics, and a 'deep' processing based on situated simulations and the activation of embodied experiences and therefore grounded information encoded in the sensorimotor neural system. According to the authors, the activation of these two different types of word meaning representation depends on temporal dynamics, with the linguistic, shallow, and amodal representations peaking before the deep, situated and modality-specific representations. It remains however to be confirmed whether the latter type of representation is necessary at all, to enable comprehension.

Another proposal aimed at integrating symbolic and embodied theories has been formulated by Dove (2009), who suggests a representational pluralism, in which word meaning results from different streams of information, encoded in different types of representations. Some are perceptual (i.e., embodied and modality-specific) and others are not (i.e., symbolic and amodal). For any given word, both sensorimotor simulations *and* linguistic representations are activated, to different extents, depending on, for example, the type of concept: while concrete concepts may rely mostly on perceptual (embodied) representations, abstract concepts may rely more on linguistic (amodal) representations. This pluralistic view of cognition takes inspiration from a classic view according to which concepts are encoded in at least two general types of semantic representations: one type that is perception and motor based and another that is language based (similarly to what is suggested by the Dual Coding Theory, Paivio, 1990, 2010).

More recently, another pluralist view of cognition that encompasses both, amodal and grounded symbols, has been proposed by Zwaan (2014). In his view, the contribution of these two types of symbols to language comprehension varies from contexts to context, depending on the degree to which language use is embedded in the environment. The author distinguishes and exemplifies five different levels of embeddedness, which characterize five different types of contexts: demonstration, instruction, projection, displacement, and abstraction. These five types

of embeddedness can be placed on a scale, with the first type being more heavily in need of grounded simulations, and the last type being mostly related with amodal representations. In a following elaboration, Zwaan (2016) suggests that the co-existing sensorimotor and symbolic representations of word meaning mutually constrain each other during natural discourse comprehension. In particular, while semantic representations based on linguistic co-occurrence lead to predictions of upcoming linguistic constructions during language comprehension, and triggers the associated perceptual representations, perceptual simulations may lead to the prediction of upcoming perceptual aspects related to discourse processing and the associated linguistic constructions.

Overall, an increasing number of theoretical and empirical studies suggests that both, amodal (linguistic) representations of word meanings and modal (embodied) representations may co-exist and that their activation may depend on the type of task at hand, and the type of context in which the words may be processed. Moreover, the representations of word meaning change also depending on the type of speaker, with native speakers and language learners relying on different types of information, and therefore different types of representations, as described in the previous chapter. Words get their meaning from both, perceptual experiences and from language (thus, from other words), depending on the task, the context, and the type of speaker. Both word-to-world and word-to-word associations are strategies used to construct and represent word meaning in the human mind. But is this the case also for the artificial mind? Can distributional models integrate perceptual information in the semantic representations that they construct?

7.2 Multimodal representation of word meaning

In order to address the symbol grounding issue raised against text-based distributional models, information derived from perceptual experience had to be integrated in the semantic (vectorial) representations of word meaning. This operation was attempted in an early work by Andrews and colleagues (2009), who added to word vectors information derived from semantic features elicited from speakers who were asked to imagine and list salient properties of given referents. These authors used the database of semantic features collected by McRae and colleagues (2005). For example, to create the vector that represents the meaning of the word *fork*, the linguistic contexts in which the word *fork* is used were concatenated with perceptual features such as 'has four prongs', which was indicated by speakers as a salient property of the concept FORK in McRae's database. This feature arguably does not come up as one of the linguistic contexts in which *fork* is typically used, because it does not happen frequently to mention in written language that forks

have four prongs. However, it is arguably an important piece of information that defines forks. Andrews and colleagues, using a Bayesian probabilistic model to construct word vectors, demonstrated how word meanings can be modelled by treating linguistic and perceptual data as a single joint distribution. Their results showed that the representations of word meaning obtained in this way are more realistic and more similar to those provided by humans than the representations available from either stream of data type used individually. Language and perceptual experience are streams of semantic information that complement one another to construct rich and human-like semantic representations of word meaning, rather than being one (language) parasitic on the other (experience).

In a more recent attempt to combine linguistic and perceptual information into semantically richer and psychologically plausible word representations, Bruni and colleagues (Bruni, Tran and Baroni, 2014) exploited computer vision techniques that automatically identify discrete visual words in images (based on visual features), rather than relying on speaker-generated features that encode visual information (as in Andrews et al., 2009). For the extraction of visual features, the authors adopted a technique for image analysis called bag-of-visual-words, which discretizes the image content and produces visual units somehow comparable to words in text, called visual words. Bruni and colleagues also proposed a way to integrate features coming from language and from images into multimodal vectors. The method employed by them is arguably a more direct way to extract visual features from images, because it is not mediated by the verbalizations of the speakers, and in this way, the authors argue, the resulting representations that integrate text- and image-based distributional information may be cognitively more plausible. Even though vision is the most prominent sensory modality from which we extract meaning from extra-linguistic contexts, recent studies have attempted to integrate also sound data to learn word vectors (Kiela and Clark, 2015; Lopopolo and Miltenburg, 2015) and even olfactory data (Kiela, Bulat and Clark, 2015).

In the last decade, thanks to the availability of new large-scale multimodal datasets and of faster computers that could process high-level visual features, multimodal research reached new highs in various tasks. Among these, great attention has been devoted to the automatic recognition of emotions expressed by human faces, to the machine-generated description of images and videos (image and video captioning), to the machine-generated answering tools in which an algorithm has to answer to a question by analyzing the content of an image, and to the automatic recognition of events (e.g., Baltrušaitis, Ahuja, and Morency, 2019). As indicated by Baltrušaitis and colleagues, there are several open challenges in multimedia research, which derive from the fact that different modalities typically encode different information in different formats, and all the different streams need to be translated into a machine-readable format. The most common method used in

deep learning is to combine high-level embeddings from the different sources of information by concatenating them and then applying some mathematical transformations (e.g., softmax). The problem of balancing the information coming from different streams remains however open and constitutes one of the main challenges in this field: how much information shall be retrieved from the visual stream and from the linguistic stream respectively, and why? Moreover, does the merging technique make sense from a cognitive perspective? Finally, this operation seems to lean too much toward a strictly binary distinction between visual vs linguistic features (respectively retrieved from two separate streams), and leaves aside other possible sources of information (e.g., emotional responses, cognitive operations, other sensory reactions that are not captured by purely visual or purely linguistic corpora).

The visual information extracted from images often is based on training sets of annotated images where the annotations were initially collected through real-time "games with a purpose", created ad-hoc for collecting data from internet users (e.g., see Thaler, Simperl, Siorpaes and Hofer, 2011). Games with a purpose are an increasingly common tool used in cognitive science and data science to collect information from online gamers, by inviting them to collaborate in a task that is presented to them by means of a game. One of the most popular games used to collect visual features has been the ESP game (von Ahn and Dabbish, 2004, then licensed by Google in 2006 and used in Google Image Labeler to improve the retrieval of online images until 2016). This game was developed to harvest image-based metadata by exploiting the computational power of humans. In order to play, two remote participants that do not know each other have to associate words to an image that they both see on the screen. The two gamers are invited to coordinate their choices and try to produce the same associations as fast as possible to make points and win the game: when they produce the same tag (i.e., they associate the same keyword to an image) they make points. When they produce different tags, they do not. The way in which the game is constructed forces each participant to implicitly negotiate the information to be tagged, and predict how the other participant would tag the image. The entertaining nature of these games is crucial to keep the participants motivated during the task, and has little or no expense, but the goal of the game can constrain the range of associations that a user might attribute to a given stimulus, and trigger ad-hoc responses that provide only partial insights on the content of semantic representations. As Weber, Robertson, and Vojnovic show (2009), ESP gamers tend to match their annotations on colors, or to produce generic labels to meet quickly the other gamer, rather than focusing on the actual details and peculiarities of the image. In addition, ESP as well as other databases of annotated images harvest annotations provided by people that are not familiar with the images: images are provided by the system. Arguably, such annotations reflect semantic knowledge about the concepts represented, which are processed

as categories (concept types), rather than individual experiential instances (concept tokens). Thus, such images cannot be fully acknowledged to be a good proxy of salient perceptual information, because they are not based on perceptual experiences: the annotator has not experienced the situation captured by the image.

7.3 Flickr Distributional Tagspace, a distributional model based on annotated images

Flickr Distributional Tagspace (henceforth, FDT, Bolognesi, 2014, 2017a) addresses these issues, by proposing a simple and hybrid distributional model that (1) is based on a unique but intrinsically variegated source of semantic information, so to avoid the artificial and arbitrary merging of linguistic and visual streams; (2) contains spontaneous and therefore richer data, which are not induced by specific instructions or time constraints such as in the online games; (3) contains perceptual information that is derived from direct experiences; (4) contains different types of semantic information (perceptual, conceptual, emotional, etc.) provided by the same individuals in relation to specific stimuli; (5) is based on a dynamic, noisy, and constantly updated source of big data.

FDT is a distributional semantic space based on the tags associated with the personal images uploaded on Flickr, the image hosting service powered by Yahoo!. Because all the visual contents hosted on Flickr are user-contributed, these images tend to represent personal experiences. Each image can be considered as a visual proxy for the actual experience lived by the photographer, captured in the picture and then tagged with relevant keywords. As shown in Bolognesi (2014; 2017) tags used on Flickr are not mere descriptors of the image, but they often denote cognitive operations, associated entities, and emotions experienced during the actual experience, or triggered later on by the picture itself.

The implementation of FDT starts from the automatic retrieval of tagsets (i.e., lists of tags, each associated to an image) through the Flickr API services,[1] following the procedure described in Bolognesi (2014; 2017). Figure 18 shows, for example, three images and relative tagsets retrieved for the tag *summer*. Each of the co-tags of *summer*, such as *sunset, Austria, mountains, green*, etc., contributes to shape the semantic representation of *summer*. Of course, each co-tag contributes with a different weight, depending on the number of pictures in which it appears together with *summer*, across the pictures on Flickr. As in bags-of-words distributional models like LSA, a tagset stands for a document or an episode, within which

1. https://www.flickr.com/services/api/

the co-occurrence of a target word with the other words can be calculated. The differences between LSA and FDT are summarized in Table 3.

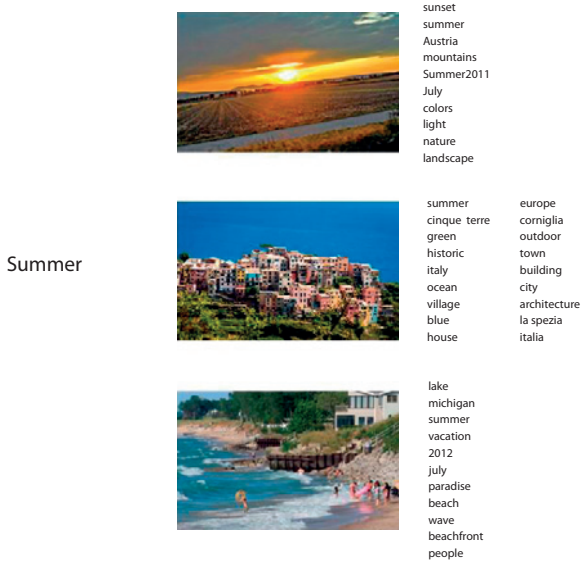


Figure 18. Three images retrieved for the tag *summer* and their relative tagsets. The images are released under a CC license

Table 3. Extracted from Bolognesi (2017a). The three main differences between LSA and FDT, pertaining context type (of the co-occurrence matrix), measure of association between an element and a context, and dimensionality reduction applied before the computation of the cosine

	LSA	**FDT**
Context	Documents of text (the matrix of co-occurrences is word by document)	Tagsets (the matrix of co-occurrences is word by word)
Measure of association	typically tf-idf (term frequency-inverse document frequency)	SPMI
Dimensionality reduction	SVD (singular value decomposition), used because the matrix is sparse.	None, the matrix is dense.

Through the Flickr API services, hundredths of thousands of tagsets can be downloaded for each target word, in order to implement a distributional semantic space with FDT. Each downloaded tagset needs to feature a target word among the first three tags, in order to be considered a relevant context for that word. After cleaning up the dataset, the word by context (all tags) contingency matrix is then constructed. In the cells, the frequencies of tag-tag co-occurrence are transformed

in more informative measures of associations,[2] and only the contexts with high measures of associations are retained, for each target word, to reduce the sparseness of the matrix. Finally, similarities between tag pairs are computed as cosine similarities between vectors.

The semantic spaces generated with FDT have been evaluated in different ways. First, in a correlation study in which the word representations were compared to those emerging from speaker-generated semantic features (McRae, Cree, Seidenberg and McNorgan, 2005) and against the similarity metrics derived from WordNet, the lexical database developed at Princeton University (Fellbaum, 1998). Results showed that speaker-generated features (e.g., the feature "has wheels" or "is a transportation" for the target word *car*) and Flickr associated tags tend to express different types of information. In particular, speaker-generated features tend to express: (1) functions, (2) external surface properties, (3) external components, (4) superordinates, and (5) entity behaviors, related to the concept denoted by the target word. Conversely, the contextual tags that co-occur with target words in FDT (e.g., the tags *road, city, family, trip*, generated for pictures that feature also the tag *car*) tend to be related to the target by means of relations that express: (1) locations, (2) associated entities, (3) superordinates, (4) functions, (5) external surface properties. Nonetheless, the average correlation scores between the semantic representations of word meaning obtained from speaker-generated features and from FDT are fairly high and they are comparable to those obtained using other distributional models (see Bolognesi, 2017a for further details). The word representations obtained from FDT show also medium and high correlation scores with the semantic representations based on semantic similarities extracted from WordNet. In WordNet, the similarity between two word meanings is based on information contained in the WordNet hierarchy. For example, *car* might be considered more similar to *boat* than *tree*, if *car* and *boat* share *vehicle* as a common immediate hypernym in the WordNet hierarchy, while *car* and *tree* do not. The similarities can be computed with the Perl module WordNet::Similarity (Pedersen, Patwardhan, and Michelizzi, 2004). Three similarity metrics were chosen for comparison with FDT: PATH and WUP, which are both based on the idea that the similarity between two meanings is a function of the length of the arch that

2. As described in Bolognesi (2014, 2017a) the measure used for this distributional semantic space is an adaptation of the Pointwise Mutual Information (Bouma, 2009), in which the joint co-occurrence of each tags pair is squared, before dividing it by the product of the individual occurrences of the two tags. Then, the obtained value is normalized by multiplying the squared joint frequency for the sample size (N). This double operation (not very different from that one performed in Baroni and Lenci 2010) is done in order to limit the general tendency of the mutual information, to give weight to highly specific semantic collocates, despite their low overall frequency.

links the two words in the WordNet taxonomy, and JCN, which is an information content-based measure, based on the idea that the more information two words share (i.e., the more they tend to appear in similar synsets), the more similar they are. The average correlations are reported in Table 3 and show that on FDT correlates well with most WordNet-based metrics of similarity as well as with the speaker-generated semantic features collected in property generation tasks.

Table 4. Extracted from Bolognesi (2017a). The average Pearson's correlation coefficients between semantic representations in FDT, McRae's features norms, and three metrics of similarity/relatedness based on WordNet (JCN, WUP, and PATH). All coefficients significant at $p < 0.005$

	FDT	McRae feature norms	JCN	WUP	PATH
FDT	1				
McRae f.n.	.69	1			
JCN	.62	.57	1		
WUP	.46	.47	.22	1	
PATH	.79	.72	.65	.65	1

The semantic representations of word meaning emerging from FDT were evaluated also in a categorization task. In a cluster analysis ($K = 11$) the data extracted from FDT clustered automatically into semantically coherent classes (Figure 19), showing accurate intra-category distinctions between different types of vehicles

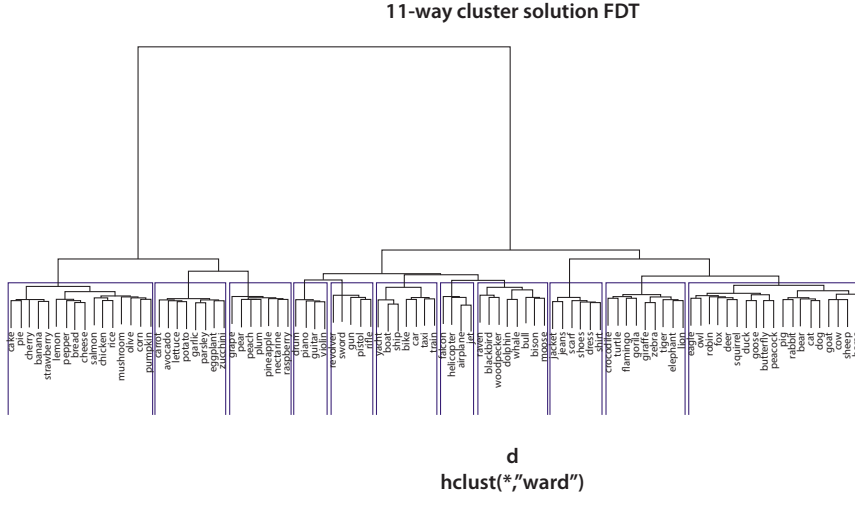


Figure 19. Extracted from Bolognesi (2017a). Cluster analysis performed in R with the function *hclust* (R-core R-core@R-project.org) on FDT data. The function *cutree* shows the solution for an 11-way partitioning (red lines around the six supported clusters)

(air, ground, and water transportation respectively), as well as different types of animals (farm animals vs. wild animals, respectively).

Finally, the ability of FDT to model an inherently perceptual domain (i.e., the domain of words denoting colors) was compared to the same classification performed by two text-based classic distributional methods: LSA and DM. The idea behind this analysis was to explore whether FDT, based on annotated images, could harvest and model word representations that encompass more perceptual information, compared to 'blind' distributional models based on solely linguistic information (Bolognesi, 2014). The results, in Figure 20, show that the distributional similarities constructed by FDT reflect the distribution of the wavelengths perceived by the three types of cones that characterize the human eye, which make us sensitive to three different spectra of light: the blue light, the green light and the red light. This physical distribution of the color waves reflects the order in which the colors appear in the rainbow.

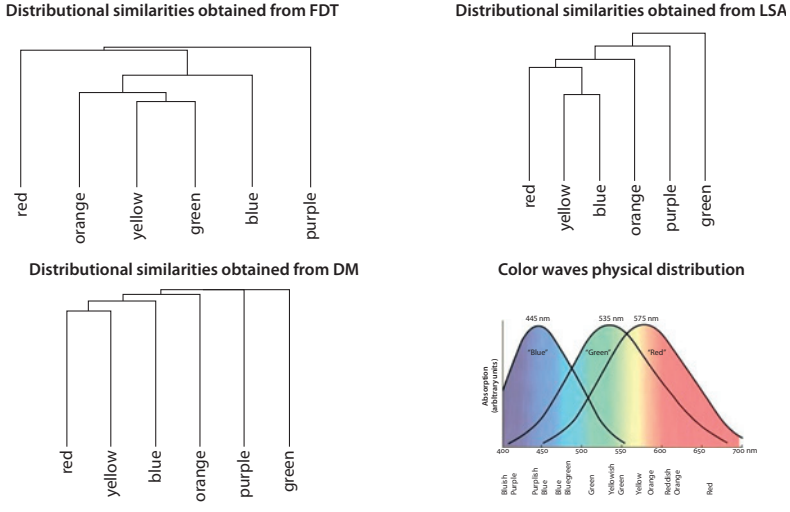


Figure 20. Adapted from Bolognesi (2014). The dendrograms showing the distributional similarities between color terms obtained from DFT, LSA and DM, and compared to the distribution of the color waves in the physical domain. In the three dendrograms, the shorter the arch, the higher the distributional similarity between two color terms

As shown in Figure 20, the three distributional models harvest different types of information and generate different types of semantic spaces, clustering the color terms according to different parameters. On one hand, the two models based on linguistic corpora produce fairly similar word spaces, where the three primary colors (*red, yellow*, and *blue*) seem to cluster together and are followed by the three secondary colors. In both models, *green* is the 'odd one out', which means that is the least similar to the others in terms of linguistic distributions. As a matter

of fact, *green* is probably the most polysemous word, among the 6 terms: as an adjective, *green* can indicate a shade of color, but also an area with lots of plants, something not ready to be eaten, or (metaphorically) an unexperienced person. Moreover, *green* can be used as a noun to designate a large area of grass, and even a member of the Green political party (both meaning extensions are derived from the basic meaning by means of metonymic chains). All these meanings are conflated in the rich semantic representation of *green* that emerges from Flickr tag distribution, making the overall representation distributionally different from the representations of the other color terms.

As discussed in Bolognesi (2014), the very high coefficient of correlation between LSA and DM supports the overall consistency and robustness of these two language-based models: although they are based on different linguistic corpora and different techniques, which allow the retrieval and analysis of different linguistic contexts and the construction of different vectorial representations, these two models provide very similar outcomes. Moreover, both LSA and DM show fairly high degrees of correlation with the representations obtained from FDT. This suggests that language by itself provides a rich source of information that to a certain extent reflects the perceptual information triggered by the visual contexts gathered in Flickr with a good degree of accuracy. However, the distribution that emerges from FDT shows peculiarities that do not appear in the two models based on word co-occurrences. In particular, the FDT distribution of color terms suggests that the distributional hypothesis applied to information related to perceptual experiences that we live through our bodies and through our senses shapes word representations that reflect perceptual similarities (e.g., the perceived similarities between colors). Conversely, when we look at linguistic contexts in which we use the words that denote such colors, and model word representations using distributional methods that capture linguistic information only, we obtain different distributions, that seem to reflect better our encyclopaedic knowledge about primary and secondary color terms.

7.4 From word-to-world to world-to-world modelling

Classic distributional models based on word frequencies are based on corpora of text. Therefore, the word meaning representations that they construct are based on word-to-word references. Multimodal distributional models, instead, integrate extra-linguistic information in the construction of word meaning, thus integrating word-to-world references. The third type of reference that allows us to construct and represent meaning, besides word-to-word and word-to-world references, is based on world-to-world associations, as anticipated in Chapter 2 and Chapter 4.

This type of relation does not involve linguistic symbols (i.e., words) at all, but only perceptual symbols (terminology borrowed from Barsalou, 1999). The representations that emerge from world-to-world associations are grounded directly in perceptual experience and, in principle, are independent from language.

While the focus of this book is on word meaning, and therefore more emphasis is put on the word-to-word and word-to-world relations, this third type of relation, between elements in the world, constitutes a core aspect of meaning making, based on which, for example, infants start making inferences and perform categorizations. As I will further elaborate in Part 3, the mechanism that supports the construction of word meaning (that is, the associative principle between elements that occur next to each other and the paradigmatic shift that allows categories to emerge) is the same that supports the construction of conceptual content starting from associations in perceptual experiences. Therefore, for the sake of clarity and exhaustiveness, I will now briefly focus on how world-to-world relations are modelled by means of computational techniques based on the detection of statistical regularities among perceptual elements.

Interestingly, this challenge became the central endeavour of a community of scientists that was mainly based in the United States, back in the Seventies. This community, which features the pioneers of the contemporary artificial neural networks, was interested in solving problems related to the creation of artificial intelligence (AI). One of the core problems turned out to be modelling computer vision, in tasks such as image recognition, and in particular the identification of objects within images. This task, in fact, is based on the identification and classification of features that are inherently perceptual, rather than linguistic. The main problem in this endeavour is that in real life the same object can take very different shapes and colors, depending on the context in which it appears. For example, the same white porcelain coffee cup can look very different if it is seen from above, from below, or from a frontal angle. Moreover, different levels of light can change the perception of its color. While the human eye is able to group different visual instantiations of a coffee cup within the same conceptual category (the COFFEE CUP conceptual category), for a computer this turns out to be a major problem because it cannot easily rely on the detection of a predefined shape or color hue. A funny anecdote that exemplifies this problem is commonly told within the AI community (Sejnowski, 2018). Back in the Sixties, the MIT AI Lab was awarded with a large grant to build a robot that could play ping pong. The PI of the project, Marvin Minsky, funder of the MIT AI Lab, apparently forgot to list in the grant budget some funds dedicated to the implementation of a software capable of understanding vision, for the robot. Because back in the Sixties the problem of computer vision was naively considered to be a fairly easy issue to solve, the Lab decided to assign it to undergraduate students. As a matter of fact, this problem, which seems quite an easy one

for the human eye, kept generations of top AI researchers in computer vision busy for the next decades. The reliability of computer vision algorithms was recently achieved only thanks to the modern neural networks, which are used nowadays in extremely sophisticated systems such as self-driving cars.

While describing the functioning of modern (deep) neural networks lies beyond the scope of this book, the pioneering system that led to their creation can usefully exemplify the way in which perceptual features are extracted from visual stimuli by means of computer vision techniques that mirror the human vision system, to learn to classify similar objects within the same conceptual category, based on world-to-world relations, and therefore bypassing the information encoded in language. One of the pioneers in this field, who implemented the direct antecedent of current neural networks used for image recognition, is Frank Rosenblatt. His algorithm, called Perceptron (1958), is the simplest neural network possible: a computational model of a single neuron that can classify an input in a binary way (e.g., 0 or 1, which could stand for cat vs. dog, black vs. white, etc.). The single-layered Perceptron consists of an input layer made of input values with their own weights;[3] a hidden layer that processes the clues obtained from the input layer by aggregating them (summing them up) and applying to them a (linear) function; and a single output, which is the classification of the input into one of the two desired categories. It is called a feed-forward neural network because the information flows in one direction from input to output, without making loops (like in recurrent networks). The function, called activation function, is the key to classify the information obtained from the input layer into a yes or a no, a cat or a dog, or any other desired binary classification. Typically, this function returns 1 only if the aggregated sum obtained from the input layer is more than some threshold; otherwise it returns 0.

For example, Perceptron may be used to predict whether a picture of a fruit displays an orange or a banana.[4] To do so, ideally just two discriminatory features are needed: shape (round for orange and non-round for banana) and color (orange for orange and non-orange for banana). Given a bunch on (unambiguous) pictures of oranges and bananas, labelled as oranges and bananas, Perceptron will first deduct a way to separate the two groups, and thus it will arguably learn that the two discriminatory features are size and color (without actually labelling them as "size" and "color" though). Then, given new instances of pictures displaying oranges or bananas, it will correctly classify them as oranges or non-oranges (i.e., bananas), based on the values measured on the two identified features, which are zeroes if the feature is not observed and ones if the feature is observed.

3. Inputs are multiplied by their own weights

4. This is a conceptual example. In reality, the distinction between the two fruits in natural photographs would require a more sophisticated neural network.

This approach is crucially different from previous approaches, in which the engineers had to manually configure the features based on which the classification of a new instance was performed. In the algorithm proposed by Rosenblatt (1958), and then adopted by contemporary scholars in deep learning, the algorithm learns features from raw data. The learning process requires not only a number of instances to which the algorithm needs to be exposed in order to learn the relevant features, but also a fair amount of computer power.

Contemporary neural networks scaled up this basic idea and, by relying on a much larger computer power, a much larger dataset for the training phase, and a deeper structure of hidden layers and latent features, can approach pattern recognition problems in situations that are more realistic. For example, self-driving cars nowadays employ these types of algorithms to determine whether a pedestrian is crossing the road, even if her face is not perfectly visible and frontally displayed to the car vision system.

This brief digression on the functioning of Perceptron exemplified how world-to-world associations can be computationally modelled, using methods that rely on the idea that meaning is distributed across features and classifications can be performed on the basis of shared features. In this case, the system exploits visual features, because the task to be performed is perceptual rather than linguistic: a picture needs to be classified into the ORANGE or the BANANA category. Once the relevant visual features are learned (e.g., fruit is yellow: yes/no; fruit is round: yes/no) the system can process a new series of weights (i.e., coordinates) from a new picture used as input, and determine whether the fruit represented in the new picture is a banana or an orange. Interestingly, the Perceptron was developed years before LSA, the pioneering distributional model based on word-to-word associations, which then prompted the debate on the grounded nature of meaning. From a historical point of view, it must be mentioned that neural networks like Perceptron had an alternate fortune, and became popular only in very recent times, when the incredible potential encapsulated in its architecture could be better appreciated. Today, thanks to the exponential increase in available computer power, contemporary (deep) neural networks consist of several input layers, sophisticated activation functions and additional mechanisms that evolved from the basic organization of Perceptron.

For language processing and in relation to the modelling of word meaning, the classic approach to language processing and modelling in those years (i.e., from the Fifties through the Eighties) was mainly focused on grammar and rules that were imposed on the data in a top-down manner, supported by theories of language such as the Chomskian generative grammar (Chomsky, 1957, 1975). Conversely, neural networks, as well as the distributional models that emerged in the late Nineties, take a completely different approach to language and meaning

constructions, which is directed from the bottom-up and does not start from sets of pre-determined rules, such as grammatical ones. As the pioneering scholar in neural network Terrence Sejonwski explains:

> Chomsky's emphasis on word order and syntax was the dominant approach in linguistics in the latter part of the 20th century. But even a bag-of-words model neural network that throws away word order does remarkably well at determining the topic of an article, such as sports or politics, and its performance can be improved by taking the immediate neighbouring words in the article. The lesson from deep learning is that even though word order carries some information, semantics, based on the meaning of words and their relations with other words, is more important. (Sejnowski, 2018)

From a theoretical point of view both Perceptron and the classic distributional models are based on the same principles that can be summarized as follows: meaning is distributed across features and can be formalized as a series of coordinates, each carrying some type of information about a target item. In classic distributional models (e.g., LSA) word meaning is represented by coordinates that consist of other words, used as contexts. Meaning is therefore constructed and represented through word-to-word relations. In multimodal distributional models (e.g., FDT) word meaning is represented by coordinates that express multimodal (including perceptual) information. Meaning is therefore constructed and represented through word-to-world relations.

7.5 Summary

Recent views, supported by empirical data, suggest that word meaning and word processing rely on different types of representational structures: symbolic amodal ones, as well as embodied and modality-specific ones. That is, word meaning is determined by both, word-to-word references, as well as word-to-world references. The different representations of word meaning (i.e., those based on word co-occurrences and on linguistic information retrieved from other words, and those based on information retrieved from perceptual experiences) may be both activated, or compete for activation, depending on the nature of the task to be performed, the context in which the word is processed, and the type of speaker.

Word meaning is therefore both, embodied and symbolic, and the two streams of information may contribute in slightly different ways to construct a word representation. Moreover, extra-linguistic information can be integrated in the vectorial representations constructed by distributional models in various ways, thereby generating representations that are more cognitively plausible.

The last part of this book is dedicated to the elaboration of the converging evidence that cognitive scientists and psychologists (Part 1) and computer scientists (Part 2) contributed to bring to the surface in the past decades. Such converging evidence will be discussed within the field of linguistic and language sciences. In particular, I will further explain some core aspects related to how words get their meaning, and the implications that such process has for our ability to perform classifications, understand one another in natural communication settings, and learn foreign languages.

The claims that emerged from the first two parts of this book and that I will take one by one and elaborate in the last part, can be summarized as follows:

- Words get their meaning from both language and experience.
- Word meaning is distributed across both language-based and experience-based features.
- Word meaning is dynamic and changes according to context and task conditions.
- Different types of words (e.g., words denoting abstract and concrete concepts) get meaning from language and from experience to different extents and in different percentages.
- Different types of speakers (e.g., native speakers and language learners) construct meaning by relying on perceptual and linguistic information to different extents and in different percentages.
- The distributional hypothesis, which led to the implementation of contemporary models of distributed word meaning, has deep cognitive foundations and is based on cognitive principles that have been overlooked for too long. Such principles constitute core mechanisms that explain how words acquire meaning.

PART 3

Converging evidence in language and communication research

CHAPTER 8

Where words get their meaning

8.1 How language and experience construct categories

Words are useful labels that we use to glue together items within a same category. For example, the word *dog* labels the conceptual category of dogs, which includes individual members of this category. As described in Chapter 2, however, words do not simply and parasitically support conceptual categories formed on the basis of perceptual experiences. Words have the power to force the construction of conceptual categories, and to attract new members within such category. Moreover, words have the power to override previous categorizations that were established on the basis of solely perceptual experiences. In this sense, words can construct categories that theoretically may differ from the categories that arise from perceptual similarities observed in experience. Consider the following example:[1] What is the difference between *soup* and *smoothie*? Instinctively, one could say that these two words label two different conceptual categories, SOUP and SMOOTHIE, and that the difference between items in the first and in the latter category are mainly due to the following features: temperature and sweetness. Soups tend to be warm and savory, smoothies tend to be cold and sweet. Moreover, one could argue that the SOUP category might have much more variability among its members, while the SMOOTHIE category might be more tightly clustered around the prototype. As a matter of fact, not even these two simple features (temperature and sweetness) are consistently configured as semantic traits by all members of the two categories: there are cold soups (*holodnik* for example is a cold beetroot Russian soup), sweet soups (*corn chowder* and *melon soups* are examples of sweet soups), as well as savory smoothies made with vegetables. One could then argue that the discriminating feature between SOUP and SMOOTHIE is not an inner feature (or a combination of entity-related features) but rather an external relation: soups tend to be consumed at the beginning of a meal, possibly as starters, while smoothies tend to be consumed as desserts or as snacks outside the main meals. Still, this does not solve the search for a common meaning that can be used to discriminate soups from smoothies. It can therefore be suggested that the differentiation boils

1. I am grateful to Paul Minda, Monica Gonzalez-Marquez and Lisa De Bruine for the fruitful message exchanges on Twitter, in which this example was discussed (prompted by Paul Minda).

down to a linguistic label: imagine finding in a restaurant or at a café abroad a typical edible fluid: if it is called *soup*, it is a soup, and if it is called *smoothie*, it is a smoothie. This example shows that while the categories of smoothies and soups can be intuitively organized around prototypical items, there are members of such categories that are peripheral in relation to the prototype, up to the point that their inclusion in one of the two categories boils down to a convention captured by the linguistic label used to name the referent. In this sense, words are tools that enable us to attract and include new experiences, new items, and new members within a given category and, as a consequence, enable us to expand the content and the meaning of such category.

In general, the construction of conceptual categories starting from the buzzing confusion of individual perceptual experiences, is based on a sequence of steps.

The first step is an associative process thanks to which we establish associations between words and referents (or words and other words) that appear or fail to appear in the same situation. This associative mechanism is supported by behavioral evidence showing that both children and adults tend to learn associations by crossing situations (Chapter 2.2) and by learning from co-occurrences as well as from missed (but expected) occurrences (Chapter 6.2). In this view, learning associations is a dynamic process, which is constantly updated by exposure to new perceptual contexts as well as linguistic contexts.

The second step consists of a pattern detection mechanism, also widely documented as a hallmark of human cognition, thanks to which we identify recurring configurations: we can observe, for example, that the objects (or perceptual features, or words) A, B, and C tend to appear together with X, and that the same objects (or perceptual features, or words) A, B, and C tend to appear also together with Y. We then detect a repeated pattern between the elements that surrounds X and the elements that surround Y, which is the pattern: A, B, C.

The third step consists of a category construction and the recognition of a paradigmatic similarity between X and Y: because these two items (or words) display a similar pattern, then the two items are similar to one another and can be grouped together into one category. In particular:

– If X and Y are two objects and A, B, and C are also objects, or perceptual properties of objects, then world-to-world associations are established and based on these, X and Y form a pre-linguistic category grounded fully in perceptual experience.[2] For example, X = soccer ball, Y = tennis ball, A = round shape,

2. Note that, as I explained in Chapter 4, A, B, and C can be whole items that tend to occur with X and with Y, or they can be components or perceptual features of X and Y. In principle, these can construct two different types of similarity, one called relational similarity and the other called attributional similarity (see Bolognesi, 2016b; Bolognesi, 2017b).

B = floor, C = hands holding the ball. Then, the tennis ball and the soccer ball are perceived to be similar and may become part of the same category, the category of BALLS, without (technically) any linguistic intervention or any linguistic label needed to form such category. This category is driven by perceptual properties and it is independent from language. According to the literature, this way to construct categories can be easily overridden by linguistic labels. For example, a child may infer that, based on the perceptual similarities listed above, a light, soft, air-inflated ball attached to a string is also a member of the category of BALLS, together with the tennis ball and the soccer ball. However, when she hears that such object is named *balloon*, then she will construct, on the basis of such linguistic label, a new category for BALLOONS, which is separate from the category of BALLS.

- If X and Y are two words and A, B, and C are objects, or perceptual properties of objects, then word-to-world associations are established. Word X and word Y both become associated with A, B, and C. Notably, if X and Y are known words, it means that they already represent categories of instances. The pattern of X and Y is then observed to be the same. Thereby, the words X and Y are grouped in the same category. For example, X = *marble*, Y = *tennis ball*, A = round shape, B = floor, C = hands holding the ball. If the words marble and tennis ball are already known, and therefore they already stand for the categories MARBLE and TENNIS BALL, then these two categories become part of the same superordinate category, the generic but concrete category of BALLS, thanks to the fact that they share the same patterns of word-to-world associations (i.e., the A, B, and C perceptual elements).[3] A superordinate category is formed, the category of BALLS, which encompasses the categories of MARBLES and TENNIS BALLS, which share word-to-world patterns of associations.
- If X and Y are two words and A, B, and C are also words, then word-to-word associations are established. Based on these patterns of associations, X and Y become members of the same category in virtue of the fact that these two words tend to be used in the same linguistic contexts. For example, X = *ball*, and Y = *teddy bear*, A = *playing*, B = *my*, C = *children*. Then, the category of BALL and the category of TEDDY BEARS become members of the same su-

3. Note that cross-situational learning works on the basis of this same type of association, but it is used to learn word meaning by solving referential ambiguity in context: given a word X (e.g., *ball*) and two possible referents, like a teddy bear and a ball, over multiple exposures to situations in which balls and teddy bears appear or do not appear, children (and adults) learn the correct association between the word *ball* and the referent ball.

perordinate category, the generic (but still concrete) category of TOYS, thanks to the same patterns of word-to-word associations observed in language use.[4]

From a developmental perspective, it has been shown that at least from the age of two months onwards, infants can form perceptual categories based on world-to-world associations (Quinn, Eimas and Rosenkrantz, 1993; Westermann and Mareschal, 2014). Then, more or less around the completion of the first year, language starts to become part of the way in which such semantic representations are shaped, and labels (words) used in conjunction with information extracted from perceptual experience open up the way to the structuring of a more sophisticated conceptual knowledge (Westermann and Mareschal, 2014). The transition from prelinguistic to language-mediated object categorization is characterized by an interesting phenomenon, mentioned in various parts of this book, according to which language can actually override previous categorizations based solely on perceptual similarity. This incredible phenomenon demonstrates that language constructs meaning, and that words are not simple labels used to name conceptual labels that are constructed exclusively on the basis of perceptual experience. Words are powerful tools that enable categorizations, which are in turn forms of abstraction. Even the concrete category BALL is a form of abstraction, from the individual items labelled as balls.

There is a rationale behind the order in which the different types of association that enable the construction of conceptual categories (i.e., world-to-world, word-to-world, and word-to-word associations) have been consistently presented throughout this book. That is: such order reflects different degrees of abstraction and therefore different degrees of groundedness of the symbols that are involved. When an association is established between two elements that co-occur in perceptual experience, the symbols in the mind used to represent such entities are arguably fully grounded in perceptual experience, because they directly reflect the perceptual properties on which such symbols, sometimes called perceptual symbols (Barsalou, 1999) is constructed. Conversely, when an association is established between a word and an object, or a word and another word, the symbols in the mind used to represent this meaning can be grounded or not, depending on context situation, goals, and task at hand, as described in relation to the grounded nature of word meaning in Chapter 6. This means that words can trigger deep conceptual

4. Note that the semantic categorization of word meanings based on word-to-word associations is probably less accurate when we consider function words, because of the too wide array of possible contexts in which such words can be used. For example, it would arguably be quite difficult to learn, based on solely word-to-word associations, that the preposition *on* and the preposition *above* are semantically similar because they tend to be used to name spatial relations between two items displayed on the vertical axis.

simulations or can be processed without accessing such simulations. In particular, word-to-world associations are more likely to trigger mental simulations of perceptual experiences, compared to word-to-word associations. It follows that if two words are distributionally similar because they share similar patterns of word-to-word associations, their processing is less likely to involve the activation of grounded representations and perceptual features. This seems to be, typically, the case for words denoting abstract and generic concepts. For instance, while writing this book, I often used the two expressions *for example* and *for instance*, to exemplify my claims. The words *example* and *instance* are distributionally similar, because they tend to be used in the same linguistic contexts (typically preceded by the preposition *for*). The words *example* and *instance* share therefore word-to-word patterns of associations and are likely to be processed linguistically, rather than by means of mental simulations involving perceptual experiences.

The literature on cognitive and lexical development in children explains that the ability to form categories starts from the construction of perception-based categories and subsequently moves onto language-mediated categories. World-to-world associations precede word-to-world associations. I argue that, in turn, word-to-world associations precede word-to-word associations for the same reasons: the latter type of associations requires a higher cognitive ability to mentally manipulate symbols (words) that are more abstract than those used to represent perceptual categories. The introduction of language in children's input boosts dramatically their ability to abstract, by pushing them to construct categories not only on the basis of world-to-world associations, but also on the basis of word-to-world and eventually word-to-word associations. Moving from the construction of categories on the basis of word-to-world associations to the construction of categories on the basis of word-to-word associations is by itself a process of abstraction, based on analogical reasoning. The processes and steps involved in the construction of word meaning in cross-situational learning, based on word-to-world associations is then applied elsewhere, it is transferred by means of analogical mapping to a different way of constructing meaning, which is based on word-to-word associations.

This ability becomes particularly interesting when we look at how concrete vs. abstract conceptual categories are formed, and at how the meaning of words denoting concrete and abstract concepts is acquired. As observed in Chapter 2, words denoting abstract concepts tend to appear later than words denoting concrete concepts in children's vocabulary development. This is arguably due to the fact that abstract categories are more strongly shaped by language than by perceptual experience, and the ability to detect patterns in language develops after the ability to detect patterns in perceptual experience.

8.2 Word-to-world associations in constructing the meaning of words denoting concrete and abstract concepts

In Chapter 2 I provided a first overview of the points that are developed in this book, and after introducing the cross-situational learning paradigm, and explaining how this is used by children (and adults) to solve the problem of referential ambiguity and associate the correct referent to the correct word, I continued by explaining that such a paradigm presents obvious problems when it comes to words denoting abstract concepts. In particular, how are words denoting abstract concepts associated to their correct meaning, when there is no referent to be associated across multiple exposures to experiential input? If word-to-world associations learned in cross-situational settings were the only mechanism used to learn word meaning, then how are words denoting abstract concepts learned? I argued that, in line with a large body of literature, information about word meaning comes from two different streams of knowledge: perceptual experience *and* language.

The meaning of words denoting abstract concepts consists of more linguistic than perceptual information, whereas the meaning of concrete concepts consists of more perceptual than linguistic information, roughly speaking. This is supported by empirical evidence showing a greater engagement of the verbal system for the abstract words and a greater engagement of the perceptual and mental image generation systems for the concrete concepts (Hoffman, 2016). Note that the meaning of words denoting abstract concepts does not consist of *exclusively* linguistic information (i.e., word-to-word associations). Similarly, the meaning of concrete words does not consist of *exclusively* perceptual information (i.e., word-to-world associations). The meaning of words denoting abstract concepts may in principle encompass also information extracted from perceptual experience, and the meaning of words denoting concrete concepts can in principle encompass also information extracted from language.

But how is the information coming from perceptual experience encoded in the meaning of words denoting abstract concepts, given that it cannot be linked to the word via direct association to a single tangible referent, uniquely associated to the abstract word? The debate around this issue is still open and quite heated (for a review, see Bolognesi and Steen, 2018). The most promising view (as argued also by Pecher, 2018) is the situation-based view, which suggests that for abstract concepts (and therefore for the meaning of the related words) the information connected to perceptual experience may come from word-to-world associations in which the referentiality does not pertain to a single tangible referent (as for concrete words, such as *bike*, which is associated with the object 'bike'). Instead, such association involves the abstract word and whole situations and properties of the whole situations, rather than properties of an individual referent (Barsalou and

Wiemer-Hastings, 2005; McRae, Nedjadrasul, Pau, Lo and King, 2018). For example, Figure 21 displays an instance of an experience, in which the word-to-world associations afforded by the meaning of the concrete word *bike* and the meaning of the more abstract word *excursion*, are highlighted.

From a cross-situational learning perspective, a child that does not know the meaning of the word *excursion* has to disambiguate this meaning thanks to the exposure to multiple situations in which the word *excursion* is uttered. However, it is unlikely that such a pattern emerges from repeated concrete entities and perceptual elements, because an excursion can involve one or many participants (solo, family, group, etc.), different possible ways of transportation (bike, car, train, feet, horse etc.), different destinations (mountains, beach, cities, underwater, etc.), and so on. The situations in which the word *excursion* is uttered are very diverse, and do not share perceptual features detectable from perceptual experiences. What different experiences of excursion share are *categories* of perceptual features, such as PARTICIPANTS (excursions need participants), TRANSPORTATIONS (excursions require a way of transportation), DESTINATIONS (excursions require a destination), ITINERARIES (a route) and so on. Understanding the meaning of words that denote abstract concepts, such as *excursion*, from a cross-situational perspective, requires understanding that there isn't a single referent repeated consistently across experiences (as for concrete concepts) but a configuration of elements, such as participants + transportation + destination + itinerary. Such elements that are repeated consistently across situations seem to be expressed at a taxonomically higher level of genericity: an excursion requires a medium of transportation, which is a rather generic category, compared to the more specific category of buses, cars, trains and so on. An additional layer of categorizations are therefore required in order to understand the meaning of an abstract concept a in cross-situational setting, compared to the simple word-referent match that can be learned for concrete concepts. For the construction of the meaning of a word denoting an abstract concept in a cross-situational setting, a viewer/speaker needs to understand first of all that the word-to-world association between the abstract word and something out there in the world needs to be constructed between the abstract word and many elements within the experience where the word is uttered, not with just one element on the scene. Moreover, the elements in the world to which the abstract word is associated are not the individual instances present in a specific scene, but the categories of (concrete) elements therein, expressed already at a higher level of abstraction. In this sense that pattern repeated across situations, required to construct the content of the abstract concept, is not the exact pattern of concrete elements observable in individual experiences (e.g., 1 biker, a bike, and a mountain, in Figure 21), but is a pattern of elements expressed at a taxonomically higher level of abstraction (e.g., participants, transportation, and destination,

in Figure 21). All these experience-based elements together, expressed at a more generic taxonomic level (e.g., participants, destination, etc) than the individual observable categories (e.g., biker, bike, etc), enable the construction of the content of the abstract concept and its related word (e.g., transportation). Such form of categorization, from the basic level lexicon to the superordinate level (e.g., from BIKE to TRANSPORTATION) requires therefore a form of abstraction from more specific to more generic categories. This, in turn, suggest that abstract concepts, which lack a tangible referent, may also be on average more generic than concrete concepts, a hypothesis that has been recently tested by Bolognesi, Burgers and Caselli (2020), and briefly elaborated in the next chapter.

Figure 21. The word-to-world associations entertained by a word denoting a concrete vs a word denoting an abstract concept

8.3 Word-to-word associations in constructing the meaning of words denoting concrete and abstract concepts

It is relatively easy to argue that word-to-world associations construct semantic representations in different ways, depending on whether a word denotes an abstract or a concrete concept. This is due to the fact that concrete concepts have tangible referents in the world, that can be experienced through our bodies, while abstract concepts do not have tangible referents that can be directly experienced. As indicated above, recent cognitive and neuropsychological evidence shows that even the neural substrates that are involved in the processing of words denoting abstract vs. concrete concepts are different, with evidence supporting a greater role played by the systems involved in perception and action for the processing of concrete words. For example, studies on patients with brain damage in the cortex areas involved in the representation of visual-sensory aspects of semantic knowledge have difficulties in processing and understanding words denoting concrete concepts, while they do not have problems with words denoting abstract concepts (see Hoffman, 2016 for an extensive review).

It is probably less straightforward to argue that word-to-word associations contribute in different ways to shape the meaning of words denoting abstract vs. concrete concepts. In what sense would the linguistic information retrieved from word co-occurrences differ for abstract and concrete meanings respectively? And *why* would it differ for abstract vs. concrete meanings?

Behavioral and neuroscientific evidence shows that the linguistic context in which a word is presented plays a greater role in determining and disambiguating the meaning of words denoting abstract concepts, compared to concrete concepts (e.g., Hoffman, Jefferies and Lambon Ralph, 2010). Besides the psychological evidence, in a recent study that employed distributional semantic techniques the degree of contextual variability associated with abstract and concrete words was compared (Hoffman, Lambon Ralph and Rogers, 2013). In this study the authors measured the semantic diversity of contexts in which concrete and abstract words are used. They found that concrete words (e.g., *spinach*) tend to be used in a restricted, inter-related set of linguistic contexts while abstract words (e.g., *life*) appeared in a wider range of diverse contexts. Moreover, because of this wider array of possible contexts of use for abstract words, the strength of association with any one context may be very weak. This is consistent with the behavioral results previously reported by Schwanenflugel and Shoben (1983) and Schwanenflugel, Harnishfeger, and Stowe (1988) showing that participants find it harder to think of one specific linguistic context in which they could use an abstract word, while they find it easier to think of one specific linguistic context in which they could use a concrete word.

From a neuropsychological perspective, a number of studies reported a greater activation of the superior anterior temporal lobe (superior temporal sulcus and gyrus) for words denoting abstract vs. concrete concepts (Binder, Desai, Graves and Conant, 2009; Noppeney and Price, 2004; Sabsevitz, Medler, Seidenberg and Binder, 2005; Wang, Conder, Blitzer and Shinkareva, 2010). This brain area is associated with the comprehension of speech and text, particularly at the sentence level (Humphries, Binder, Medler and Liebenthal, 2006; Scott, Blank, Rosen and Wise, 2000; Spitsyna, Warren, Scott, Turkheimer and Wise, 2006). The fact that such an area is more involved during the processing of abstract vs. concrete meanings supports the idea that the comprehension of abstract words places strong demands on linguistic aspects of semantic knowledge.

Besides the fact that the meaning of words denoting abstract concepts is more deeply shaped by language (and by the linguistic contexts of use) compared to the meaning of concrete words, it is interesting to compare the representational frameworks underlying the organization of word meanings denoting abstract and concrete words respectively. This type of research shows that the way in which the relationships among concepts are organized differs, for concrete vs. abstract

concepts (Crutch and Warrington, 2005). In particular, while concrete words are organized primarily in terms of paradigmatic relations of semantic similarity (e.g., *dog* is associated with *wolf*: the two words are co-hyponyms) abstract ones are organized primarily in terms of thematic and broadly speaking syntagmatic relations (e.g., *justice* is associated with *law*: the two words are linked by a broadly speaking thematic connection that is hard to pin down). As reported in (Hoffman, 2016), evidence for this view comes mainly from two sources: on one hand from aphasic patients with semantic deficits who experience interference between semantically related concepts (Crutch, Ridha and Warrington, 2006; Crutch and Warrington, 2005, 2010; Warrington and Cipolotti, 1996; Warrington and Shallice, 1979), and on the other hand from healthy participants involved in psycholinguistic tasks in which, given a list of related words, they were asked to detect the semantically anomalous one (Crutch, Connell and Warrington, 2009; Crutch and Jackson, 2011). For concrete words, participants are faster to spot the odd-one-out when the other words were semantically (paradigmatically) similar, while for abstract words they were faster to spot the odd-one-out when the other words were related by means of thematic (syntagmatic) associations.

Most importantly, the paradigmatic similarity that tends to govern the organization of the meanings of concrete words in the studies reported above seems to arise from the fact that the referents of the related concrete concepts share many of the same perceptual features. The paradigmatic similarity is thus based on shared perceptual features (Hoffman, 2016). In terms of cognitive development, word-to-world associations precede word-to-word associations: they are established during early development, in order to learn word meanings based on perceptual experiences. Conversely, word-to-word associations are established later, because they require a more sophisticated level of cognitive development and ability to manipulate exclusively abstract symbols. Arguably, once the meanings of concrete words are constructed and organized in the mental lexicon, starting from word-to-world associations, in principle there is no need to re-structure them from scratch, once the word-to-word associations become a viable option for meaning construction. Similarities between word meanings can be easily constructed thanks to the shift between syntagmatic and paradigmatic levels: *dog* is similar to *wolf* because the two referents share many perceptual features (i.e., many word-to-world associations).

Conversely, for words denoting abstract concepts there aren't shared perceptual features on which a paradigmatic similarity can be directly constructed. Thus, abstract word representations cannot be organized in the mental lexicon by means of paradigmatic similarities constructed on the basis of word-to-world associations involving shared perceptual features. A different organizational framework must characterize the representations of abstract words. From a developmental perspective, when word-to-word associations become a viable option to construct

semantic representations, then the meaning of words denoting abstract concepts can be constructed. However, the word-to-word syntagmatic associations for abstract words are typically numerous but weak (Schwanenflugel and Shoben, 1983; Schwanenflugel et al., 1988; Hoffman, Lambon Ralph and Rogers, 2013). As a consequence, there is not enough strength of association between an abstract word and other co-occurring words in language, to enable the construction of proper paradigmatic relations of similarity between semantic representations in the mental lexicon. For this reason, abstract words remain organized in the mental lexicon mostly by means of (weak) syntagmatic word-to-word associations, rather than by genuine and strong paradigmatic similarities.

To conclude, the meaning of abstract words is strongly dependent on language and on word-to-word *syntagmatic* associations between words, and it is susceptible to context variation (meaning varies deeply as a function of the linguistic context in which the word is used). Conversely, the meaning of concrete words is strongly dependent on perceptual experience and on word-to-world associations on which *paradigmatic* relations of similarity are constructed.

8.4 Word meaning organization in the L1 and L2

The principles that govern the organization of word meaning representations in the mental lexicon vary not only as a function of the type of word (we saw the differences between words denoting abstract and concrete concepts) but also as a function of the type of speaker, as described in Chapter 4. I explained that the organization of word meanings in the L2 tends to be more strongly driven by linguistic factors, compared to the organization of word meanings in the L1. This difference emerges, for example, from free word association tasks reported in the literature, showing that word associations in a L2 seem to be driven by syntagmatic relations, translations from L1 words, and phonological similarities. In Chapter 6 I then reported a study in which distributional semantic techniques were used to model the organization of word meanings in the L1 and L2 respectively, and showed that the semantic representations in the L2 seemed to rely more deeply on word-to-word associations, compared to the semantic representations in the L1 (Bolognesi, 2011; 2016). As a consequence, the semantic representations emerging from word-to-word co-occurrences, modelled on the basis of a text-based distributional model, tend to be overall more similar to the semantic representations constructed by language learners than to those constructed by native speakers. This supports the linguistic nature of the semantic representation of word meaning in the L2, compared to the L1.

From a theoretical point of view, the idea that a foreign language learned in adulthood follows different principles for word meaning construction, compared to the principles that govern word meaning construction in the L1, can be easily motivated by a variety of factors that have been widely discussed in the literature. Most importantly, adult foreign language learners have already a fully developed linguistic system (their L1) and a fully developed cognitive system used for categorization purposes and for structuring word meanings and their mutual relations. In particular, the semantic representations constructed in the L1 during childhood are based on both word-to-world associations as well as word-to-word associations, as I discussed in previous chapters. Conversely, the semantic representations constructed in the L2, when the L2 is learned in adulthood, appear to be more language-driven, and therefore more strongly affected by word-to-word associations.

But how do word-to-word associations construct and explain the organization of L2 word meanings? As previously illustrated, words in the L2 seem to be connected to one another by means of: 1. Direct translations of associations established in the L1; 2. Associations between words that construct collocational patterns in the L2; 3. Associations based on the phonological similarity between two words (Meara, 2009).

The direct translations of associations established in the L1 allow non-native speakers to associate, in the mental lexicon, words that are paradigmatically related to one another in relation to their meaning. For example, adult learners of English as a foreign language may associate the word *dog* to the word *cat*, because such association holds between the equivalent words in their native language (e.g., between *cane* e *gatto* in Italian). Such semantic relation between *dog* and *cat* is arguably constructed in the L1 on the basis of word-to-world (as well as word-to-word) associations, thanks to which the speakers infer that there is a distributional similarity between the contexts (both linguistic and experiential) in which dogs and cats tend to appear. The paradigmatic relation between *cane* and *gatto* is then transferred in the English L2 by means of analogy: a lexical translation *cane*-to-*dog*, and a translation *gatto*-to-*cat*. This implies that the concept (and therefore the category) DOG and the concept CAT in the English L2 mirrors, by analogy, the conceptual representations that speakers constructed for these categories in their L1. Such categories are simply labelled with a different (but semantically equivalent) word, which is expressed in the L2. The relation between *dog* (English) and *cane* (Italian), in the example above, is based on a semantic relation which is comparable to the relation of synonymy: the two words share a large number of features.

The associations between words that construct collocational patterns in the L2 allow non-native speakers to connect word meanings that are syntagmatically related to one another. For example, adult learners of English as a L2 may associate the word *dog* with the word *bite*. These syntagmatic associations constitute the

first step for the construction of semantic representations organized on the basis of paradigmatic relations, which is the principle that governs the organization of the mental lexicon of native speakers. The fact that in L2 speakers the syntagmatic relations appear among the most frequently produced associations supports the idea that this associative mechanism based on word-to-word linkages is an intermediate step that eventually leads L2 speakers to structure word meanings in the L2 on the basis of paradigmatic (semantic) relations, as native speakers do. In other words, if the goal of L2 speakers is that of constructing semantic representations of L2 words that mirror those of native speakers, which are organized in terms of paradigmatic similarity between meanings, then the syntagmatic associations based on word-to-word linkages in the L2 are indeed a preliminary step to construct such organization in the mental lexicon. And the fact that syntagmatic relations determine the organization of word meanings in the L2 supports the cognitive foundations of these mechanisms, which start from syntagmatic associations to construct paradigmatic relations. This is the same mechanism predicated by the distributional hypothesis and implemented by distributional semantic methods. For example, a language learner that associates *dog* with *bite*, and *mosquito* with *bite* (both word pairs are based on syntagmatic, collocational relations) will eventually construct a paradigmatic similarity between *dog* and *mosquito*, based on the shared collocation *bite*, and will eventually learn that dog and mosquito are to some extent distributionally more similar to one another than *dog* and *umbrella*, or *mosquito* and *umbrella*, because *umbrella* does not share the syntagmatic relation with *bite*.

Finally, the associations based on phonological similarity allow L2 speakers to connect words in the L2 that have a similar phonetic/phonological structure. For example, given the word *dog*, L2 speakers are likely to produce words that have similar sounds, such as *fog*. How do these associations relate to the mechanism of word meaning construction based on the syntagmatic to paradigmatic switch, which I argue is a core principle that explains how word meanings are constructed and organized in the mind? The association between *dog* and *log* is paradigmatic, but such paradigmatic relation emerges not from shared features or shared linguistic contexts between dogs and logs. This peculiar type of paradigmatic relation emerges from shared phonemes between the word forms *dog* and *fog*. L2 speakers associate *dog* and *fog* because these two words share two out of three phonemes. The same type of phonology-based word associations has been documented in young children. Interestingly, the ability to detect statistical prosodic patterns in language input has been also documented in pre-lingual children (infants) and it is claimed to be the basic principle that leads young children to the discovery of phonemes and words in their native language (Kuhl, 2004).

To conclude, L2 speakers do follow the same cognitive mechanisms predicated by the distributional hypothesis in structuring word meanings in the L2. However,

they apply the steps (i.e., syntagmatic associations followed by construction of paradigmatic relations) in different ways compared to native speakers, as shown by the word associations that they provide. In particular, they construct associations between semantic representations of equivalent words in the L1 and L2 (e.g., *dog* in English and *cane* in Italian), based on the shared word-to-world associations (the two words are used to name the same referents). They construct associations between words that construct collocations in the L2 (e.g., *dog* and *bite*) because this is the first step of meaning construction, which precedes the organization based on paradigmatic relations between meanings. Finally, tthey construct associations between words with similar phonetic structures (e.g., *dog* and *fog*), based on the pattern of shared phonemes that these word forms display.

8.5 Summary

This chapter described in further detail the three steps that enable the construction and representation of word meaning, and their organization in the mind.

In relation to the first step, the associative mechanism, in this chapter I focused on how the associations between words and experiential contexts (word-to-world) and words and other words (word-to-word) contribute to the construction and organization of different types of words: words that denote concrete concepts and words that denote abstract concepts.

Then, I described how different types of speakers (native speakers vs. adult language learners) seem to rely on the mechanisms described in the three steps, to structure semantic representations for word meaning in the L1 and L2 respectively. By describing how the three steps described above can be applied to the construction of word meaning for different types of words and for different types of speakers, I provide evidence to support the cognitive foundations of the distributional hypothesis, which was described in Chapter 5, and widely used to implement computational, corpus-driven models of semantic representations by means of distributional semantics.

In the next chapter I will explain how the distributional hypothesis was largely misunderstood, and how its broader interpretation can help researchers move on, abandon the controversies related to the symbol grounding problem, and address new exciting research questions aimed at bridging embodiment and abstraction in a comprehensive theory of meaning construction and representation. Such endeavors require a plurality of approaches and methods, and require cognitive scientists, computer scientists, and linguists to join forces, to answer questions such as how humans evolve the ability to manipulate purely abstract symbols and, in turn, how AI's purely symbolic processing can be grounded in perception and action.

CHAPTER 9

The cognitive foundations of the distributional hypothesis

9.1 Leaving the Chinese room and climbing the ladder of abstraction

A decade ago, the distributional hypothesis was put under scrutiny by several cognitive scientists and cognitive psychologists, who were debating the embodied vs. symbolic nature of words and meaning in human cognition (De Vega, Glenberg and Graesser, 2008). Some crucial bottlenecks emerged during the debate, such as the processing and representation of meanings denoting abstract concepts (more recently discussed in Bolognesi and Steen, 2018) and the role that context and task conditions play in determining whether language processing relies on grounded or symbolic representations (more recently discussed by Zwaan, 2014, 2016). Although these aspects are key to understanding whether the grounded and the symbolic accounts of cognition can be integrated in a more encompassing theory of meaning, I believe that it is useful to take a step back and look at this theoretical debate from a distance, to understand why it seems to have reached an impasse and how such impasse can be overcome, thanks to the most recent developments achieved in cognitive/neuro- sciences and computer sciences/artificial intelligence.

Within the debate on the nature of word meaning, the arguments in favor of a symbolic nature of meaning are typically maintained by supporters of the distributional hypothesis, and criticized by supporters of the embodied/grounded views of cognition with the argument summarized by Harnad (1990) as the symbol grounding problem, previously exemplified by Searle with the Chinese room analogy (Searle, 1980), described in Chapter 6. Such critiques, however, are based on the assumption that the distributional hypothesis can be applied to words to model word meaning by looking at word co-occurrences only. However, this is not the case.

The distributional hypothesis is a general mechanism of cognitive processing that explains how meaning is constructed. This hypothesis can be applied to various types of associations, including associations derived from co-occurrences and missed co-occurrences of items in extra-linguistic contexts. In this book I focused on word meaning, and thus on word-to-world and word-to-word co-occurrences, but I briefly explained also how world-to-world co-occurrences may function.

The cognitive foundations of the distributional hypothesis and the concrete mechanisms that determine its implementation described in this volume find empirical support in large bodies of empirical literature on language acquisition, such as the studies reviewed in Chapter 2 on cross-situational learning and more generally on statistical learning, and the studies reviewed in Chapter 4 on incidental vocabulary learning. Moreover, indirect support can be found in the way in which the most successful computational models of meaning extension model polysemy, metaphor, and metonymy, which I reviewed in Chapter 3: these models are based on methods and algorithms that rely on the distributional hypothesis.

The mechanisms that underlie the implementation of the distributional hypothesis are supported by basic cognitive mechanisms widely documented in the empirical literature: the broadly defined associative mechanism (which is actually based on positive and negative feedback), the pattern detection, and the feature matching process, described in Chapter 4 and elaborated in Chapter 8. From a computational perspective, the applicability of the distributional hypothesis to extra-linguistic contexts has been described extensively in Chapter 7. Therefore, across these chapters, I discussed on one hand how the distributional hypothesis (originally adopted by computer scientists) is rooted in cognitive mechanisms widely supported by psychologists and cognitive scientists, and on the other hand how computational models that rely on the distributional hypothesis are flexible to the point that they can integrate extra-linguistic information in the construction of the word vectors, thus mirroring the functioning of the human mind in its ability to construct and organize word meanings. With these two operations, I have defended the cognitive foundations of the distributional hypothesis.

What can hardly be defended, instead, is the one-to-one correspondence between the specific engineering tricks (aka algorithms and formulas) used to implement the vector spaces and the cognitive mechanisms that the human mind undertakes to perform such steps. For example, the formula that determines the strength of the association between a word and its contexts of use (e.g., the Mutual Information formula), which is used to fill the contingency matrix in vector spaces (Chapter 5) can hardly be compared to the exact way in which humans construct the weights through which they evaluate the association between a word and its contexts of use. Similarly, the algorithms used to reduce the dimensionality of the contingency matrices in vector spaces, such as the SVD algorithm, can hardly be matched with a cognitive mechanism that works in the exact same way. In this sense, predictive models such as word embeddings, based on neural networks, which take into account positive feedback (co-occurrences) as well as negative feedback (missed co-occurrences) and update the association strength between two items based on different algorithms (e.g., the Rescorla-Wagner rule) are probably more likely to mirror human behavior.

However, I believe, this is not the interesting part of the equivalence between human and artificial mechanisms of word meaning construction. In the end, machines 'read' natural language by translating each soundwave (in spoken language) or visual symbol (in written language) into a number, while humans arguably do not do so. Therefore, there are inherent undeniable differences in the architecture of the human and the artificial minds in constructing and representing word meaning.

Leaving aside the obvious differences that characterize the human and the artificial mind allows us to focus on the general mechanisms that make the two systems comparable, and allows us to explore new interesting parallels that can help us gain a better understanding on how human and artificial intelligence respectively function, and how the functioning of each of these two systems can inform us on the functioning of the other one, in a synergistic manner. One specific example is the following: in language acquisition research it remains an open (and debated) issue whether within the cross-situational learning paradigm speakers really keep track of each word-object co-occurrence and gradually proceed, as parallel accumulators of statistical regularities, to disambiguate the correct word-object reference (Vouloumanos, 2008; McMurray, Horst and Samuelson, 2012; Yurovsky, Fricker, Yu and Smith, 2014), or whether instead they first form a single hypothesis about each word-referent association, and then proceed to test it in future encounters, in order to confirm it or reject it and form a new one, as in the *Propose-but-Verify* account (Medina, Snedeker, Trueswell and Gleitman, 2011; Trueswell, Medina, Hafri and Gleitman, 2013). From a computational perspective, both these mechanisms can eventually produce successful word-referent mapping, but they will do so at very different rates. The first mechanism requires a larger working memory capacity, because it keeps track of each and every experience in which a word is heard in a situational context. Each situation is processed once, and the new information is compared and contrasted online, with all the other possible word-object associations. The latter mechanism requires less working memory capacity. As Trueswell and colleagues (2013) explain, the initial association word-object is established by guessing at chance. Consequently, at every subsequent occurrence of the word, the guessed probability of association with a given referent is recalled and if the referent is actually present in the episode, the association word-object is strengthened, otherwise the connection word-object initially formed is dropped and a new object is selected at random from the context of the episode, to be associated to the word, and the process starts again. It remains however an open issue whether the subject reconsiders past events and past possible associations, to establish a new hypothesis, or whether instead she starts anew, as if she has never encountered the word before.

This and other issues involved in different types of statistical learning configurations, are hard to test in behavioral experiments, because it is difficult to control the input to which children are exposed, in order to test different theoretical models that would explain how the word-object associations are established. For this reason, often the behavioral evidence is collected using invented words and objects (e.g., Ramscar, Dye and Klein, 2013). However, it cannot be excluded that some resemblance between the invented words and existing words, or the invented objects and existing ones, may influence the associations established between words and objects, for example. An alternative strategy to the use of behavioral data to test different learning mechanisms is that of running computer simulations.

For example, as described in Chapter 2, Cassani and colleagues (2016) compare four theory-driven configurations of cross-situational learning, in their ability to predict behavioral learning data: a simple associative model that takes into account only co-occurrences to learn associations (reinforcement learning); a discriminative learning mechanism that learns from positive and negative feedback, using the Rescorla-Wagner updating rule; a single hypothesis model along the lines of the *Propose-but-Verify* account; and a probabilistic learner that computes a posterior probability distribution over referents for each word, updating the probability mass allocated to each referent in the light of new evidence. The authors show that the simple associative model based on co-occurrences alone as well as the *Propose-but-Verify* model fail to account for behavioral evidence. Conversely, the discriminative model based on the naïve discriminative learning algorithm, which uses the Rescorla-Wagner rule, and the probabilistic approach, both fit the behavioral data and they are therefore more likely to reflect actual human learning processes. In this sense, implementing different possible mechanisms of statistical learning processes in the artificial mind and running computational simulations, to see which model fits behavioral data, can inform us on the actual functioning of the human learning process.

The claim that the distributional hypothesis has cognitive foundations and it can be seen as a general mechanism for meaning construction and representation that start from associations between elements in experience, words with experiences, and finally words with other words, can also shed light on the general process of abstraction, "a central construct in cognitive science" (Barsalou, 2003: 1177). Despite being a core notion in cognitive science, the very definition of abstraction is non-trivial, and Barsalou, for example, identifies six different senses of abstraction. Among them, a broad distinction can be made between two senses in which abstraction seems to be intended: on one hand there is categorical abstraction (the distinction between specific categories such as TABLE and more generic categories such as FURNITURE), and on the other hand there is conceptual abstractness (the distinction between concrete categories such as TABLE and

abstract categories such as THEORY). As recently demonstrated by Bolognesi and colleagues (2020), these two phenomena are partially correlated, but theoretically distinct. Categorical abstraction is a process of categorization that involves both abstract and concrete concepts alike, by which "knowledge of a specific category has been abstracted out of the buzzing and blooming confusion of experience" (Barsalou, 2003: 389). Conceptual abstractness (or better, concreteness, as it is usually defined) is a variable that measures the degree to which the referent in the real world is associated with a specific concept that can be perceived through our sensory experiences. In this sense, while the concept TABLE is rather specific and concrete, the concept FURNITURE is more generic, because it defines a category expressed at a higher level of abstraction. The members that belong to such category, however, are still quite concrete (tables, chairs, lamps, etc.). Conversely, the concept THEORY is rather abstract. The question arises of whether THEORY is perceived to be also more generic (and therefore less specific) than TABLE, because of its higher degree of abstractness.

Although abstraction and abstractness are theoretically distinct, one describing the construction of conceptual categories starting from experiences, and the other describing the perceptibility of the referents designated by given concepts, they happen to be conflated with one another in some studies, precisely because of the polysemous nature of the very term *abstraction*. For example, Burgoon and colleagues (Burgoon, Henderson and Markman, 2013) provide an extensive overview of the different ways in which abstraction has been defined in the literature and then suggest their integrative definition of abstraction as "a process of identifying a set of invariant central characteristics of a thing" (Burgoon et al. 2013: 502). They then claim that abstraction operates on a continuum, in which:

> lower levels of abstraction (i.e., higher levels of concreteness) capture thoughts that are more specific, detailed, vivid, and imageable […], often encompassing readily observable characteristics (e.g., furry dog, ceramic cup; Medin and Ortony, 1989). Higher levels of abstraction (i.e., lower levels of concreteness), on the other hand, include fewer readily observable characteristics and therefore capture thoughts that are less imageable (e.g., friendly dog, beautiful cup).
>
> (Burgoon et al., 2013, p. 503)

In this integrative definition, the two variables that we described above are conflated.

The APA Dictionary of Psychology (Van den Bos 2007: 4), a valuable source of information within the communities of cognitive scientists and psychologists, provides three different definitions for the notion of *abstraction*. Of these, the first two suggest that the same label *abstraction* may be applicable to both, categorical abstraction and conceptual concreteness as they have been described above:

1. The formation of general ideas or concepts, such as "fish" or "hypocrisy," from particular instances.
2. Such a concept, especially a wholly intangible one, such as "goodness" or "beauty".[1]

While definition 1. seems to refer to the process of categorization, definition 2. refers to the perceptibility of the referent associated to a concept. In a recent attempt to clarify the relation between these two variables, Borghi and Binkofski (2014: 3) suggest that "concepts as ANIMAL and FURNITURE (on top of the abstraction hierarchy) are more abstract than DOG and CHAIR, but their category members are all concrete instances". Here the authors seem to suggest that abstract concepts are also more generic (i.e., higher on the taxonomy of categorical abstraction) compared to concrete concepts and that, as a consequence, we should be able to find a positive and significant correlation between abstractness and abstraction. This view can be motivated by the fact that generic categories (e.g., FURNITURE) are by definition more inclusive (Rosch, Mervis, Gray, Johnson and Boyes-Braem, 1976) and therefore less rich in defining features, specifically in perceptual features. Being low in perceptual features, such generic concepts might also be less tangible, or less concrete, and therefore more abstract.

Bolognesi and colleagues (Bolognesi, Burgers, Caselli, 2020) operationalized categorical abstraction by means of specificity measures extracted from the lexical taxonomy of WordNet[2] and conceptual abstractness by means of concreteness

1. Definition number 3 relates to conditioning, and defines abstraction as "discrimination based on a single property of multicomponent stimuli".

2. WordNet is a large lexical database of English words created in the Cognitive Science Laboratory of Princeton University in 1985. Entries cover the major parts-of-speech, such as verbs, nouns, adjectives and adverbs, and are organized via sets of synonyms, called synsets. Each synset represents a distinct concept – or, as stated by Miller (Miller, 1998), an instance of a lexicalized concept – and is inter-linked to other synsets through lexical and conceptual-semantic relations. Entries covering nouns in WN are primarily structured by two main semantic relations: (i.) synonymy, and (ii.) subsumption/subordination or hypernymy/hyponymy. The latter relation links more generic concepts to more specific ones (e.g., FURNITURE is a hypernym of a TABLE). The hypernymy/hyponymy relation, usually abbreviated as IS-A (e.g., DOG IS-A MAMMAL), is hierarchical, asymmetric and transitive: all properties of super-ordinate elements are directly inherited by their subordinate nodes. In the study conducted by Bolognesi and colleagues (2020) the measure of specificity was formalized as follows: if we imagine WN as an upside-down tree, in which the top root nodes constitute the most generic concepts, and the nodes at the very bottom of the tree (i.e., the leaves) constitute the most specific concepts, then the relative position of a concept within the tree (i.e., the number of nodes to the top root and the average number of nodes from the concept to each of the leaves) provides a good approximation of how specific a concept is, compared to all the other concepts represented in WordNet.

ratings extracted from established resources,[3] and compared the two variables on a set of 13,518 nouns. As a result, they found overall positive and significant correlations between the two variables (the more concepts are concrete the more they tend to be specific), although not particularly strong. The distribution of the four types of concepts, obtained by crossing the two variables, are displayed in Figure 22.

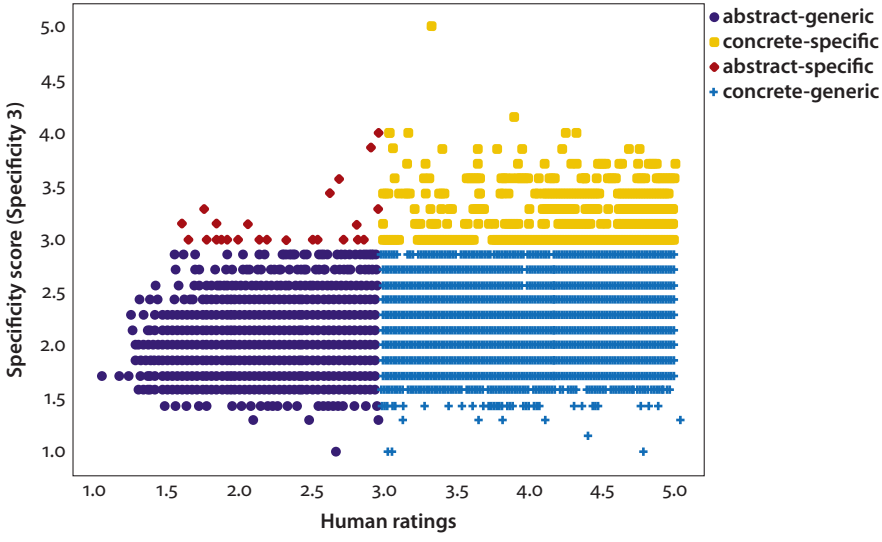


Figure 22. Extracted from Bolognesi et al. (2020). The distribution of 13,518 nouns across the four quadrants, obtained by crossing the variables Specificity and Concreteness. Both variables are measured on numeric scales from 1 to 5. The 4 colors used in this figure may give the impression that the variables are categorical (i.e., a concept is either concrete or abstract, either generic or specific). This is however not the case: The variables are continuous and the colors have been used to differentiate between more abstract, more concrete, more generic and more specific

In a subsequent qualitative exploration of the data, the authors analyzed the concepts found in each of the four quadrants: concepts that are generic and abstract, concepts that are specific and abstract, concepts that are generic and concrete, and concepts that are specific and concrete. The authors showed that prototypical abstract concepts are also highly generic (e.g., ABSURDITY, BELIEF, IDEA); that concepts that are abstract but specific seem to relate mainly to specific notions within the spiritual domain or to the socio-political one, thus belonging to the so-called 'social reality', which emerge thanks to social interactions of humans with other humans (e.g., FUNDAMENTALISM, MONOTHEISM, SUMMONS); that

3. The concreteness scores were extracted from the database of concreteness ratings released by Brysbaert and colleagues (2014).

typical concrete concepts are also highly specific (e.g., ASPIRIN, PEAR, SCALPEL); and that concrete concepts that are highly generic often denote concrete referents in the world that do not have clearly defined boundaries (mass nouns such as FOREST, PEOPLE, SEAFOOD).

Overall, the study proposed by Bolognesi and colleagues demonstrates and supports with a large-scale empirical analysis the idea that categorical abstraction and conceptual abstractness are two different (though related) phenomena. A crucial theoretical distinction between them is that while generic terms (or abstract categories) typically refer to groups of concrete objects that share functional but not sensory features, abstract concepts never refer to concrete objects but share different kinds of features altogether (Barsalou and Wiemer-Hastings, 2005; Wiemer-Hastings and Xu, 2005). Because generic categories (e.g., FURNITURE), compared to specific ones (e.g., TABLE), typically refer to concrete objects that do not share perceptual properties like shape, components or colors (think of tables, couches and lamps, within the category of furniture), the role that language plays in 'gluing' together the various members of the generic category is stronger than for specific categories, which in turn share also perceptual properties (think of instances of tables). Categorical abstraction, therefore, is a variable that is more deeply affected by language than conceptual concreteness which, in turn, is more deeply shaped by perceptual experience. In other words, language and of word-to-word associations seem to play a crucial role in constructing generic (vs. specific) categories, because such categories, like FURNITURE, encompass elements that do not share similarities that can be easily constructed on the basis of perceptual features.

A specific category[4] (e.g., the category TABLE), can in principle be constructed on the basis of world-to-world associations only: various instances of tables share perceptual entity-related properties (e.g., flat horizontal surface, vertical legs etc.) and associations with other objects (e.g., plates and cups are typically placed on tables, lamps might hang over them etc.). World-to-world associations are established between tables and table components, as well as tables and objects that co-occur with tables. On the basis of these similar patterns of associations, the various instances of tables are grouped together to form the (concrete and specific) category of tables. Conversely, for the construction of a generic category (e.g., the category FURNITURE) world-to-world associations are arguably not sufficient, because the category is not based on perceptual features shared by its members that allow us to cluster together lamps, couches, tables and book shelves.

4. Note that the variable categorical specificity is not binary but continuous. For the sake of simplicity, I am hereby referring to generic and specific categories, but the degree of genericity/specificity is relative to other categories rather than absolute.

Word-to-world and word-to-word associations play a more prominent for the construction of generic (vs. specific) categories.

As I will describe in the next section, generic categories play a crucial role in the processing and comprehension of figurative language (metaphor and metonymy). By virtue of the transitive property, if language plays a crucial role in the construction and representation of generic categories, and generic categories play a crucial role in the comprehension of metaphor, it follows that language (linguistic information) plays a crucial role in the processing and comprehension of metaphor.

9.2 The distributional hypothesis applied to metaphor

In Chapter 3 I provided an overview on how metaphor functions as a mechanism of meaning extension and illustrated some peculiarities related to the ways in which metaphor is typically expressed in texts (i.e., indirectly, as in *I devoured the book*, rather than directly in *x is y* form) and the way in which metaphorical expressions are processed. I also explained that there seems to be a crucial difference in the way conventional metaphoric expressions vs. novel metaphoric expressions are processed. In particular, when metaphors are conventionalized in language use, metaphorical word meanings are typically understood by means of semantic categorization, illustrated in Figure 23 (e.g., Glucksberg and Keysar, 1990; Giora, 1997; Glucksberg, 2001; Bowdle and Gentner, 2005). The metaphorical meaning gets categorized as element of a higher-level category selected by the literal meaning. Instead, for words used metaphorically in a creative way, the comprehension works by means of a comparison, in which the literal meaning needs to be activated, in order to be compared to the metaphorical meaning, and semantic features need to be projected from the literal to the metaphorical meaning (e.g., Bowdle and Gentner 2005).

Figure 23 shows how the metaphorical meaning of *invest* is understood within the conventional expression *invest effort*, according to the categorization view. In the case of *invest*, the hypernym (the generic category) of the literal meaning, inherited by the metaphorical meaning of *invest*, could be identified in the verb *put*. Both the literal and metaphoric meanings of *invest* entail a transfer of possession by which something that before was owned by the agent is (intentionally) *put* somewhere else, with the aim of obtaining something in return.

The categorization view therefore predicts that given the verb *invest*, its meaning can be understood by means of vertical alignment (as defined by Bowdle and Gentner, 2005) in which the common feature shared by the two meanings of *invest* is their hypernym *put*.

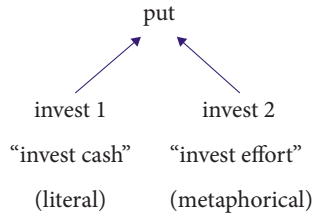


Figure 23. Categorization view of metaphor processing, based on a possible hypernym of both the literal and the metaphorical meaning of *invest*

Within the vertical alignment, the relation between the metaphorical meaning of *invest* and the hypernym *put* is the same type of taxonomic relation existing between specific and generic categories such as *table* and *furniture*. Both these vertical relations are strongly influenced by language. Moreover, the type of similarity entertained by the literal and the metaphorical meaning of *invest* is also linguistic: it is based on the fact that the two meanings share a common hypernym (*put*), or a common generic category, which is molded by language. Likewise, the type of similarity entertained by *table* and *couch*, for example, is based on the fact that the two meanings share the common generic category of *furniture*. Having a common hypernym, or a common generic category, pushes the comprehension of the metaphoric expression toward the activation of linguistic (rather than perceptual) information. Thus, the comprehension of the metaphorical expression *invest effort* arguably proceeds by means of word-to-word associations between *invest* and its hypernym *put*, and the similarity between the metaphorical and the literal meanings of *invest* derives from word-to-word associations that each meaning entertains with the common hypernym *put*.

Conversely, within the horizontal alignments that characterize the processing of novel metaphoric expressions, the activation of shared features between literal and metaphorical meanings, which are not yet lexicalized and therefore do not pertain to the linguistic system yet, depends on perceptual experiences and their mental simulation. For example, the less conventional[5] metaphoric expression *inject effort*, based on the literal expression *inject cash*, according to the Career of Metaphor is likely to be processed and understood by means of comparison, that is, by means of a horizontal alignment of the two meanings of *inject*, and the activation of inferences that highlight the features to be mapped onto the metaphorical use of *inject*. Such features can be identified for example in the physical force required to introduce (something) under pressure into a passage. This experience-based feature is mapped onto the metaphorical meaning of *inject*, enabling its

5. I am not claiming that this is a completely novel expression, but it is definitely less conventional than invest effort, according to corpus searches.

comprehension. The relation between the two meanings of *inject* and the mapping 'physical force required to introduce (something) under pressure into a passage' is based on word-to-world associations, that is, the association between the literal meaning of *inject* and the experience-based feature, and the metaphorical meaning of *inject* and the experience-based feature. If the metaphoric expression *inject effort* is used repeatedly to the point that it becomes highly conventional and fully lexicalized, then its processing moves from horizontal alignment (by comparison) based on word-to-world associations, to vertical alignment (by categorization) based on word-to-word associations between each meaning of *inject* and a common hypernym (generic category) shared by the two word meanings. In this sense, word meaning processing by means of word-to-word associations replaces word meaning processing by means of word-to-world associations when the metaphorical meaning becomes extremely conventionalized, that is, when the speaker acquires a high degree of expertise and familiarity with said word meaning. This view is consistent with the idea that word-to-word associations appear later in development, compared to word-to-world associations, as observed already in first language learners. Word-to-word associations, which in the case of metaphor comprehension are realized in the hierarchical associations between a metaphorical meaning and its hypernym as predicted by the categorization view, require the ability to combine abstract symbols (words) with one another, which develops at a later stage in first language learners, compared to the ability of combining symbols with objects or other elements pertaining to the realm of perceptual experiences. Similarly, the processing of metaphor by means of categorization (i.e., based on word-to-word associations) comes at a later stage and requires more expertise and familiarity with a given metaphor (which needs to become conventional), compared to the processing by comparison, based on word-to-world associations.

Thus, like for other types of paradigmatic relations of similarity between word meanings, also metaphorical similarity can be viewed as a relation between two meanings, constructed on the basis of the distributional hypothesis. Word meanings that are aligned in a metaphorical comparison (e.g., the two meanings of *invest*, or the two meanings of *inject*, in indirect metaphors) need to have something in common for them to be comparable, some latent piece of semantic information that allows the comparison to emerge and be meaningful. However, the similarity that characterizes two metaphor terms is not as consistent and stable as the similarity that characterizes, for example, two synonyms. In particular, while the similarity between two synonyms is arguably based on several common features, two meanings aligned in a metaphor might share only one or a few features. The nature of these shared features, which accounts for the similarity between two metaphor terms, is still debated. In a recent series of investigations Bolognesi (2016; 2017) and Bolognesi and Aina (2019) showed that applying the distributional hypothesis

to the modelling of metaphorical similarity, it is possible to highlight different patterns of features shared by metaphor terms. In particular, looking at a sample of linguistic metaphors and a sample of pictorial metaphors, identified and formalized into *x-is-y* statements through established procedures, the authors showed that the distributional similarity between metaphor terms aligned in linguistic and pictorial metaphors differs, when we look at the linguistic contexts (word-to-word associations), at the semantic features, and at the relational properties shared by the two terms (word-to-world associations). The linguistic metaphors, on average much more conventionalized than the pictorial ones, show higher distributional similarity when looking at the shared linguistic contexts, but lower distributional similarity (compared to pictorial metaphors) when looking at shared semantic features and related entities. Thus, while the metaphorical similarity that characterizes two metaphor terms aligned in typical linguistic metaphors is based on word-to-word shared associations, the metaphorical similarity that characterizes two metaphor terms aligned in typical pictorial metaphors, which tend to be on average more creative and less conventionalized than linguistic ones, is based on word-to-world shared associations, operationalized as semantic features (such as those collected by McRae, Cree, Seidenberg and McNorgan, 2005) and relational properties (such as those captured by FDT, Bolognesi, 2017a).

9.3 The distributional hypothesis applied to metonymy

Unlike metaphor, metonymy is a mechanism of meaning construction that does not imply the establishment of a paradigmatic relation of similarity between two terms. While in metaphors the metaphorical meaning of a polysemous word may be understood by means of its relation (and contrast) with the basic literal meaning, in metonymy there isn't a comparison between the literal and the figurative (metonymic) word meaning. When speakers refer to the war in Vietnam by simply mentioning *Vietnam*, in a sentence like *Vietnam is a shame for American history,* they use the name of the country to refer to a specific (but extremely salient) event that took place in that country. In this example, *Vietnam*-the-country and *Vietnam*-the-war are not compared to one another, with features belonging to the first meaning being mapped onto the second one. The semantic shift that allows the second meaning of Vietnam to emerge is based on contiguity, and in particular on an event-for-location type of metonymy.

It remains to be clarified how this mechanism of meaning construction relates to the general principle of word meaning construction that underlies the distributional hypothesis. In the case of metaphor, as argued above, all the three steps take place, with the semantic relation of similarity that characterize the metaphorical

and literal meaning of a word being based on a (broadly speaking) feature matching process that can involve different types of shared properties (linguistic hypernyms in the case of the categorization view or various types of semantic features in the case of the comparison view). In the case of metonymy, the situation is different. There is no feature matching process that constructs similarity between semantic representations, between the literal meaning of a word and its metonymic extension (e.g., the two meanings of *Vietnam*). Instead, the literal meaning of a word and its metonymic extension appear to be based on associations that remain on the syntagmatic level. Vietnam, as a country, is arguably associated with the war that took place there. Because metonymy is based on contiguity (syntagmatic relation, broadly speaking) rather than on similarity (paradigmatic relation), I argue that the first associative step that characterizes the meaning construction process based on the distributional hypothesis is the only one step that takes place.

Word-to-world and word-to-word associations by themselves seem to be able to explain the relations of contiguity between literal meanings and their metonymic extensions. In particular, one could argue that word-to-world associations characterize metonymies in which the metonymic meaning is clearly based on referential transfer, such as the circumstantial metonymies used within specific contexts. For example, in a sentence like *table 23 needs the bill*, the metonymic meaning of *table* (i.e., the customer sitting at that table) derives from a metonymic extension of the literal meaning of *table* (i.e., the concrete object). The association between these two meanings (the object table and the customer sitting at the table) is based on the fact that the two entities co-occur in experience (a type of relation which I called syntagmatic, broadly speaking). Conversely, word-to-word associations may characterize those metonymies in which the metonymic meaning is based on a lexical transfer, that is, metonymies based on logical polysemy. In Chapter 3 I exemplified these metonymies with the sentence *the gentleman began Dickens*, where Dickens stands for the books he wrote (author-for-book or more broadly producer-for-product type of regular metonymy), and the action expressed by *began* refers to the sub-event of writing, which is not expressed, and which is an action that can have *book* (also not expressed) as a typical object. In this case, the relation between the literal meaning of *began* (i.e., *to initiate*) and the metonymic meaning of *began* (i.e., *initiate the process of writing*) can be defined as a lexical transfer, because it is based on a taxonomic relation between a more generic and a more specific event. Being based on a relation recoverable from the lexical structures of the generative lexicon (Pustejovsky, 1991), as opposed to extra-linguistic knowledge retrievable from the context, logical metonymy works on the basis of word-to-word associations, while conventional (referential) metonymy works on the basis of word-to-world associations.

Metonymies based on referential transfers and metonymies based on lexical transfer are processed in different ways (McElree, Frisson and Pickering, 2006), and are also discussed typically in different disciplines. In particular, while computational linguists tend to talk about lexical transfers when analyzing metonymic shifts, and to model this type of metonymy (e.g., the work by Pustejovsky, 1995), cognitive linguists tend to talk about referential transfers when analyzing metonymic shifts (e.g., Littlemore, 2015). Future empirical work, therefore, may need to test whether the two different types of metonymy are indeed based on qualitatively different types of associations (i.e., word-to-word associations in the case of logical metonymy and word-to-world associations in the case of referential metonymy). I suspect that, as for metaphor, conventionality may also play an important role in determining whether a metonymy is processed by means of lexical or referential transfer (and therefore by means of word-to-word or word-to-world associations). In particular, it could be the case that unconventional metonymies are processed by means of referential transfer, while conventional metonymies commonly used in language, such as logical metonymies, may be processed by means of lexical transfers (and therefore by means of word-to-word associations, which require less cognitive effort). However, this observation remains speculative, awaiting empirical evidence.

To conclude, metaphor constructs a paradigmatic similarity between two meanings on the basis of (different kinds of) shared features between the two words, which are the two metaphor terms. Such relation of similarity evolves from the syntagmatic relations that each of the metaphor terms entertains with its features, which can be linguistic (e.g., hypernyms) or perceptual (e.g., components, shape, etc.). Conversely, metonymy relies on the (broadly speaking) syntagmatic relation of contiguity between two meanings, which are associated to one another on the basis of linguistic or perceptual contiguity. These associative relations on which metonymies are constructed do not evolve into paradigmatic similarities, as in metaphor. From a cognitive perspective, therefore, metonymy may be based on syntagmatic associations that, in principle, may be the starting blocks on which metaphors are then derived by means of pattern detection and feature matching processes. While this possibility and its investigation lie beyond the scope of this book, it should be mentioned that some scholars in cognitive linguistics and figurative language have defended precisely this position, showing that (at least some, if not all) metaphors are motivated by metonymies, or, in other words, that metonymy may be a conceptual prerequisite on which metaphor is then constructed (Barcelona, 2000; Kövecses, 2013). As a matter of fact, at least primary metaphors, which are defined as metaphors motivated by correlations in experience (Grady, 1997, 1999, 2005), such as AFFECTION-IS-WARMTH are based by definition on associations established between entities that co-occur in experience. Starting

from childhood, we experience parental love and care by means of physical proximity, which transfer physical, bodily warmth. The repeated association in experience between these two concepts connects them on the basis of world-to-world associations. However, if this association is considered to be based on a metaphor, it should be characterized instead by paradigmatic similarity between the two concepts, which derives from a feature matching process. Yet, it remains unclear what are the features shared by AFFECTION and WARMTH, on the basis of which metaphorical similarity can be constructed. I believe that such association, being by Grady's definition based on correlation in experience, which implies contiguity, remains at the syntagmatic level of world-to-world association, and therefore it should be labelled as a metonymic relation, rather than a metaphorical one. More generally, all metaphors that are based on contiguity in experience (i.e., on world-to-world associations, as primary metaphors are) should probably be labelled as metonymies, rather than metaphors.

9.4 The power of language as a driving force to abstraction

In the past three sections I have defended that the distributional hypothesis is much more flexible, cognitively motivated, and powerful than its mere application to the computational modelling of word co-occurrences. I have explained how such a hypothesis, grounded in established cognitive mechanisms, can be applied to extra-linguistic contexts, and I showed how the steps that are involved in its implementations relate to established mechanisms of meaning construction, such as metaphor and metonymy. By doing so, I argued in favor of the distributional hypothesis as a fundamental mechanism of meaning construction, on the basis of which we perform categorizations from instances of experiences and from word co-occurrences alike, and organize the resulting semantic representations in our mind. In this final section of Chapter 9 I would like to elaborate in greater detail the implications that such claims have for the grounding of language (and of word meaning) in perception and action. This aspect is particularly relevant for figurative language processing, because much of the research conducted on figurative language in the past 40 years relates to the theoretical framework fathered by Lakoff and Johnson (1980), the Conceptual Metaphor Theory, according to which metaphors are cognitive devices thanks to which we ground abstract concepts in perception and action indirectly, by mapping them onto concrete concepts that can be directly experienced through our bodies (Lakoff and Johnson, 1999). Supporters of this theory have provided numerous examples over the past four decades, both in terms of linguistic (corpus-based) analyses of conceptual metaphors (see Hampe, 2017 for a recent review) as well as in terms of psycholinguistic evidence

in which the activation of both source and target domains in the participants conceptual system is reported (Gibbs, 2006a, Gibbs, 2006b; Gibbs, 2011 for reviews). Within the Conceptual Metaphor Theory framework, metaphor is considered to be grounded in the sensorimotor system to the extent that the metaphor terms as well as the inferences that these concepts produce are based on associations that take place in the sensorimotor system (Lakoff and Johnson, 1999, p. 29). According to the Conceptual Metaphor Theory, human metaphorical thought is embodied in perception and action thanks to the mechanism of Embodied Simulation, that is, the recruitment of a type of representation that is directly anchored in perceptual experience and in bodily reactions to perceptual experience.[6] A classic example is provided by empirical studies showing that when we read a metaphorical sentence such as *to grasp an idea*, the activation of the neural circuit of "grasping" takes place in the premotor cortex, similarly to when we read a literal sentence like *to grasp a cup* (Boulenger, Hauk and Pulvermuller, 2009). This is commonly taken as evidence to support the argument that both meanings of *grasp* (the physical action and the abstract concept of understanding) are implemented in the same neural areas. However, as pointed out by Cuccio (2018), there are empirical findings showing diverging results, and suggesting that embodied simulations do not appear to be always recruited during metaphor processing. Some fMRI studies, for example, show that motor simulations are activated during literal uses of action language, but not during metaphorical uses (e.g., Desai, Conant, Binder, Park and Seidenberg, 2013; Raposo, Moss, Stamatakis and Tyler, 2009; Rüschemeyer, Brass and Friederici, 2007). Moreover, as Casasanto and Gijssels (2015) point out in a thought-provoking, rigorous and lucid article, ultimately there is no strong evidence demonstrating that even mental (primary) metaphors are embodied, and that the representation of the (concrete) source domains of many metaphors is grounded in perceptual and motoric neural substrates:

> there is, therefore, a Grand Canyon-sized gap between the strength of many researchers' belief in "embodied metaphors" and the strength of the evidence on which their beliefs should be based. (Casasanto and Gijssels, 2015)

The ultimate lack of empirical evidence supporting the embodied nature of metaphor processing, and supporting the processing of concrete concepts that act as source domains in metaphors, suggests that even concrete concepts can be

6. Cuccio conveniently distinguishes two types of Embodied Simulation: a narrow type, which mirrors the configuration of sensorimotor circuits automatically activated during the embodied processing of language, within the first 200 ms, and a broader type of embodied simulation, which develops beyond the 200 ms, and involves the activation of physical sensations. According to Cuccio, this distinction helps in making sense of the diverging evidence observed in relation to the embodied processing of figurative language.

processed bypassing embodied simulations. Concrete concepts, in other words, can be processed on the basis of semantic representations that recruit exclusively non-modality-specific brain areas. In Figure 4 (Chapter 2), the information encoded in these representations is illustrated in terms of 'linguistic information', which is then elaborated as word-to-word associations throughout this book. When word meaning denoting either a concrete or an abstract concept is processed by relying on word-to-word associations, then embodied simulations are not recruited. Conversely, when word meaning is processed by relying on word-to-world associations, then embodied simulations are arguably involved, because this type of processing relies on semantic representations constructed on associations between words and perceptual experiences. As argued by Barsalou and colleagues (2008), language-based representations (and therefore word-to-word associations) tend to peak first because they are less energy-consuming. Word meaning processing based on linguistic information can be followed by the activation of situated simulations, i.e., the recruitment of representations constructed on word-to-world associations. However, linguistic processing alone provides a "shallow heuristics that make correct performance easily possible. When the retrieval of linguistic forms and associated statistical information is sufficient for adequate performance, no retrieval of deeper conceptual information is necessary." (Barsalou et al., 2008, p. 249). Therefore, the processing of word meaning on the basis of word-to-word co-occurrences is a stand-alone cognitive strategy that can be adopted for processing concrete and abstract concepts alike, for literal and figurative language uses, without relying on deeper mental simulations. The recruitment of situated simulations, or embodied representations constructed on the basis of word-to-world associations is not a necessary requirement of language comprehension and word meaning processing. The linguistic system has evolved in such a way as to enable language comprehension and word meaning processing also when grounding is absent.

Semantic representations of word meaning derived from word-to-word co-occurrences, as widely discussed in this book, are based on the same principles that underlie the construction of semantic representations derived from word-to-world co-occurrences, but are 'lighter', in the sense that they require less cognitive effort, as argued by Barsalou and colleagues (2008). Being lighter, they peak first, and can often be sufficient to enable language comprehension. Within the classic definition of embodiment, which relies on the notion of Embodied Simulation and the recruitment of sensorimotor neural circuits, these language-based semantic representations are 'disembodied', stricto sensu. However, they do not come out of the blue: they are based on the same cognitive mechanisms that are used to generate grounded semantic representations (i.e., those representations that are based on word-to-world associations). The cognitive mechanisms that characterize the

construction of both types of representations are summarized by the distributional hypothesis. Last but not least, the neuroeconomic motivation for the emergence of these 'light' language-based representations may not be the sole motivation that explains their emergence and their use. In fact, on one hand linguistic representations peak first and require less cognitive effort, compared to semantic representations that recruit deep embodied simulations. On the other hand, linguistic representations encode linguistic information which does not necessarily mirror the information encoded in perceptual experience. Therefore, depending on the task at hand, the context of use, the speaker's intentions and communicative goals, relying on language-based representations may be a strategy that enables speakers to attract listeners' attention on specific aspects of word meaning that are better encoded in language than in experience. Conversely, relying on embodied representations of word meaning may be necessary and strategically useful in those cases in which information that is more clearly encoded in perceptual experience (rather than in language) needs to be recruited to perform a given task.

To conclude, the same word affords (at least) two different types of semantic representations: one based on its co-occurrence with other words, and another based on its co-occurrence with elements in experiential contexts. Therefore, it is misleading to talk about grounded or symbolic processing of word meaning in general. Any word, in principle, can be processed by activating its grounded or its symbolic representation, depending on the context, on the task to be performed and on the speaker.

9.5 Summary

The distributional hypothesis has deep cognitive roots. This hypothesis has been largely misunderstood by cognitive scientists who limited its usability and applicability to the modelling of word-to-word co-occurrences only. As a general mechanism, the distributional hypothesis consists of three steps, which are all documented in cognitive science as basic principles of human cognition: the associative mechanism, the pattern detection, and the similarity construction by means of feature matching. Embracing the parallel between the mechanisms involved in the distributional hypothesis and the mechanisms used by the human mind to construct and organize word meaning enables us to develop new inferences that can shed light on how humans and machines (AI) respectively function.

The different types of associations that are involved in the first step of the implementation of the distributional hypothesis can shed light onto the mechanisms of abstraction, a hallmark of human cognition. In this chapter, in particular, I distinguished between categorical abstraction (the construction of generic categories

starting from specific ones) and conceptual abstractness (a variable that determines the degree of perceptibility of the referend denoted by a word meaning). While these two variables are positively correlated, they capture different phenomena, and while categorical abstraction is tightly related to language and to linguistic processing (based on word-to-word associations), conceptual abstractness is tightly related to perceptual experience (based on word-to-world associations).

In the second part of the chapter I focused on the role that word-to-word and word-to-world associations play in the construction of the similarity that characterizes two meanings aligned in a metaphor and to meanings aligned in a metonymy. For metaphor, I described how the categorization view of metaphor comprehension affords the construction of metaphorical similarity based on word-to-word associations, while the comparison view of metaphor comprehensions affords the construction of metaphorical similarity based on word-to-world associations. Then I explained that metonymy works in a different way: the literal and the metonymic senses of a polysemous word are not compared to one another (as for meanings aligned in metaphors). The metonymic relation between two meanings of a polysemous word is based on syntagmatic associations rather than on paradigmatic similarity like in metaphor. In particular, while word-to-world associations may motivate the relation between literal and figurative meanings in referential metonymies (i.e., metonymies based on a referential shift), word-to-word associations may motivate the association between word meanings in lexical metonymies (i.e., logical metonymies).

Finally, I related the mechanisms underlying the distributional hypothesis, and in particular the associative mechanism that can be configured into different types of associations, to the cognitive grounding of word meaning. On one hand I described in what sense language (and the semantic representations based on word-to-word associations) allows us to process meaning without necessarily relying on deep embodied simulations. On the other hand, I explained how embodied simulations (and therefore the activation of modality-specific neural circuits dedicated to processing perception and action) enables us to access aspects of word meaning that are related to experience and encoded in word-to-world associations.

In the final chapter of this book, I will focus on the practical implications that arise from the acknowledgement that the distributional hypothesis has cognitive foundations and explains how the human mind constructs, represents and organize word meaning.

CHAPTER 10

Conclusions and outlook

10.1 AI behaviorism: Learning how the mind constructs word meaning by looking at how machines do it

> Computation is still the best, indeed the only, scientific explanation we have of how a physical object like a brain can act intelligently. (Allison Gopnik, 2015)

Throughout this book I defended the idea that the distributional hypothesis, on which several computational models of word meaning are based, has solid cognitive foundations. That is to say, I used empirical evidence reported in the field of cognitive science to defend and support a computational hypothesis of word meaning construction and representation. The two disciplines that I bridged with this cross-disciplinary work are therefore cognitive science and computer science, which I compared in their ways of addressing the question of how and where words get their meaning, a topic that is highly relevant in linguistics and communication sciences.

The two architectures compared across these two disciplines are the human mind and the artificial mind implemented by AI, where AI is a general term used to label that field of research that aims at constructing computer intelligence. Because one of the main ways in which intelligence is manifested is the ability to learn, one of the main challenges of AI is creating artificial intelligence that can learn. The branch of AI that deals with this specific ability is called machine learning, and in recent years a specific application of machine learning that uses neural networks, called deep learning, has gained popularity among computer scientists and computational linguists. Deep learning is a family of machine learning algorithms that are capable of learning (among other things) word meanings. Such algorithms are called 'deep' because they rely on various (and increasingly more abstract) levels of representations, constructed on the basis of raw (linguistic) input. The various levels of representation on which deep learning relies to construct word meaning are the hidden layers within the neural network, which consist of features. In Chapter 5 we observed the functioning of word embeddings, which are implemented with shallow neural networks, which have only one layer of hidden features. Without having previous knowledge regarding a word's meaning, a neural network automatically generates the identifying characteristics of that

meaning, by looking at the instances of word occurrences, and weights their relevance for the meaning construction. The resulting representation is a vector of weights, which together profile a word meaning, also defined as word embedding. Relations between word meanings are then determined by comparing the relative word vectors, as in classic distributional models that rely on the distributional hypothesis.

Despite their ambiguous name, neural networks are computing systems, rather than biological systems, that are only inspired by the functioning of biological networks of neurons. Sometimes, they are called *artificial* neural networks to disambiguate their computational nature, but more often they are simply called neural networks. The original goal of the neural networks' approach was to solve problems in ways that mirrored human brains not only in terms of mere outputs but also in terms of structure and functioning mechanisms, assuming that there was a (metaphorical) equivalence between the way in which the networks of neurons function in the human brain and the way in which information is transferred between the nodes of an artificial neural network. However, the equivalence between the human and the artificial architecture of such networks has been criticized to the point that in the past 50 years the two disciplines (cognitive/neurosciences and computer sciences/AI) grew apart and the achievements obtained within each field have been rarely discussed in relation to the other discipline. The lack of communication between the two fields of research was then sealed by the increasing popularity of the grounded and embodied accounts of cognition, which became mainstream in the Nineties. According to these theoretical cognitive frameworks, a bodyless artificial mind like a computer could never properly mirror the functioning of the human cognitive architecture, because this one relies heavily on the information recruited through the body. Research on (artificial) neural networks, word embeddings and other computational systems of word meaning construction and representation, therefore, proceeded on a parallel track, without attempting to establish a mirrored image of how the human brain and mind construct and represent word meaning. One of the most prominent exceptions is represented by the introduction of Latent Semantic Analysis (Landauer and Dumais, 1997), the pioneering computational model of word meaning based on the distributional hypothesis, discussed in this book. The claim that Thomas Landauer and Susan Dumais made, proposing LSA as a cognitively legitimate theory of meaning, was heavily discouraged by cognitive scientists and neuroscientists (as described in De Vega, Glenberg and Graesser, 2008). In the past 10 years, however, the outstanding achievements and performances obtained by computational models of word meaning and language processing within the fields of distributional semantics and neural networks prompted a new wave of curiosity toward the actual possible bridges that could be built between the two disciplines: computer science

and computational linguistics on one hand, and cognitive science and cognitive psychology on the other.

In this book I explained how this new wave of mutual interest between the two disciplines manifested, and how the empirical findings achieved within each one of these two communities informed the other community, and enabled new exciting discoveries in various subfields, such as first language acquisition (e.g., the cross-situational learning mechanism), second language acquisition (e.g., incidental vocabulary learning mechanisms), word meaning extension and processing (e.g., comprehension of figurative meaning), distributional semantics (e.g., the integration of extra-linguistic information to construct multimodal semantic spaces).

To briefly summarize the rationale behind the structure of this book, in the first part I focused on the open challenges and problems within first and second language acquisition research, in relation to the acquisition, construction and representation of word meaning, and the organization of word meanings in the mind. These were followed by an overview of the theoretical and computational models that have been suggested to address such open issues. In the second part of the book, I visited the other field of research, delving into the mechanisms of meaning construction and representation that characterize the distributional view. I explained how word meanings are constructed from scratch, on the basis of their associations with other words as well as with extra-linguistic entities. Then, in the third part of this book I elaborated the converging evidence that emerged from the two fields, and claimed that the distributional hypothesis, the cornerstone of many contemporary computational models of word meaning, has solid cognitive foundations and, when applied to linguistic as well as extra-linguistic contexts, can explain how humans construct and represent word meaning. In this process, abstraction is a key component that enables the construction of categories of items and relative conceptual representations. The way in which humans are capable of extracting and constructing meaning starting from the great blooming and buzzing confusion of perceptual experience is a complex process that necessarily requires abstractions, categorizations, and classifications. As I argued, language is one of the main and most powerful tools we have to perform such abstractions.

In an evolutionary perspective, cognitive historian Jeremy Lent recently suggested that the ability to abstract and construct meaning that is detached from everyday sensory experiences was a major break-through in human evolution, and that such evolution was enabled by the appearance of language. In his recent award-winning book on the cultural history of humanity's search for meaning, Lent (2017) suggests that the emergence of human language and of specific linguistic features played a key role in determining this cognitive leap forward in human evolution. For example, the fact that ancient Greek had definite articles, which were absent in many other ancient languages, allowed speakers to construct

abstract concepts from concrete ones: from the adjective *good*, used to describe sensory properties of objects, Greek speakers could derive *the Good*, a much more abstract concept. Lent argued that the ability to abstract, enabled by language, led the pre-Socratic thinkers to develop the ideas that constitute the basis of human Western thought, and eventually led to the Scientific Revolution. His work is partially inspired by previous work conducted by psychologist William Noble and anthropologist Iain Davidson (1991, 1993), who argued that the emergence of language in human evolution enabled a much more sophisticated flow of information and generation of new ideas that led to the crescendo of the upper Paleolithic, in which art, music, religion, and tools construction are first documented. All these complex forms of expression require a cognitive architecture capable of abstracting representations from individual experiences, a cognitive ability that was boosted by language.

Compared to humans, computers are faster and more accurate at performing generalizations from large amounts of raw data. Such mechanisms are exemplified, for example, by the subroutines, which are pieces of code written to perform very basic operations. In order to make computer programming more efficient, early scholars in computer sciences (e.g., Wheeler, 1952) developed several subroutines, which could be labelled and stored in a computer library. Then, the scientist would program a computer to translate the label of the subroutine into the list of operations coded within the subroutine. When writing a new piece of code, then, subroutines would be embedded within the new piece of code by simply mentioning their label. This basic process of abstraction thanks to which a machine reads a label and runs the list of (more specific) operations described enabled the sophisticated developments achieved in the next decades in machine learning, which exploit language statistics in a bottom-up manner, to abstract word meaning representations. For both humans and machines, abstraction is a cognitive operation that involves language.

To conclude with a thought-provoking observation, the parallel between the functioning of the human mind and the artificial mind seems, in recent years, to have reached a tipping point: while scholars in computer science traditionally aimed at the development of computational models of the human mind that were inspired by the actual functioning of human mind and brains, recently this tendency might have overturned. As a result of the lack of mutual communication between the two communities, and facilitated by the exponential growth of available computer power, the achievements reached within AI and machine learning are today the objects of study of cognitive linguists and cognitive scientists. To put it more simply, the impressive outputs obtained by contemporary computational systems such as neural networks (and therefore also word embeddings) which are able to learn without any prior instruction and in ways that are still not completely

clear to humans, stimulated the curiosity of scholars working on human learning to try to understand and test the mechanisms employed by the artificial systems to solve problems that humans are typically faced with. This could be the beginning of a new era that I would call AI behaviorism, in which cognitive scientists, linguists and philosophers may focus on how AI works, in order to gain a clearer understanding of how the human cognitive system works, exploiting the analogy with the artificial cognitive system, constructed by artificial intelligence. Nevertheless, what is still an open and critical issue in AI research is precisely the ability to transfer what is learned in a specific domain onto another domain. In other words, the critical bottleneck of AI learning abilities is that of becoming capable of abstracting knowledge and, by analogy, learning to apply what has been learned in one domain, onto another domain (Mitchell, 2020). This implies the ability to perceive a degree of sameness at some level of abstraction between two meanings, or two tasks, and thus applying by means of analogical reasoning, knowledge and features from the known domain onto the new one. Thinking through analogies is the barrier that AI will need to crack, in order to enable what is called "transfer learning", that is acquiring new information on the basis of already known mechanisms, and learning new ways of learning, by analogy with already known ways of learning.

10.2 Practical implications for the study of human creativity

What is the relation between the construction of word meaning representations described in this book and the creative linguistic behavior manifested in many linguistic utterances, in which words are used in non-canonical ways? Constructing categories and semantic representations in the way suggested by the distributional hypothesis and described throughout this book is a way of learning. Learning can be seen, in a way, as a process that hinders creativity. Learning implies forgetting about all the possible options that could potentially exist, to focus on one 'correct' solution. Learning can be seen as leaving the exploration of multiple possible associations between two entities, two ideas, two objects or two words, and focusing on exploiting one single association, which is the target of the learning process. For example, children in kindergarten tend to draw human figures using all the possible colors available, including green, purple and blue. Slowly, then, they learn to limit the range of colors used to draw human faces to the colors afforded by human pigmentation. Similarly, episodes of overextensions of word meanings are commonly found during language development: children may overextend animal names such as *dog* to all animal species with four legs, including cats and goats. This is a creative behavior, because it establishes unconventional (thus creative)

associations between features and entities, or words and entities. When children learn the correct association between a word and a referent, the other associations gradually get pruned and disappear. Thus, the process of learning conventionally accepted categories within a linguistic community implies a loss in creative behavior, which is necessary to favor the strengthening of conventional (learned) associations.

Despite this initial loss in creative behavior, which is necessary for the construction of categories and word meaning representations that can be used to conduct successful social interactions within a community, creative behavior and creative linguistic constructions appear also later in development and throughout adulthood. In particular, creative uses of language and of words in particular, can be commonly found in many contexts, ranging from natural conversations, to slogans, political speeches, and even academic texts. A typical way in which words are used creatively is by embedding them in a context in which they are not usually found (Veale, 2012). As explained by Veale, even a very creative metaphor (or simile), such as the one used by designer Karl Lagerfeld to describe sunglasses, which he defined *burqa for my eyes* can be seen as a categorization. In line with the Aristotelian view of metaphor, Veale explains that this simile can be understood by identifying the structure of a category that embeds both items, the sunglasses and the burqas. This is the general category of WEARABLE OBJECTS. Within this category, burqas and sunglasses, which are quite far apart from one another in terms of shared core features, although they both belong to the category of wearables, are forced into a closer proximity that on one hand highlights their membership within the category of wearables, and on the other hand invites the activation of features that belong typically to burqas, which are mapped onto the sunglasses, such as the sense of protection from external viewers when walking in public places and the related ability to see without being seen.

While, technically, burqas and sunglasses already belong to the generic category of wearables and their proximity is only strengthened by the figurative linguistic alignment, it is possible to use language creatively in such a way that an entity is forced into a category in which it would not be found at all. For example, imagine the generic category FURNITURE, used in various examples in previous chapters. By saying *the room was furnished with a table, two chairs, a couch, and Nora's smile*, the reader is forced to interpret the meaning of *smile* as a member of the category FURNITURE. As a consequence, *smile* acquires (temporarily) features that characterize the members of that category, such as their function, which is that of making an environment comfortable, welcoming, or suitable for living. Such creative construction of the meaning of *smile* is driven by language, and relies on the creative integration of a word meaning within a non-default generic category. For this type of creative construction, learning the default meaning of

the category FURNITURE is a necessary step for the reader, because she needs to be acquainted with the features that construct such a category, in order to be able to integrate the 'alien' member (in this example, *smile*) within said category. Creativity, therefore, is in this sense a cognitive operation driven by language, which enables the meaningful integration of an entity within a non-typical (previously learned and established) category.

In a similar way, it is also possible to force the creative construction of ad-hoc categories that include two different word meanings, by aligning them in a metaphor or simile. For example, reading the creative metaphor *a lawyer is a lighthouse*, the reader needs to construct an ad-hoc category based on which such comparison makes sense. Such ad-hoc category, forced by the linguistic alignment of two meanings, may be labelled as *entities that can warn us about risks*. In this example, the category is creatively constructed, while in the previous example a member is creatively included in an established and conventional category, and in the very first example two entities that can be seen as members of the same generic category (but they are far apart) are pushed in closer proximity. Thus, linguistic creativity can affect word meaning construction and interpretation at different levels and in different ways. The mechanisms that underlie such construction are the same mechanisms involved in the construction and representation of word meaning described in this book and supported by the distributional hypothesis: associations between words and experience, or words and other words, pattern detection and similarity construction, based on feature matching processes.

In this sense, generalizations are key for the creative construction of word meanings, because the processes of abstraction involved in the construction of generic categories by definition leaves aside perceptually-derived details, which instead characterize the more specific categories. Being less characterized by perceptual features, generic categories allow the creative integration of new members within the category. By generalizing, we construct categories that are more ambiguous than specific categories, and being so, they can be (creatively) applied to new meanings. Therefore, categorical abstraction, a cognitive process enabled and led by language, is crucial not only for learning but also for creative thinking and creative constructions of word meanings.

An interesting empirical question that arises in relation to the study of human creativity, is whether the different ways in which new word meanings are creatively constructed, described above, afford different types of cognitive processing and involve different neural circuits. Moreover, it would be useful to know whether the different ways to construct word meanings creatively, based on the different types of categorizations described above, correlate with different cognitive styles classified in the literature. This could help educators and professionals developing creative thinking in pupils and students even in an institutional school setting,

thus helping to fight the trend described by educationalist Sir Ken Robinson,[1] who recently claimed in his TED talk that "schools kill creativity", and argued that "we don't grow into creativity, we grow out of it. Or rather we get educated out of it".

10.3 Practical implications for the study of first language acquisition

In the previous section I anticipated how the overextensions performed by children during the early stages of language acquisition can be seen as creative processes of meaning constructions, which are then pruned off once they learn to configure categories in the way that is conventionally shared by their linguistic community. The creative behavior that underlies the initial overextensions can be seen as a process based on probabilistic inferences that I am going to discuss here.

The construction of a category on the basis of word-to-world and word-to-word associations, as described in previous chapters, is fundamentally a bottom-up process that does not require rules or constraints imposed in a top-down manner. However, it can be argued that some top-down operations do take place in the process of category construction. These operations are the inferences that children make when they guess the meaning of a word by trying it out in a new context, to test whether the category in which the meaning was constructed is acceptable in natural communication. While in the previous section I used an example of overextension involving a concrete concept (e.g., naming various animals with the same label *dog*, to test the acceptability of the corresponding referents within the category of dogs), let me now provide an anecdotal example with a word denoting an abstract concept. A few weeks ago, I asked my 5 year old son whether he was hungry and whether he wanted to eat some…butterflies. As a matter of fact, I asked him in Italian whether he wanted to eat *farfalle* (in English, butterflies) when I meant to say *fragole* (strawberries): a simple slip of the tongue. He laughed. I explained that I mixed up *fragole* with *farfalle*, and used the wrong word to phrase my question, and he concluded that what I did was tell him a *lie*. For him, using the incorrect word in a given context was a way of lying. He was testing his own definition for the category LIE, trying to include within that category the recent communicative experience with my slip of the tongue. Clearly, the meaning of *lie*, for him, was generically speaking something 'that is not true', something 'to be corrected'. Within this generic category, my slip of the tongue would have found a good fit. Unfortunately, he was overextending the meaning of *lie*, or otherwise stated, his construction of the meaning of *lie* was based on a category that was too generic: it was overextended. This example shows that once a category is constructed, on

1. https://youtu.be/iG9CE55wbtY

the basis of word-to-world and word-to-word associations (e.g., the meaning of *lie*) children start making inferences, aimed at using the newly constructed word meanings in natural communication. These attempts can be seen as tests that children use to verify the category they made, and to possibly enrich such category (such meaning) with new associations (word-to-world and word-to-word linkages). Had I confirmed my son's intuition that the meaning of *lie* is applicable to the description of slips of the tongue, the category that he constructed for the concept LIE, on the basis of the common feature 'wrong things to be said in a given context', would have acquired a new member, namely the type of communicative experience we just had. And such experience would have arguably enriched the category of LIE by bringing within that category other features that characterized our communicative experience. For example, the funny aspect involved in the communicative experience we just had could have become a characterizing feature of the category LIE. If this feature was repeatedly encountered across various experiences, it could have become a core defining feature of lies.

Overall, the inferences that children make when they use a newly constructed category in a novel situation, and in particular the inferences that they make when they use a newly constructed word meaning in a novel linguistic context, suggest that children engage in probabilistic logic that allows them to handle the uncertainty they have toward the exact word meaning, and at the same time combine such uncertainty with the prior knowledge they aggregated from (limited) associations that they collected, which involved that word. This probabilistic approach resembles the Bayesian inferencing model, according to which the probability for a hypothesis (in this case, a word meaning) is based on the expectations constructed on a state of knowledge (the limited amount of experience that children have with a word), and expressed as a probabilistic measure instead of an exact frequency measure. A core aspect of such approach is that the inferences (and the related word meaning representations) are updated as more evidence or information becomes available. Interestingly, probabilistic approaches to word meaning construction are used in word embeddings to express the weights that characterize the dense word vectors, and they are used (in a different way) in classic distributional semantic spaces, in the measures of associations[2] used to weight the strength (or entropy, or amount of shared information) between a word and a contextual entity.

2. For example, as described in Chapter 5, the Mutual Information between two variables (two words, or a word and a context) measures the mutual dependence between them, by weighting how much knowing one of these variables reduces uncertainty about the other. This is not technically speaking a probabilistic measure, but nonetheless it describes a relation that encompasses not only the raw co-occurrence frequencies but also the missed co-occurrences (that is, the occurrences of each of the two items alone)

This suggests another parallel between the human and the artificial minds, based on the probabilistic approach used to construct word meaning in distributional modes and word embeddings and the probabilistic approach in which children engage when they try out newly constructed categories and word meanings.

Another interesting empirical question that arises from the parallel between the human mind and the artificial mind in their process of word meaning construction is whether the unlikely inferences that children produce when testing a newly constructed word meaning are the same that an artificial mind produces. As a matter of fact, the performance of artificially intelligent systems of word meaning construction is typically evaluated by comparison with the 'correct' way, set by humans who were performing the same task. This does not allow us to test whether the way in which computational systems 'make mistakes' is the same way that humans show, for example when they overextend newly constructed word meanings. Comparing the way in which the two systems 'make mistakes', instead, would provide further argument to support the comparability between the two learning systems.

Finally, the discovery that both animals and humans learn not only from positive feedback (i.e., observed co-occurrences) but also from negative feedback (i.e., missed but expected co-occurrences) has important theoretical implications on the general mechanisms of language acquisition. In particular, the fact that the association between (for example) a word and a referent can decrease in strength and can even be unlearned, if enough negative feedback is provided (see Rescorla-Wagner rule), contrasts with the long established logical problem of language acquisition (LPLA, see Baker 1979), as recently explained by Ramscar, Dye and McCauley (2013). The LPLA argument is based on the observation that children make systematic grammatical mistakes (such as over-regularizing past tenses and plurals) and typically they are not corrected explicitly by adults. Given this "poverty of the stimulus" as Chomsky defined it (Chomsky, 1980), children could not learn the correct word forms from experience alone, but must have innate constraints on what is learned (e.g., Chomsky 1980; Pinker 1991, 1999, 2004). However, as Ramscar and colleagues (2013) demonstrate, children can learn to correct themselves solely on the basis of evidence available in the environment, thanks to the mechanisms of discriminative learning. Therefore, behavioral evidence collected using discriminative learning mechanisms and commonly implemented in neural networks, show that the LPLA argument does not hold.

10.4 Practical implications for learning and teaching a foreign language

A crucial aspect that was investigated throughout this book was the difference between the meanings of words denoting concrete and abstract concepts respectively. I argued and explained that abstract concepts, compared to concrete ones, are more deeply shaped by word-to-word associations, that is, by linguistic information, for their lack of direct referents in the world that can be uniquely associated with them. It follows that abstract word meanings are more deeply affected by linguistic variability. In other words, the meaning of abstract concepts is arguably more varied than the meaning of concrete concepts across languages. The meaning of abstract words relies mainly on word-to-word associations, and such associations vary greatly across languages because they depend on the structures of that given language and the collocations typically found therein. Conversely, the meaning of concrete words relies to a greater extent on word-to-world associations, which are arguably less varied across languages. As a matter of fact, even though the range of experiences that humans may perceive is incredibly vast, the fact that we share similar bodies and similar perceptual systems through which we approach such experiences limits the variability of word-to-world associations that can be afforded by our cognitive system. This has been empirically tested in a recent study conducted by Vivas, Montefinese, Bolognesi and Vivas (2020), in which the authors show through a property generation task that the semantic representations of concrete words, emerging from the features generated by native speakers of English, Spanish and Italian about the underlying concepts, are relatively stable across languages. For example, the way in which humans construct the meaning of the word denoting a fork through word-to-world associations is arguably quite similar because we tend to associate forks with features and other entities that we experience together with forks. For example, we all perceive forks to have an elongated shape, and pointy prongs. Similarly, we typically use forks for eating and thus these objects appear often together with food, bowls and plates, possibly tables, and so on. All these entities become associated to the word that denotes a fork, across different languages. As a consequence, the meaning of *fork* (English), *forchetta* (Italian), *Gabel* (German), *vilka* (Russian), *kaanta* (Hindi), and *çatal* (Turkish) is relatively similar and quite easily translatable across languages. Conversely, the meaning of a word denoting an abstract concept, such as *goodness*, is hardly translatable in a straightforward way across languages, because its translation depends very much on the linguistic context in which the word *goodness* is used. By simply looking at how this word can be translated through Google translator, this phenomenon becomes clearer. *Goodness* can be translated into Italian as *bontà, cortesia*, or *generosità* (among others) depending on the context; in German, as *Güte, Tugend, Qualität* or *Nettheit*, depending on the context;

in Russian as *dobrota, dobrodetel'*, or *velikodushiye*, among other translations; in Hindi as *bhalaee*; and in Turkish as *iyilik, cevher*, or *öz*. The greater variability of possible translations for words denoting abstract concepts across languages, compared to words denoting concrete concepts, can also be tested by checking the mutual translatability of the words used in this example: *fork* and *goodness*, using Google translator. For example, by translating *fork* into any of the equivalent words in the languages listed above, and then back into English, we obtain *fork* again. Conversely, by translating an abstract concept into another language and then back to English,[3] often the translation lands onto a different word, compared to the original one. For example, *goodness* translated into German and back into English becomes *quality*. When translated into Russian and back into English, *goodness* becomes *kindness*. The translation into Turkish and back into English returns *favor*. And so on. Moreover, as briefly described in Chapter 2, words denoting concepts that describe emotions, or feelings, are particularly abstract and they vary greatly across languages, with many languages displaying lexical gaps when it comes to direct translations. For example, the English word *awkwardness*, according to SketchEngine, appears together with words such as *embarrassment, shyness*, and *nervousness*. In Italian, the word *awkwardness* is Google translated in various ways, such as *imbarazzo* or *goffaggine*. These words, however, do not capture the same meaning that *awkwardness* captures in English, because they tend to be associated, in Italian, with other words such as for example with *vergogna* (*shame*), which does not appear to be that close to *awkwardness* in English. In German *awkwardness* is translated in multiple ways too, including *Unbeholfenheit* or *Verlegenheit*, but such translations are not reciprocal either: *Unbeholfenheit* translates back to English mainly as *ineptitude*, while *Verlegenheit* as *embarrassment*. In Russian *awkwardness* can be translated as *nelovkost'* or *neuklyuzhest'*, where both translations are formed with a negative prefix that underlies the lack of dexterity, so the Russian translations can be roughly translated back to English as *clumsiness*. In Hindi, *awkwardness* is translated as *bhaddaapan* which, back into English, gets translated as *clumsiness*, rather than *awkwardness*, and into *ajeeb*, which is a general concept for *strange*. Finally, in Turkish *awkwardness* gets translated as *beceriksizlik*, which is translated back into English as *incompetence*, clearly not a straightforward semantic equivalence. Examples like this are extremely frequent and the translatability of words denoting abstract concepts is often problematic even between languages that are typologically similar and belong to the same linguistic family. *Voorpret*, in Dutch, denotes a concept which can be described as fun perceived in advance of a fun thing. In English it can be translated as *anticipation*, although this

3. The same test can be done with any language used as starting point.

is an overextension of *voorpret*, which is clearly not a perfect translation because it gets translated as *verwachting*, back in Dutch. And the list goes on.

Because the meaning of words denoting abstract concepts is more language-specific (i.e., determined by linguistic factors) than the meaning of words denoting concrete concepts, these words are more difficult to learn, for both children who acquire their first language, as well as non-native speakers who study a foreign language. Especially for adult foreign language learners, who are arguably more inclined to encounter and to use words denoting abstract concepts compared to children, such words are quite challenging, because their meaning is shaped by the linguistic contexts in which they are used in the target language. While a language learner can in principle construct the meaning of a concrete word with word-to-world associations retrieved from her own experience in her own cultural community, she will find it more difficult to construct the meaning of an abstract concept, because her experience with the linguistic contexts in the target language are limited. A practical advice for language practitioners and researchers in applied linguistics would be to focus on teaching these abstract words by exposing the learners to as many linguistic contexts as possible, to enable them to establish the word-to-word associations that they need to construct the meaning of such words.

To conclude, new exciting open questions in the field of second/foreign language acquisition arise in relation to the construction of word meaning and the role that word-to-word associations play in such process. For example, if non-native speakers are more sensitive to the linguistic information retrieved from word-to-word associations, compared to native speakers, it might be possible to observe non-native speakers paradoxically outperforming native speakers in linguistic tasks involving words in the target language (which is the first language for native speakers) that are equally new to both the groups. Moreover, it would be interesting to compare in an empirical way the reliance on word-to-word associations vs. word-to-world associations in language learners who learn the foreign language in different circumstances, namely, in institutional settings while being in their home country and thus not being exposed to perceptual experiences within the target language/culture vs. in the setting of the target language. This latter setting could be tested in study-abroad students, who are exposed on a daily basis to both linguistic input in the target language as well as perceptual experiences within the hosting culture, which may favor the construction of word meanings by means of word-to-world associations, pushing the learners toward a learning strategy that better resembles the strategies employed by children when they learn their first language.

10.5 Outlook

Let me conclude this manuscript with a (motivated) joke.

A man goes to prison. The first night there, after the lights in the cell are turned off, his cellmate goes over to the bars and yells, "twelve!". The whole cell block breaks out laughing. A few minutes later, somebody else yells, "four!" and again, the whole cell block breaks out laughing.

The man asks his cellmate to explain why inmates were laughing at random numbers and the experienced prisoner explains: "we've all been in this here prison for so long, we all know all the same jokes. So, after a while we just started giving them numbers and yelling those numbers is enough to remind us of the joke instead of telling it".

Wanting to fit in, the new prisoner walks up to the bars and yells, "SEVEN!" But instead of laughter, a dead silence falls on the cell block. He turns to the older prisoner and asks: "What's wrong? Why didn't I get any laughs?".

His cellmate replies: "sometimes it's not the joke, but how you tell it!"

This joke summarizes in a paradoxical way the power that generic categories and processes of abstraction play in natural communication: they are shortcuts that allow us to communicate about concepts without providing complete and detailed descriptions that would in turn trigger fleshy mental simulations of the related perceptual experiences. Moreover, this joke points out an important aspect of meaning construction: the fact that word meaning is affected by extra-linguistic information such as pragmatic aspects related to the situation in which meaning is processed and prosody, both related to the *way* in which words are used. The very same word, uttered in different situational contexts and in different types of speech-acts, triggers different patterns of activation in the brain. For instance, hearing the word *water* in a context in which a speaker is naming a referent, such as a glass of water, activates in the listener brain circuits implicated in linking information about word forms and related reference objects, while hearing the word *water* in a context in which a speaker is requesting water from a peer, activates areas known to support action-related and social interactive knowledge (Egorova, Shtyrov and Pulvermüller, 2016). The different patterns of activation suggest that processing the same word, *water*, in different communicative contexts triggers different types of information related to that word, and therefore different types of semantic representations. Similarly, one could argue that different prosodic patterns in which the same word can be uttered afford different interpretations of its meaning. For example, if *water* is uttered with a rising prosodic pattern it may be associated with a request for water while if it uttered with a falling prosodic pattern it may be associated with an affirmative statement. Once again, while these arguments defending the role played by extra-linguistic information in shaping word

meaning could have been raised against the distributional theories of word meaning representation assuming a narrow view of how the distributional hypothesis could be applied, within a broader and deeper view of the distributional theory of meaning it is certainly possible to construct word meaning representations that integrate pragmatic and prosodic information retrieved from the communicative contexts in which words are used. The converging evidence in language and communication research, once again, comes from psychological evidence collected in the field of first language acquisition, and from computational evidence from the field of distributional semantics and AI. In the first field, it has been shown that young infants are sensitive to subtle differences between phonetic units and that they analyze the statistical distributions of sounds that they hear to form phonemic categories in a cross-situational manner (Kuhl, 2004). The same statistical regularities in sound patterns that are tracked and used to construct phonemic categories and to predict word segmentation on the basis of a continuous stream of sound are arguably used to differentiate different uses of the same word. Combined with clues extracted from social interactions and encoded in word-to-world patterns of associations, such regularities help children interpret the communicative goal encoded in the pragmatic and prosodic information that accompanies a word uttered in context. In the field of computational modelling and AI, the integration of extra-linguistic information in distributional semantic models of word meaning has been described in Chapter 7, in relation to multimodal distributional semantics and to the ability of vector representations to account for the construction of meaning that goes beyond the simple word co-occurrences.

Thus, even the information that contributes to shape a word meaning that we extract from the extra-linguistic, pragmatic contexts in which language is used, is learned and can be modelled on the basis of the same principles afforded by the distributional hypothesis, that is: associative processes, pattern detection, and similarity construction by means of feature matching.

In conclusion, while the mechanisms described throughout this book evolved from the converging evidence coming from cognitive sciences and computer sciences provide the scaffolding for a general theory of word meaning construction based on the distributional hypothesis, the potential future applications of such theory and such mechanisms were only briefly anticipated. I strongly believe, in this regard, that in the coming years we are going to witness fast and huge leaps forward in the fields of human and artificial intelligence research, especially in relation to natural language processing and production. Moreover, I believe that these achievements will be reached thanks to the collaboration between human and artificial intelligence, from which both human and artificial intelligence will benefit. Already today, human and artificial intelligence work side by side to solve open challenges, such as the detection of fake news and hate speech on social

media: on the basis of neural network algorithms (programmed by humans) AI works to identify the content of billions of texts, but human curators ultimately decide whether such texts should be flagged. Similarly, inside companies like Google, linguists are hand-labelling vast amounts of data to help train neural networks to understand natural language. The strengths and weaknesses of the two types of intelligence complement one another and should be seen as compatible and necessary to improve one another, rather than as in competition with one another.

References

Abutalebi, J., & Green, D. (2007). Bilingual language production: The neurocognition of language representation and control. *Journal of Neurolinguistics*, 20(3), 242–275. https://doi.org/10.1016/j.jneuroling.2006.10.003

Ahmad, J. (2012). Intentional vs. Incidental vocabulary learning. *ELT Research Journal*, 1(1), 71–79.

Althaus, N., & Westermann, G. (2016). Labels constructively shape object categories in 10-month-old infants. *Journal of Experimental Child Psychology*, 151, 5–17. https://doi.org/10.1016/j.jecp.2015.11.013

Amrami, A., & Goldberg, Y. (2019). *Towards better substitution-based word sense induction*. arXiv:1905.12598v2.

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, 116(3), 463–498. https://doi.org/10.1037/a0016261

Apresjan, Ju. D. (1974). Regular Polysemy. *Linguistics*, 12(142). doi: https://doi.org/10.1515/ling.1974.12.142.5

Asher, N. (2011). *Lexical meaning in context: A web of words*. Cambridge: Cambridge Univ. Press. https://doi.org/10.1017/CBO9780511793936

Asher, N., & Pustejovsky, J. (2013). A Type Composition Logic for Generative Lexicon. In James Pustejovsky, P. Bouillon, H. Isahara, K. Kanzaki, & C. Lee (Eds.), *Advances in Generative Lexicon Theory* (Vol. 46, pp. 39–66). doi: https://doi.org/10.1007/978-94-007-5189-7_3

Asmuth, J., & Gentner, D. (2017). Relational Categories are More Mutable than Entity Categories. *Quarterly Journal of Experimental Psychology*, 70(10), 2007–2025. https://doi.org/10.1080/17470218.2016.1219752

Axelrod, R., & Keohane, R. O. (1985). Achieving Cooperation under Anarchy: Strategies and Institutions. *World Politics*, 38(1), 226–254. https://doi.org/10.2307/2010357

Baayen, R. H. (2010). Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon*, 5, 436–461. https://doi.org/10.1075/ml.5.3.10baa

Baayen, R. H., Hendrix, P., & Ramscar, M. (2013). Sidestepping the combinatorial explosion: An explanation of ngram frequency effects based on naïve discriminative learning. *Lang Speech*, 56(3), 329–347. https://doi.org/10.1177/0023830913484896

Baker, C. (1979). Syntactic theory and the projection problem. *Linguistic Inquiry* 10. 533–81.

Balaban, M. T., & Waxman, S. R. (1997). Do Words Facilitate Object Categorization in 9-Month-Old Infants? *Journal of Experimental Child Psychology*, 64(1), 3–26. https://doi.org/10.1006/jecp.1996.2332

Baldwin, D. A. (1991). Infants' Contribution to the Achievement of Joint Reference. *Child Development*, 62(5), 875–890. https://doi.org/10.1111/j.1467-8624.1991.tb01577.x

Baltrušaitis, T., Ahuja, C., Morency, L. (2019). Multimodal Machine Learning: A Survey and Taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence. February 2019. https://doi.org/10.1109/TPAMI.2018.2798607

Bambini, V., Bertini, C., Schaeken, W., Stella, A., & Di Russo, F. (2016). Disentangling Metaphor from Context: An ERP Study. *Frontiers in Psychology*, 7. https://doi.org/10.3389/fpsyg.2016.00559

Barca, L., Burani, C., & Arduino, L. S. (2002). Word naming times and psycholinguistic norms for Italian nouns. *Behavior Research Methods, Instruments, & Computers*, 34(3), 424–434. https://doi.org/10.3758/BF03195471

Barcelona, A. (2000). On the plausibility of claiming a métonymie motivation for conceptual metaphor. In A. Barcelona (Ed.), *Metaphor and Metonymy at the Crossroads* (pp. 31–58). doi: https://doi.org/10.1515/9783110894677.31

Baroni, M., & Lenci, A. (2010). Distributional Memory: A General Framework for Corpus-Based Semantics. *Computational Linguistics*, 36(4), 673–721. https://doi.org/10.1162/coli_a_00016

Baroni, M., & Lenci, A. (2011). *How we BLESSed distributional semantic evaluation*, 1–10.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. Context-predicting semantic vectors. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 238–247. https://doi.org/10.3115/v1/P14-1023

Baroni, M., Evert, S., & Lenci, A. (Eds.). (2008). *Lexical semantics: Bridging the gap between semantic theory and computational simulation*. Retrieved from http://wordspace.collocations.de/lib/exe/fetch.php/workshop:esslli:esslli_2008_lexicalsemantics.pdf

Barrett, L. F. (2017). *How emotions are made: The secret life of the brain* (Paperback edition). London: PAN Books.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(04). doi: https://doi.org/10.1017/S0140525X99002149

Barsalou, L. W. (2003). Abstraction in perceptual symbol systems. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 358(1435), 1177–1187. https://doi.org/10.1098/rstb.2003.1319

Barsalou, L. W. (2008). Grounded Cognition. *Annual Review of Psychology*, 59(1), 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Barsalou, L. W., & Wiemer-Hastings, K. (2005). Situating Abstract Concepts. In D. Pecher & R. A. Zwaan (Eds.), *Grounding Cognition* (pp. 129–163). doi: https://doi.org/10.1017/CBO9780511499968.007

Barsalou, L. W., Santos, A., Simmons, K., & Wilson, C. (2008). Language and simulation in conceptual processing. In *Symbols and Embodiment: Debates on meaning and cognition* (pp. 245–283). Retrieved from https://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199217274.001.0001/acprof-9780199217274-chapter-13. https://doi.org/10.1093/acprof:oso/9780199217274.003.0013

Beigman Klebanov, B., Wee Leong, C., Gutierrez, D., Shutova, E. & Flor, M. (2016). Semantic classifications for detection of verb metaphors. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 101–106. https://doi.org/10.18653/v1/P16-2017

Bengio, Y., Schwenk, H., Senécal, J.-S., Morin, F., & Gauvain, J. L. (2006). Neural Probabilistic Language Models. In D. E. Holmes & L. C. Jain (Eds.), *Innovations in Machine Learning* (Vol. 194, pp. 137–186). doi: https://doi.org/10.1007/3-540-33486-6_6

Beretta, A., Fiorentino, R., & Poeppel, D. (2005). The effects of homonymy and polysemy on lexical access: An MEG study. *Cognitive Brain Research*, 24(1), 57–65. https://doi.org/10.1016/j.cogbrainres.2004.12.006

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, 109(9), 3253–3258. https://doi.org/10.1073/pnas.1113380109

Berlin, B., Breedlove, D. E., & Raven, P. H. (1973). General Principles of Classification and Nomenclature in Folk Biology. *American Anthropologist*, 75(1), 214–242. https://doi.org/10.1525/aa.1973.75.1.02a00140

Bialystok, E. (2011). Reshaping the mind: The benefits of bilingualism. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, 65(4), 229–235. https://doi.org/10.1037/a0025406

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where Is the Semantic System? A Critical Review and Meta-Analysis of 120 Functional Neuroimaging Studies. *Cerebral Cortex*, 19(12), 2767–2796. https://doi.org/10.1093/cercor/bhp055

Blasko, D. G., & Connine, C. M. (1993). Effects of familiarity and aptness on metaphor processing. *Journal of experimental psychology: Learning, memory, and cognition*, 19(2), 295–308.

Blythe, R. A., Smith, K., & Smith, A. D. M. (2010). Learning Times for Large Lexicons Through Cross-Situational Learning. *Cognitive Science*, 34(4), 620–642. https://doi.org/10.1111/j.1551-6709.2009.01089.x

Bolognesi, M. (2011). Il lessico mentale bilingue e gli spazi semantici distribuzionali: E similarità tra i verbi in L1, in L2 e nei corpora. *Studi di Glottodidattica*, 5(2), 51–72.

Bolognesi, M. (2014). Distributional Semantics meets Embodied Cognition: Flickr® as a database of semantic features. *Selected Papers from the 4th UK Cognitive Linguistics Conference*, 18–35.

Bolognesi, M. (2016a). Metaphors, bilingual mental lexicon and distributional models. In E. Gola & F. Ervas (Eds.), *Metaphor in Language, Cognition, and Communication* (Vol. 5, pp. 105–122). doi: https://doi.org/10.1075/milcc.5.06bol

Bolognesi, M. (2016b). Modeling Semantic Similarity between Metaphor Terms of Visual vs. Linguistic Metaphors through Flickr Tag Distributions. *Frontiers in Communication*, 1. https://doi.org/10.3389/fcomm.2016.00009

Bolognesi, M. (2017a). Flickr® distributional tagspace: Evaluating the semantic spaces emerging from Flickr® tag distributions. In (ed.). M. Jones, *Big data in cognitive science*. (pp. 144–173). New York, US: Routledge/Taylor & Francis Group.

Bolognesi, M. (2017b). Using semantic feature norms to investigate how the visual and verbal modes afford metaphor construction and expression. *Language and Cognition*, 9(3), 525–552. https://doi.org/10.1017/langcog.2016.27

Bolognesi, M., & Aina, L. (2019). Similarity is closeness: Using distributional semantic spaces to model similarity in visual and linguistic metaphors. *Corpus Linguistics and Linguistic Theory*, 15(1), 101–137. https://doi.org/10.1515/cllt-2016-0061

Bolognesi, M., Burgers, C., & Caselli, T. (2020). On Abstraction: Decoupling conceptual concreteness and categorical specificity. *Cognitive Processing*. https://doi.org/10.1007/s10339-020-00965-9

Bolognesi, M., & Steen, G. (2018). Editors' Introduction: Abstract Concepts: Structure, Processing, and Modeling. *Topics in Cognitive Science*, 10(3), 490–500. https://doi.org/10.1111/tops.12354

Bolognesi, M., & Steen, G. (Eds.). (2019). *Perspectives on abstract concepts: Cognition, language and communication*. Amsterdam; Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/hcp.65

Bonnaud, V., Gil, R., & Ingrand, P. (2002). Metaphorical and non-metaphorical links: A behavioral and ERP study in young and elderly adults. *Neurophysiologie Clinique/Clinical Neurophysiology*, 32(4), 258–268. https://doi.org/10.1016/S0987-7053(02)00307-6

Borghi, A. M., & Binkofski, F. (2014). *Words as social tools: An embodied view on abstract concepts*. New York: Springer. https://doi.org/10.1007/978-1-4614-9539-0

Boroditsky, L. (2001). Does Language Shape Thought? Mandarin and English Speakers' Conceptions of Time. *Cognitive Psychology*, 43(1), 1–22. https://doi.org/10.1006/cogp.2001.0748

Boroditsky, L. (2011). How Language Shapes Thought. *Scientific American*, 304(2), 62–65. https://doi.org/10.1038/scientificamerican0211-62

Boroditsky, L., & Prinz, J. (2008). What Thoughts Are Made Of. In G. R. Semin & E. R. Smith (Eds.), *Embodied Grounding* (pp. 98–116). doi: https://doi.org/10.1017/CBO9780511805837.005

Borowsky, R., & Masson, M. E. J. (1996). Semantic ambiguity effects in word identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(1), 63–85. https://doi.org/10.1037/0278-7393.22.1.63

Boulenger, V., Hauk, O., & Pulvermuller, F. (2009). Grasping Ideas with the Motor System: Semantic Somatotopy in Idiom Comprehension. *Cerebral Cortex*, 19(8), 1905–1914. https://doi.org/10.1093/cercor/bhn217

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. Processing texts automatically. In *Proceedings of the Biennial GSCL Conference*.

Bowdle, B. F., & Gentner, D. (2005). The Career of Metaphor. *Psychological Review*, 112(1), 193–216. https://doi.org/10.1037/0033-295X.112.1.193

Bruni, E., Tran, N. K., & Baroni, M. (2014). Multimodal Distributional Semantics. *Journal of Artificial Intelligence Research*, 49, 1–47. https://doi.org/10.1613/jair.4135

Brysbaert, M., & Duyck, W. (2010). Is it time to leave behind the Revised Hierarchical Model of bilingual language processing after fifteen years of service? *Bilingualism: Language and Cognition*, 13(3), 359–371. https://doi.org/10.1017/S1366728909990344

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3), 510–526. https://doi.org/10.3758/BF03193020

Bullinaria, J. A., & Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming, and SVD. *Behavior Research Methods*, 44(3), 890–907. https://doi.org/10.3758/s13428-011-0183-8

Burgoon, E. M., Henderson, M. D., & Markman, A. B. (2013). There Are Many Ways to See the Forest for the Trees: A Tour Guide for Abstraction. *Perspectives on Psychological Science*, 8(5), 501–520. https://doi.org/10.1177/1745691613497964

Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., & Moore, C. (1998). Social Cognition, Joint Attention, and Communicative Competence from 9 to 15 Months of Age. *Monographs of the Society for Research in Child Development*, 63(4), i. https://doi.org/10.2307/1166214

Casasanto, D. (2008). Who's Afraid of the Big Bad Whorf? Crosslinguistic Differences in Temporal Language and Thought. *Language Learning*, 58, 63–79. https://doi.org/10.1111/j.1467-9922.2008.00462.x

Casasanto, D., & Gijssels, T. (2015). What makes a metaphor an embodied metaphor? *Linguistics Vanguard*, 1(1). doi: https://doi.org/10.1515/lingvan-2014-1015

Cassani, G., Grimm, R., Gillis, S., & Daelemans, W. (2016). Constraining the search space in cross-situational word learning: different models make different predictions. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 1152–1157). Cognitive Science Society.

Çetinavcı, B. M. (2014). Contextual Factors in Guessing Word Meaning from Context in a Foreign Language. *Procedia – Social and Behavioral Sciences*, 116, 2670–2674. https://doi.org/10.1016/j.sbspro.2014.01.633

Chandler, S. (2017). The analogical modeling of linguistic categories. *Language and Cognition*, 9(1), 52–87. https://doi.org/10.1017/langcog.2015.24

Chersoni, E., Lenci, A., & Blache, P. (2017). Logical Metonymy in a Distributional Model of Sentence Comprehension. *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, 168–177. https://doi.org/10.18653/v1/S17-1021

Chomsky, N. (1957). *Syntactic structures*. Retrieved from. https://doi.org/10.1515/9783110218329

Chomsky, N. (1975). *The logical structure of linguistic theory*. New York: Plenum Press.

Chomsky, N. (1980). *Rules and representations*. Oxford, UK: Blackwell. https://doi.org/10.1017/S0140525X00001515

Church, K. W., & Hanks, P. (1989). Word association norms, mutual information, and lexicography. *Proceedings of the 27th Annual Meeting on Association for Computational Linguistics* -, 76–83. https://doi.org/10.3115/981623.981633

Clark, E. (1995). Language acquisitionThe lexicon and syntax. In *Speech, Language, and Communication* (pp. 304–337). doi: https://doi.org/10.1016/B978-012497770-9/50011-X

Clark, H. H., & Lucy, P. (1975). Understanding what is meant from what is said: A study in conversationally conveyed requests. *Journal of Verbal Learning and Verbal Behavior*, 14(1), 56–72. https://doi.org/10.1016/S0022-5371(75)80006-5

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. *Proceedings of the 25th International Conference on Machine Learning – ICML '08*, 160–167. https://doi.org/10.1145/1390156.1390177

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural Language Processing (almost) from Scratch. *ArXiv:1103.0398 [Cs]*. Retrieved from http://arxiv.org/abs/1103.0398

Connell, L. (2018). What have labels ever done for us? The linguistic shortcut in conceptual processing. *Language, Cognition and Neuroscience*, 1–11. https://doi.org/10.1080/23273798.2018.1471512

Costa, A., & Santesteban, M. (2004). Lexical access in bilingual speech production: Evidence from language switching in highly proficient bilinguals and L2 learners. *Journal of Memory and Language*, 50(4), 491–511. https://doi.org/10.1016/j.jml.2004.02.002

Coulson, S., & Van Petten, C. (2002). Conceptual integration and metaphor: an event-related potential study. *Memory & Cognition* 30, 958–968. https://doi.org/10.3758/BF03195780

Cox, T. F., & Cox, M. A. A. (2001). *Multidimensional Scaling*. Chapman & Hall.

Cruse, D. A. (1986). *Lexical semantics*. Cambridge, New York: Cambridge University Press.

Crutch, S. J., & Jackson, E. C. (2011). Contrasting Graded Effects of Semantic Similarity and Association across the Concreteness Spectrum. *Quarterly Journal of Experimental Psychology*, 64(7), 1388–1408. https://doi.org/10.1080/17470218.2010.543285

Crutch, S. J., & Warrington, E. (2005). Abstract and concrete concepts have structurally different representational frameworks. *Brain*, 128(3), 615–627. https://doi.org/10.1093/brain/awh349

Crutch, S. J., & Warrington, E. K. (2010). The differential dependence of abstract and concrete words upon associative and similarity-based information: Complementary semantic interference and facilitation effects. *Cognitive Neuropsychology*, 27(1), 46–71. https://doi.org/10.1080/02643294.2010.491359

Crutch, S. J., Connell, S., & Warrington, E. K. (2009). The different representational frameworks underpinning abstract and concrete knowledge: Evidence from odd-one-out judgements. *Quarterly Journal of Experimental Psychology*, 62(7), 1377–1390. https://doi.org/10.1080/17470210802483834

Crutch, S. J., Ridha, B. H., & Warrington, E. K. (2006). The Different Frameworks Underlying Abstract and Concrete Knowledge: Evidence from a Bilingual Patient with a Semantic Refractory Access Dysphasia. *Neurocase*, 12(3), 151–163. https://doi.org/10.1080/13554790600598832

Cuccio, V. (2018). *Attention to metaphor: From neurons to representations*. Amsterdam; Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/milcc.7

Daelemans, W. & van den Bosch, A. (2005). *Memory-based Language Processing. Studies in natural language processing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511486579

Daoutis, C. A., Franklin, A., Riddett, A., Clifford, A., & Davies, I. R. L. (2006). Categorical effects in children's colour search: A cross-linguistic comparison. *British Journal of Developmental Psychology*, 24(2), 373–400. https://doi.org/10.1348/026151005X51266

de Bot, K., Lowie, W., Thorne, S. L., & Verspoor, M. (2013). Dynamic Systems Theory as a comprehensive theory of second language development. In M. del P. García Mayo, M. J. Gutierrez Mangado, & M. Martínez Adrián (Eds.), *AILA Applied Linguistics Series* (Vol. 9, pp. 199–220). doi: https://doi.org/10.1075/aals.9.13ch10

De Grauwe, S., Swain, A., Holcomb, P. J., Ditman, T., & Kuperberg, G. R. (2010). Electrophysiological insights into the processing of nominal metaphors. *Neuropsychologia*, 48(7), 1965–1984. https://doi.org/10.1016/j.neuropsychologia.2010.03.017

De Vega, M., Glenberg, A., & Graesser, A. (2008). *Symbols and EmbodimentDebates on meaning and cognition*. https://doi.org/10.1093/acprof:oso/9780199217274.001.0001

Desai, R. H., Conant, L. L., Binder, J. R., Park, H., & Seidenberg, M. S. (2013). A piece of the action: Modulation of sensory-motor regions by action idioms and metaphors. *NeuroImage*, 83, 862–869. https://doi.org/10.1016/j.neuroimage.2013.07.044

De Saussure, F. (1916). *Cours de linguistique générale*, ed. C. Bally and A. Sechehaye, with the collaboration of A. Riedlinger, Lausanne and Paris: Payot.

Dove, G. (2009). Beyond perceptual symbols: A call for representational pluralism. *Cognition*, 110(3), 412–431. https://doi.org/10.1016/j.cognition.2008.11.016

Dove, G. (2016). Three symbol ungrounding problems: Abstract concepts and the future of embodied cognition. *Psychonomic Bulletin & Review*, 23(4), 1109–1121. https://doi.org/10.3758/s13423-015-0825-4

Dufter, P., & Schütze, H. (2019). *Analytical Methods for Interpretable Ultradense Word Embeddings*. arXiv preprint arXiv:1904.08654.

Egorova, N., Shtyrov, Y., & Pulvermüller, F. (2016). Brain basis of communicative actions in language. *NeuroImage*, 125, 857–867. https://doi.org/10.1016/j.neuroimage.2015.10.055

Elman, J. L. (2009). On the Meaning of Words and Dinosaur Bones: Lexical Knowledge Without a Lexicon. *Cognitive Science*, 33(4), 547–582. https://doi.org/10.1111/j.1551-6709.2009.01023.x

Erk, K., & Pado, S. (2010). Exemplar-based models for word meaning in context. In *'Proceedings of ACL*, Uppsala, Sweden, pp. 92–97.

Falkum, I. L., Recasens, M., & Clark, E. V. (2017). " The moustache sits down first": On the acquisition of metonymy. *Journal of Child Language*, 44(1), 87–119. https://doi.org/10.1017/S0305000915000720

Fellbaum, C. (1998). A Semantic Network of English: The Mother of All WordNets. *Computers and the Humanities*, 32(2/3), 209–220. https://doi.org/10.1023/A:1001181927857

Ferree, M. M., Gamson, W. A., Gerhards, J., & Rucht, D. (2002). *Shaping Abortion Discourse: Democracy and the Public Sphere in Germany and the United States*. https://doi.org/10.1017/CBO9780511613685

Ferretti, T. R., Kutas, M., & McRae, K. (2007). Verb aspect and the activation of event knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(1), 182–196. https://doi.org/10.1037/0278-7393.33.1.182

Ferretti, T. R., McRae, K., & Hatherell, A. (2001). Integrating Verbs, Situation Schemas, and Thematic Role Concepts. *Journal of Memory and Language*, 44(4), 516–547. https://doi.org/10.1006/jmla.2000.2728

Ferry, A. L., Hespos, S. J., & Waxman, S. R. (2010). Categorization in 3- and 4-Month-Old Infants: An Advantage of Words Over Tones. *Child Development*, 81(2), 472–479. https://doi.org/10.1111/j.1467-8624.2009.01408.x

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2001). Placing search in context: The concept revisited. *Proceedings of the Tenth International Conference on World Wide Web – WWW '01*, 406–414. https://doi.org/10.1145/371920.372094

Firth, J. (1957). A synopsis of linguistic theory 1930–1955. *Tudies in Linguistic Analysis*, 1–32.

Fitneva, S. A., & Christiansen, M. H. (2011). Looking in the Wrong Direction Correlates With More Accurate Word Learning. *Cognitive Science*, 35(2), 367–380. https://doi.org/10.1111/j.1551-6709.2010.01156.x

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. Cambridge, Mass: MIT Press. https://doi.org/10.7551/mitpress/4737.001.0001

Frisson, S., & Pickering, M. J. (1999). The processing of metonymy: Evidence from eye movements. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 25(6), 1366–1383. https://doi.org/10.1037/0278-7393.25.6.1366

Frisson, S., & Pickering, M. J. (2007). The processing of familiar and novel senses of a word: Why reading Dickens is easy but reading Needham can be hard. *Language and Cognitive Processes*, 22(4), 595–613. https://doi.org/10.1080/01690960601017013

Fulkerson, A. L., & Waxman, S. R. (2007). Words (but not Tones) facilitate object categorization: Evidence from 6- and 12-month-olds. *Cognition*, 105(1), 218–228. https://doi.org/10.1016/j.cognition.2006.09.005

Garcia, J., & Koelling, R. A. (1966). Relation of cue to consequence in avoidance learning. *Psychon Sci* 4, 123–124. https://doi.org/10.3758/BF03342209

Gass, S. (1999). DISCUSSION: Incidental Vocabulary Learning. *Studies in Second Language Acquisition*, 21(2), 319–333. https://doi.org/10.1017/S0272263199002090

Gentner, D. (1978). On Relational Meaning: The Acquisition of Verb Meaning. *Child Development*, 49(4), 988. https://doi.org/10.2307/1128738

Gentner, D. (1983). Structure-Mapping: A Theoretical Framework for Analogy*. *Cognitive Science*, 7(2), 155–170. https://doi.org/10.1207/s15516709cog0702_3

Gentner, D., & Boroditsky, L. (2001). Individuation, relativity and early word learning. In M. Bowerman & S. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 215–256). Cambridge, UK: Cambridge University Press.

Gentner, D., & Goldin-Meadow, S. (2003). *Language in mind: Advances in the study of language and thought*. Cambridge, MA: MIT Press.

Gerlach, P. (2018). *The Social Framework of Individual Decisions*. https://doi.org/10.18452/18725

Gibbs Jr., R. W., & Gerrig, R. J. (1989). How Context Makes Metaphor Comprehension Seem "Special." *Metaphor and Symbolic Activity*, 4(3), 145–158. https://doi.org/10.1207/s15327868ms0403_3

Gibbs, R. W. (1984). Literal Meaning and Psychological Theory*. *Cognitive Science*, 8(3), 275–304. https://doi.org/10.1207/s15516709cog0803_4

Gibbs, R. W. (2006a). *Embodiment and cognitive science*. Cambridge; New York: Cambridge University Press.

Gibbs, R. W. (2006b). Metaphor Interpretation as Embodied Simulation. *Mind & Language*, 21(3), 434–458. https://doi.org/10.1111/j.1468-0017.2006.00285.x

Gibbs, R. W. (2007). Experiential tests of figurative meaning construction. In G. Radden, K.-M. Köpcke, T. Berg, & P. Siemund (Eds.), *Aspects of Meaning Construction* (pp. 19–32). doi: https://doi.org/10.1075/z.136.04gib

Gibbs, R. W. (2011). Evaluating Conceptual Metaphor Theory. *Discourse Processes*, 48(8), 529–562. https://doi.org/10.1080/0163853X.2011.606103

Gildea, P., & Glucksberg, S. (1983). On understanding metaphor: The role of context. *Journal of Verbal Learning and Verbal Behavior*, 22(5), 577–590. https://doi.org/10.1016/S0022-5371(83)90355-9

Giora, R. (1997). Understanding figurative and literal language: The graded salience hypothesis. *Cognitive Linguistics*, 8(3), 183–206. https://doi.org/10.1515/cogl.1997.8.3.183

Glanzer, M., & Duarte, A. (1971). Repetition between and within languages in free recall. *Journal of Verbal Learning and Verbal Behavior*, 10(6), 625–630. https://doi.org/10.1016/S0022-5371(71)80069-5

Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. C. (2005). Hard Words. *Language Learning and Development*, 1(1), 23–64. https://doi.org/10.1207/s15473341lld0101_4

Glucksberg, S. (2001). *Understanding Figurative Language*. https://doi.org/10.1093/acprof:oso/9780195111095.001.0001

Glucksberg, S. (2003). The psycholinguistics of metaphor. *Trends in Cognitive Sciences*, 7(2), 92–96. https://doi.org/10.1016/S1364-6613(02)00040-2

Glucksberg, S., Gildea, P., & Bookin, H. (1982). On understanding non-literal speech: Can people ignore metaphors? *Journal of Verbal Learning and Verbal Behavior*, 1, 85–96. https://doi.org/10.1016/S0022-5371(82)90467-4

Glucksberg, S., & Keysar, B. (1990). Understanding metaphorical comparisons: Beyond similarity. *Psychological Review*, 97(1), 3–18. https://doi.org/10.1037/0033-295X.97.1.3

Goggin, J., & Wickens, D. D. (1971). Proactive interference and language change in short-term memory. *Journal of Verbal Learning and Verbal Behavior*, 10(4), 453–458. https://doi.org/10.1016/S0022-5371(71)80046-4

Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, 70(2), 109–135. https://doi.org/10.1016/S0010-0277(99)00003-7

Grady, J. (1997). *Foundations of Meaning: Primary Metaphors and Primary Scenes*. Retrieved from https://escholarship.org/uc/item/3g9427m2

Grady, J. (1999). A typology of motivation for conceptual metaphor (pp. 79–100). In R. Gibbs & G. Steen (Eds.), *Metaphor in cognitive linguistics*. Amsterdam, The Netherlands: John Benjamins. https://doi.org/10.1075/cilt.175.06gra

Grady, J. (2005). Primary metaphors as inputs to conceptual integration. *Journal of Pragmatics*, 37, 1595–1614. https://doi.org/10.1016/j.pragma.2004.03.012

Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition*, 1(2), 67–81. https://doi.org/10.1017/S1366728998000133

Grice, P. (1975). Logic and Conversation. Reprinted from *Syntax and semantics 3: Speech arts, Cole et al. "Logic and conversation"*, pp. 41–58, (1975), with permission from Elsevier.

Guida, A., & Lenci, A. (2007). Semantic Properties of Word Associations to Italian Verbs. *Italian Journal of Linguistics*, 19(2), 293–326.

Haastrup, K. (1989). Lexical inferencing procedures, Part 1 and Part 2. *Copenhagen: Handelshojskolen i Kobenhavn*.

Hampe, B. (Ed.). (2017). *Metaphor: Embodied Cognition and Discourse*. https://doi.org/10.1017/9781108182324

Harley, B., & Hart, D. (2000). Vocabulary Learning in the Content-oriented Second-language Classroom: Student Perceptions and Proficiency. *Language Awareness*, 9(2), 78–96. https://doi.org/10.1080/09658410008667139

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1–3), 335–346. https://doi.org/10.1016/0167-2789(90)90087-6

Harris, R. J. (1976). Comprehension of metaphors: A test of the two-stage processing model. *Bulletin of the Psychonomic Society*, 8(4), 312–314. https://doi.org/10.3758/BF03335150

Harris, Z. S. (1954). Distributional Structure. *Word*, 10(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Harris, Z. S. (1973). *A Leonard Bloomfield Anthology. Charles F. Hockett. International Journal of American Linguistics*, 39(4), 252–255. https://doi.org/10.1086/465274

Harris, Zellig S., & Mandelbaum, D. G. (1951). Selected Writings of Edward Sapir in Language, Culture, and Personality. *Language*, 27(3), 288. https://doi.org/10.2307/409757

Hill, F., Reichart, R., & Korhonen, A. (2015). SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation. *Computational Linguistics*, 41(4), 665–695. https://doi.org/10.1162/COLI_a_00237

Hoffman, P. (2016). The meaning of 'life' and other abstract words: Insights from neuropsychology. *Journal of Neuropsychology*, 10(2), 317–343. https://doi.org/10.1111/jnp.12065

Hoffman, P., Jefferies, E., & Lambon Ralph, M. A. (2010). Ventrolateral Prefrontal Cortex Plays an Executive Regulation Role in Comprehension of Abstract Words: Convergent Neuropsychological and Repetitive TMS Evidence. *Journal of Neuroscience*, 30(46), 15450–15456. https://doi.org/10.1523/JNEUROSCI.3783-10.2010

Hoffman, P., Lambon Ralph, M. A., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, 45(3), 718–730. https://doi.org/10.3758/s13428-012-0278-x

Hollis, G. (2019). Learning about things that never happened: a critique and refinement to the Rescorla-Wagner update rule when many outcomes are possible, *Memory & Cognition*, 1–16. https://doi.org/10.3758/s13421-019-00942-4

Horst, J. S., & Samuelson, L. K. (2008). Fast Mapping but Poor Retention by 24-Month-Old Infants. *Infancy*, 13(2), 128–157. https://doi.org/10.1080/15250000701795598

Huang, T. H. (2014). Social Metaphor Detection via Topical Analysis. *Computational Linguistics and Chinese Language Processing*, 19(2), 1–16.

Huckin, T., & Coady, J. (1999). Incidental vocabulary acquisition in a second language: A Review. *Studies in Second Language Acquisition*, 21(2), 181–193. https://doi.org/10.1017/S0272263199002028

Humphries, C., Binder, J. R., Medler, D. A., & Liebenthal, E. (2006). Syntactic and Semantic Modulation of Neural Activity during Auditory Sentence Comprehension. *Journal of Cognitive Neuroscience*, 18(4), 665–679. https://doi.org/10.1162/jocn.2006.18.4.665

Indefrey, P. (2006). A Meta-analysis of Hemodynamic Studies on First and Second Language Processing: Which Suggested Differences Can We Trust and What Do They Mean?: *Hemodynamic Studies of L1 and L2 Processing*. *Language Learning*, 56, 279–304. https://doi.org/10.1111/j.1467-9922.2006.00365.x

Inhoff, A. W., Lima, S. D., & Carroll, P. J. (1984). Contextual effects on metaphor comprehension in reading. *Memory & Cognition*, 12(6), 558–567. https://doi.org/10.3758/BF03213344

Jackendoff, R. (1997). *The architecture of the language faculty*. Cambridge, Mass: MIT Press.

Jung, K., Shavitt, S., Viswanathan, M., & Hilbe, J. M. (2014). Female hurricanes are deadlier than male hurricanes. *Proceedings of the National Academy of Sciences*, 111(24), 8782–8787. https://doi.org/10.1073/pnas.1402786111

Kachergis, G., Yu, C., & Shiffrin, R. M. (2014). Cross-situational word learning is both implicit and strategic. *Frontiers in Psychology*, 5. https://doi.org/10.3389/fpsyg.2014.00588

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2), 263. https://doi.org/10.2307/1914185

Kalgren, I., & Sahlgren, M. (2001). *From Words to Understanding*, 294–308. CSLI publications.

Kamin, L. J. (1968). "Attention-like" processes in classical conditioning. In M. R. Jones (Ed.), *Miami symposium on the prediction of behavior, 1967: Aversive Stimulation*. Coral Gables, FL: University of Miami Press.

Kiela, D., & Clark, S. (2014). A Systematic Study of Semantic Vector Space Model Parameters. *Proceedings of the 2nd Workshop on Continuous Vector Space Models and Their Compositionality (CVSC)*, 21–30. https://doi.org/10.3115/v1/W14-1503

Kiela, D., Bulat, L., & Clark, S. (2015). Grounding semantics in olfactory perception. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Volume 2: Short Papers) (pp. 231–236). Beijing, China: ACL.

Kiela, D., & Clark, S. (2015). Multi-and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)* (pp. 2461–2470). Lisbon, Portugal: ACL.

Kintsch, W. (2000). Metaphor comprehension: A computational theory. *Psychonomic Bulletin & Review*, 7(2), 257–266. https://doi.org/10.3758/BF03212981

Klein, D. E., & Murphy, G. L. (2001). The Representation of Polysemous Words. *Journal of Memory and Language*, 45(2), 259–282. https://doi.org/10.1006/jmla.2001.2779

Klein, D. E., & Murphy, G. L. (2002). Paper has been my ruin: Conceptual relations of polysemous senses. *Journal of Memory and Language*, 47, 548–570. https://doi.org/10.1016/S0749-596X(02)00020-7

Kövecses, Z. (2013). The Metaphor–Metonymy Relationship: Correlation Metaphors Are Based on Metonymy. *Metaphor and Symbol*, 28(2), 75–88. https://doi.org/10.1080/10926488.2013.768498

Kövecses, Z. (2017). Levels of metaphor. *Cognitive Linguistics*, 28(2), 321–347. https://doi.org/10.1515/cog-2016-0052

Krennmayr, T., Steen, G. (2017). VU Amsterdam Metaphor Corpus. In: Ide, N., Pustejovsky, J. (eds) *Handbook of Linguistic Annotation* pp. 1053–1071 Springer, Dordrecht. https://doi.org/10.1007/978-94-024-0881-2_39

Krishnakumaran, S., & Zhu, X. (2007). Hunting Elusive Metaphors Using Lexical Resources. *Proceedings of the Workshop on Computational Approaches to Figurative Language* (pp. 13–20). Rochester, New York.

Kroll, J. F., & Stewart, E. (1994). Category Interference in Translation and Picture Naming: Evidence for Asymmetric Connections Between Bilingual Memory Representations. *Journal of Memory and Language*, 33(2), 149–174. https://doi.org/10.1006/jmla.1994.1008

Kroll, J. F., Bobb, S. C., & Wodniecka, Z. (2006). Language selectivity is the exception, not the rule: Arguments against a fixed locus of language selection in bilingual speech. *Bilingualism: Language and Cognition*, 9(2), 119–135. https://doi.org/10.1017/S1366728906002483

Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, 5(11), 831–843. https://doi.org/10.1038/nrn1533

Kurzweil, R. (2012). *How to create a mind: The secret of human thought revealed*. New York: Viking.

Lai, V. T., Curran, T., & Menn, L. (2009). Comprehending conventional and novel metaphors: An ERP study. *Brain Research*, 1284, 145–155. https://doi.org/10.1016/j.brainres.2009.05.088

Lai, V. T., & Curran, T. (2013). ERP evidence for conceptual mappings and comparison processes during the comprehension of conventional and novel metaphors. *Brain and language*, 127(3), 484–496. https://doi.org/10.1016/j.bandl.2013.09.010

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by* (5. [Dr.]). Chicago, Ill.: Univ. of Chicago Press.

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

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Lapesa, G., & Evert, S. (2014). A Large Scale Evaluation of Distributional Semantic Models: Parameters, Interactions and Model Selection. *Transactions of the Association for Computational Linguistics*, 2, 531–546. https://doi.org/10.1162/tacl_a_00201

Leech, G. N. (1974). *Semantics*. Harmondsworth: Penguin.

Lemhöfer, K., Dijkstra, T., Schriefers, H., Baayen, R. H., Grainger, J., & Zwitserlood, P. (2008). Native language influences on word recognition in a second language: A megastudy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(1), 12–31. https://doi.org/10.1037/0278-7393.34.1.12

Lenci, A. (2018). Distributional Models of Word Meaning. *Annual Review of Linguistics*, 4(1), 151–171. https://doi.org/10.1146/annurev-linguistics-030514-125254

Lent, J. R. (2017). *The patterning instinct: A cultural history of humanity's search for meaning*. Amherst, New York: Prometheus Books.

Leong, C. W., Klebanov, B. B., & Shutova, E. (2018). A Report on the 2018 VUA Metaphor Detection Shared Task. *Proceedings of the Workshop on Figurative Language Processing. NAACL 2018 Workshop on Figurative Language Processing*. New Orleans, Louisiana.

Levinson, S. C. (2003). Language and mind: Let's get the issue straight! In D. Gentner and S. Goldin-Meadow (Eds.), *Language in Mind: Advances in the Study of Language and Thought*, pp. 25–46. Cambridge, MA: MIT Press.

Levy, O., & Goldberg, Y. (2014). Linguistic Regularities in Sparse and Explicit Word Representations. *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, 171–180. https://doi.org/10.3115/v1/W14-1618

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, 3, 211–225. https://doi.org/10.1162/tacl_a_00134

Liberman, V., Samuels, S. M., & Ross, L. (2004). The Name of the Game: Predictive Power of Reputations versus Situational Labels in Determining Prisoner's Dilemma Game Moves. *Personality and Social Psychology Bulletin*, 30(9), 1175–1185. https://doi.org/10.1177/0146167204264004

Littlemore, J. (2015). *Metonymy: Hidden shortcuts in language, thought and communication*. Cambridge ; New York: Cambridge University Press. https://doi.org/10.1017/CBO9781107338814

Littlemore, J., & Low, G. (2006). *Figurative Thinking and Foreign Language Learning*. https://doi.org/10.1057/9780230627567

Louwerse, M. M. (2007). Symbolic or embodied representations: A case for symbol interdependency. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.). *In Handbook of latent semantic analysis* (T. Landauer, D. McNamara, S. Dennis, W. Kintsch, pp. 107–120). Mahwah, NJ: Erlbaum.

Louwerse, M. M. (2008). Embodied relations are encoded in language. *Psychonomic Bulletin & Review*, 15(4), 838–844. https://doi.org/10.3758/PBR.15.4.838

Louwerse, M. M. (2011). Symbol Interdependency in Symbolic and Embodied Cognition: Topics in Cognitive Science. *Topics in Cognitive Science*, 3(2), 273–302. https://doi.org/10.1111/j.1756-8765.2010.01106.x

Louwerse, M. M. (2018). Knowing the Meaning of a Word by the Linguistic and Perceptual Company It Keeps. *Topics in Cognitive Science*, 10(3), 573–589. https://doi.org/10.1111/tops.12349

Louwerse, M. M., & Zwaan, R. A. (2009). Language Encodes Geographical Information. *Cognitive Science*, 33(1), 51–73. https://doi.org/10.1111/j.1551-6709.2008.01003.x

Lowder, M. W., & Gordon, P. C. (2013). It's hard to offend the college: Effects of sentence structure on figurative-language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(4), 993–1011. https://doi.org/10.1037/a0031671

Lupyan, G. (2008). The conceptual grouping effect: Categories matter (and named categories matter more). *Cognition*, 108(2), 566–577. https://doi.org/10.1016/j.cognition.2008.03.009

Lupyan, G. (2012). Linguistically Modulated Perception and Cognition: The Label-Feedback Hypothesis. *Frontiers in Psychology*, 3. https://doi.org/10.3389/fpsyg.2012.00054

Lynott, D., & Connell, L. (2013). Modality exclusivity norms for 400 nouns: The relationship between perceptual experience and surface word form. *Behavior Research Methods*, 45(2), 516–526. https://doi.org/10.3758/s13428-012-0267-0

Lyons, J. (1977). *Semantics*. Cambridge; New York: Cambridge University Press.

Mahon, B. Z. (2015). What is embodied about cognition? *Language, Cognition and Neuroscience*, 30(4), 420–429. https://doi.org/10.1080/23273798.2014.987791

Mahon, B. Z., & Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of Physiology-Paris*, 102(1–3), 59–70. https://doi.org/10.1016/j.jphysparis.2008.03.004

Malt, B. C., Gennari, S., Imai, M., Ameel, E., Tsuda, N., & Majid, A. (2008). Talking About Walking: Biomechanics and the Language of Locomotion. *Psychological Science*, 19(3), 232–240. https://doi.org/10.1111/j.1467-9280.2008.02074.x

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, 92, 57–78. https://doi.org/10.1016/j.jml.2016.04.001

Markman, E. M. (1990). Constraints Children Place on Word Meanings. *Cognitive Science*, 14(1), 57–77. https://doi.org/10.1207/s15516709cog1401_4

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, 20(2), 121–157. https://doi.org/10.1016/0010-0285(88)90017-5

McCormick, C. & Ryan, N. (2019). BERT Word Embeddings Tutorial. Retrieved from http://www.mccormickml.com

McDonald, S., & Brew, C. (2004). A distributional model of semantic context effects in lexical processing. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics – ACL '04*, 17-es. https://doi.org/10.3115/1218955.1218958

McElree, B., & Nordlie, J. (1999). Literal and figurative interpretations are computed in equal time. *Psychonomic Bulletin & Review*, 6(3), 486–494. https://doi.org/10.3758/BF03210839

McElree, B., Frisson, S., & Pickering, M. J. (2006). Deferred Interpretations: Why Starting Dickens is Taxing but Reading Dickens Isn't. *Cognitive Science*, 30(1), 181–192. https://doi.org/10.1207/s15516709cog0000_49

McElree, B., M. Traxler, M. Pickering, R. Seely, and R. Jackendoff (2001). Reading time evidence for enriched composition. *Cognition* 78(1), B17–B25.

McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, 119(4), 831–877. https://doi.org/10.1037/a0029872

McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, 119(4), 831–877. https://doi.org/10.1037/a0029872

McMurray, B., Horst, J. S., Toscano, J. C., & Samuelson, L. K. (2009). Integrating Connectionist Learning and Dynamical Systems Processing: Case Studies in Speech and Lexical Development. In J. Spencer (Ed.), *Toward a Unified Theory of Development Connectionism and Dynamic System Theory Re-Consider* (pp. 218–250). doi: https://doi.org/10.1093/acprof:oso/9780195300598.003.0011

McNeil, B. J., Pauker, S. G., Sox, H. C., & Tversky, A. (1982). On the Elicitation of Preferences for Alternative Therapies. *New England Journal of Medicine*, 306(21), 1259–1262. https://doi.org/10.1056/NEJM198205273062103

McRae, K., & Jones, M. (2013). *Semantic Memory*. In D. Reisberg (Ed.) The Oxford Handbook of Cognitive Psychology. https://doi.org/10.1093/oxfordhb/9780195376746.013.0014

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4), 547–559. https://doi.org/10.3758/BF03192726

McRae, K., Nedjadrasul, D., Pau, R., Lo, B. P.-H., & King, L. (2018). Abstract Concepts and Pictures of Real-World Situations Activate One Another. *Topics in Cognitive Science*, 10(3), 518–532. https://doi.org/10.1111/tops.12328

Meara, P. M. (2009). *Connected words: Word associations and second language vocabulary acquisition*. Amsterdam; Philadelphia: John Benjamins Pub. Co.. https://doi.org/10.1075/lllt.24

Medin, D., & Ortony, A. (1989). Comments on Part I: Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179–196). doi: https://doi.org/10.1017/CBO9780511529863.009

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, 108(22), 9014–9019. https://doi.org/10.1073/pnas.1105040108

Meng, Y., Huang, J., Wang, G., Wang, Z., Zhang, C., Han, J. (2020). Unsupervised Word Embedding Learning by Incorporating Local and Global Contexts. *Frontiers in Big Data*, 3, doi: https://doi.org/10.3389/fdata.2020.00009

Mesquita, B., Barrett, L. F., & Smith, E. R. (Eds.). (2010). *The mind in context*. New York: Guilford Press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ArXiv:1301.3781 [Cs]*. Retrieved from http://arxiv.org/abs/1301.3781

Mikolov, T., Yih, S. W., & Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013)*. Retrieved from https://www.microsoft.com/en-us/research/publication/linguistic-regularities-in-continuous-space-word-representations/

Miller, G. A. (1998). Nouns in wordnet. *WordNet: An Electronic Lexical Database*, 23–46.

Mitchell, M. (2020). On Crashing the Barrier of Meaning in Artificial Intelligence. *AI Magazine*, 41(2), 86–92. https://doi.org/10.1609/aimag.v41i2.5259

Monaghan, P., & Mattock, K. (2012). Integrating constraints for learning word–referent mappings. *Cognition*, 123(1), 133–143. https://doi.org/10.1016/j.cognition.2011.12.010

Moss, H. E., Ostrin, R. K., Tyler, L. K., & Marslen-Wilson, W. D. (1995). Accessing different types of lexical semantic information: Evidence from priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(4), 863–883. https://doi.org/10.1037/0278-7393.21.4.863

Munakata, Y., & McClelland, J. L. (2003). Connectionist models of development. *Developmental Science*, 6(4), 413–429. https://doi.org/10.1111/1467-7687.00296

Murphy, G. (2004). *The big book of concepts*. Cambridge, MA: MIT press.

Nacey, S. (2013). *Metaphors in learner English*. Amsterdam ; Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/milcc.2

Nassaji, H. (2003). L2 Vocabulary Learning from Context: Strategies, Knowledge Sources, and Their Relationship with Success in L2 Lexical Inferencing. *TESOL Quarterly*, 37(4), 645. https://doi.org/10.2307/3588216

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge ; New York: Cambridge University Press. https://doi.org/10.1017/CBO9781139524759

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2), 1–69. https://doi.org/10.1145/1459352.1459355

Noble, W., & Davidson, I. (1991). The Evolutionary Emergence of Modern Human Behaviour: Language and its Archaeology. *Man*, 26(2), 223. https://doi.org/10.2307/2803830

Noble, W., & Davidson, I. (1993). Tracing the emergence of modern human behavior: Methodological pitfalls and a theoretical path. *Journal of Anthropological Archaeology*, 12(2), 121–149. https://doi.org/10.1006/jaar.1993.1004

Noppeney, U., & Price, C. J. (2004). An fMRI Study of Syntactic Adaptation. *Journal of Cognitive Neuroscience*, 16(4), 702–713. https://doi.org/10.1162/089892904323057399

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General* 115(1), 39–61. https://doi.org/10.1037/0096-3445.115.1.39

Nunberg, G. (1995). Transfers of Meaning. *Journal of Semantics*, 12(2), 109–132. https://doi.org/10.1093/jos/12.2.109

Ortony, A., Schallert, D. L., Reynolds, R. E., & Antos, S. J. (1978). Interpreting metaphors and idioms: Some effects of context on comprehension. *Journal of Verbal Learning and Verbal Behavior*, 17(4), 465–477. https://doi.org/10.1016/S0022-5371(78)90283-9

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological Bulletin*, 49(3), 197–237. https://doi.org/10.1037/h0055737

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The Measurement of Meaning*. University of Illinois press.

Padó, S., & Lapata, M. (2007). Dependency-Based Construction of Semantic Space Models. *Computational Linguistics*, 33(2), 161–199. https://doi.org/10.1162/coli.2007.33.2.161

Paivio, A. (1983). The mind's eye in arts and science. *Poetics*, 12(1), 1–18. https://doi.org/10.1016/0304-422X(83)90002-5

Paivio, A. (1990). *Mental Representations*. https://doi.org/10.1093/acprof:oso/9780195066661.001.0001

Paivio, A. (2010). Dual coding theory and the mental lexicon. *The Mental Lexicon*, 5(2), 205–230. https://doi.org/10.1075/ml.5.2.04pai

Parel, R. (2004). The impact of lexical inferencing strategies on second language reading proficiency. *Reading and Writing*, 17(6), 847–873. https://doi.org/10.1007/s11145-004-9347-6

Park, S., Bak, J., & Oh, A. (2017). Rotated word vector representations and their interpretability. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/D17-1041

Pecher, D. (2018). Curb Your Embodiment. *Topics in Cognitive Science*, 10(3), 501–517. https://doi.org/10.1111/tops.12311

Pecher, D., & Zwaan, R. A. (Eds.). (2005). *Grounding cognition: The role of perception and action in memory, language, and thinking*. New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511499968

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). *Proceedings of the 19th National Conference on Artifical Intelligence*, 1024–1025. AAAI Press.

Peirsman, Y., & Geeraerts, D. (2006). Metonymy as a prototypical category. *Cognitive Linguistics*, 17(3). doi: https://doi.org/10.1515/COG.2006.007

Piñango, M. M., Zhang, M., Foster-Hanson, E., Negishi, M., Lacadie, C., & Constable, R. T. (2017). Metonymy as Referential Dependency: Psycholinguistic and Neurolinguistic Arguments for a Unified Linguistic Treatment. *Cognitive Science*, 41, 351–378. https://doi.org/10.1111/cogs.12341

Pinker, S. (1991). Rules of language. *Science* 253.530–35. https://doi.org/10.1126/science.1857983

Pinker, S. (1995). *The language instinct: The new science of language and mind*. London: Penguin.

Pinker, S. (1999). *Words and rules*. New York: Basic Books.

Pinker, S. (2004). Clarifying the logical problem of language acquisition. *Journal of Child Language* 31.949–53. https://doi.org/10.1017/S0305000904006439

Pirrelli, V., Marzi, C., Ferro, M., Cardillo, F. A., Baayen, H. R., & Milin, P. (2020). Psycho-computational modelling of the mental lexicon. In V. Pirrelli, I. Plag, and W. Dressler (eds.) *Word Knowledge and Word Usage*, pages: 23–82. Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110440577-002

Plat, R., Lowie, W. & de Bot, K. (2018). Word Naming in the L1 and L2: A Dynamic Perspective on Automatization and the Degree of Semantic Involvement in Naming. *Frontiers in Psychology* 8.

Plunkett, K., Hu, J.-F., & Cohen, L. B. (2008). Labels can override perceptual categories in early infancy. *Cognition*, 106(2), 665–681. https://doi.org/10.1016/j.cognition.2007.04.003

Polka, L., & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance*, 20(2), 421–435. https://doi.org/10.1037/0096-1523.20.2.421

Ponari, M., Norbury, C. F., & Vigliocco, G. (2017). How do children process abstract concepts? Evidence from a lexical decision task. *Developmental Science*, 10, 10–11. https://doi.org/10.1111/desc.12549

Ponniah, J. (2011). The Effectiveness of the Comprehension Hypothesis: A Review on the Current Research on Incidental Vocabulary Acquisition. *Journal on English Language Teaching*, 1(2), 1–4.

Pragglejaz Group (2007). MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22:1, pp1–39. https://doi.org/10.1080/10926480709336752

Pulvermüller, F. (2018). Neurobiological Mechanisms for Semantic Feature Extraction and Conceptual Flexibility. *Topics in Cognitive Science*, 10(3), 590–620. https://doi.org/10.1111/tops.12367

Pustejovsky, J. (1995). *The generative lexicon*. Cambridge, Mass: MIT Press.

Pustejovsky, James. (1991). The generative lexicon. *Computational Linguistics*, 17(4), 409–441.

Pynte, J., Besson, M., Robichon, F.-H., & Poli, J. (1996). The Time-Course of Metaphor Comprehension: An Event-Related Potential Study. *Brain and Language*, 55(3), 293–316. https://doi.org/10.1006/brln.1996.0107

Quine, W. V. O. (2013). *Word and object*. Place of publication not identified: Martino Fine Books. https://doi.org/10.7551/mitpress/9636.001.0001

Quinn, P. C., Eimas, P. D., & Rosenkrantz, S. L. (1993). Evidence for Representations of Perceptually Similar Natural Categories by 3-Month-Old and 4-Month-Old Infants. *Perception*, 22(4), 463–475. https://doi.org/10.1068/p220463

Raganato, A., Camacho-Collados, J., & Navigli, R. (2017). Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 99–110. https://doi.org/10.18653/v1/E17-1010

Ramscar, M., Dye, M., & Klein, J. (2013). Children value informativity over logic in word learning. *Psychological Science* 24, 6, 1017–1023. https://doi.org/10.1177/0956797612460691

Ramscar, M., Dye, M., & McCauley, S. M. (2013). Error and expectation in language learning: The curious absence of mouses in adult speech. *Language*, 89(4), 760–793. https://doi.org/10.1353/lan.2013.0068

Raposo, A., Moss, H. E., Stamatakis, E. A., & Tyler, L. K. (2009). Modulation of motor and premotor cortices by actions, action words and action sentences. *Neuropsychologia*, 47(2), 388–396. https://doi.org/10.1016/j.neuropsychologia.2008.09.017

Rapp, R. (2004). *A Freely Available Automatically Generated Thesaurus of Related Words*, 395–398.

Reines, M. F., & Prinz, J. (2009). Reviving Whorf: The Return of Linguistic Relativity. *Philosophy Compass*, 4(6), 1022–1032. https://doi.org/10.1111/j.1747-9991.2009.00260.x

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy, Eds., *Classical Conditioning II*, pp. 64–99. Appleton-CenturyCrofts.

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist*, 43(3), 151–160. https://doi.org/10.1037/0003-066X.43.3.151

Restrepo Ramos, F. D. (2015). Incidental Vocabulary Learning in Second Language Acquisition: A Literature Review. *PROFILE Issues in Teachers' Professional Development*, 17(1), 157–166. https://doi.org/10.15446/profile.v17n1.43957

Riordan, B., & Jones, M. N. (2011). Redundancy in Perceptual and Linguistic Experience: Comparing Feature-Based and Distributional Models of Semantic Representation. *Topics in Cognitive Science*, 3(2), 303–345. https://doi.org/10.1111/j.1756-8765.2010.01111.x

Roberson, D., Hanley, J. R., & Pak, H. (2009). Thresholds for color discrimination in English and Korean speakers. *Cognition*, 112(3), 482–487. https://doi.org/10.1016/j.cognition.2009.06.008

Robinson, C. W., & Sloutsky, V. M. (2007). Linguistic Labels and Categorization in Infancy: Do Labels Facilitate or Hinder? *Infancy*, 11(3), 233–253. https://doi.org/10.1111/j.1532-7078.2007.tb00225.x

Rodd, J., Gaskell, G., & Marslen-Wilson, W. (2002). Making Sense of Semantic Ambiguity: Semantic Competition in Lexical Access. *Journal of Memory and Language*, 46(2), 245–266. https://doi.org/10.1006/jmla.2001.2810

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104(3), 192–233. https://doi.org/10.1037/0096-3445.104.3.192

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8(3), 382–439. https://doi.org/10.1016/0010-0285(76)90013-X

Rosenblatt, F. (1957). The perceptron, a perceiving and recognizing automaton (Project Para Report No. 85-460-1). Ithaca, NY: Cornell Aeronautical Laboratory (CAL).

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386. https://doi.org/10.1037/h0042519

Rothe, S., Ebert, S., & Schütze, H. (2016). Ultradense word embeddings by orthogonal transformation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10), 627–633. https://doi.org/10.1145/365628.365657

Rüschemeyer, S.-A., Brass, M., & Friederici, A. D. (2007). Comprehending Prehending: Neural Correlates of Processing Verbs with Motor Stems. *Journal of Cognitive Neuroscience*, 19(5), 855–865. https://doi.org/10.1162/jocn.2007.19.5.855

Sabsevitz, D. S., Medler, D. A., Seidenberg, M., & Binder, J. R. (2005). Modulation of the semantic system by word imageability. *NeuroImage*, 27(1), 188–200. https://doi.org/10.1016/j.neuroimage.2005.04.012

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, 274(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word Segmentation: The Role of Distributional Cues. *Journal of Memory and Language*, 35(4), 606–621. https://doi.org/10.1006/jmla.1996.0032

Sahlgren, M. (2006). The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. Stockholm: Dep. of Linguistics, Stockholm Univ. [unpublished thesis].

Sahlgren, M., & Lenci, A. (2016). The Effects of Data Size and Frequency Range on Distributional Semantic Models. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 975–980. https://doi.org/10.18653/v1/D16-1099

Schwanenflugel, P. J., & Shoben, E. J. (1983). Differential context effects in the comprehension of abstract and concrete verbal materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9(1), 82–102. https://doi.org/10.1037/0278-7393.9.1.82

Schwanenflugel, P. J., Harnishfeger, K. K., & Stowe, R. W. (1988). Context availability and lexical decisions for abstract and concrete words. *Journal of Memory and Language*, 27(5), 499–520. https://doi.org/10.1016/0749-596X(88)90022-8

Schütze, H. (1993). Word Space. *Advances in Neural Information Processing Systems* 5, 895–902.

Schütze, H. (1992). Dimensions of meaning. In *Proceedings of Supercomputing '92*, 787–796. IEEE Press. https://doi.org/10.1109/SUPERC.1992.236684

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1), 97–124.

Scott, S. K., Blank, C., Rosen, S., & Wise, R. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain*, 123(12), 2400–2406. https://doi.org/10.1093/brain/123.12.2400

Searle, J. R. (1975). Speech acts and recent linguistics. *Annals of the New York Academy of Sciences*, 263(1 Developmental), 27–38. https://doi.org/10.1111/j.1749-6632.1975.tb41567.x

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences* 3 (3):417–57. https://doi.org/10.1017/S0140525X00005756

Sejnowski, T. J. (2018). *The deep learning revolution*. Cambridge, Massachusetts: The MIT Press. https://doi.org/10.7551/mitpress/11474.001.0001

Shutova, Ekaterina. 2010. Automatic metaphor interpretation as a paraphrasing task. In *Proceedings of NAACL* 2010, pp. 1029–1037, Los Angeles, USA.

Shutova, E., Kaplan, J., Teufel, S., & Korhonen, A. (2013). A computational model of logical metonymy. *ACM Transactions on Speech and Language Processing*, 10(3), 1–28. https://doi.org/10.1145/2483969.2483973

Shutova, E., Kiela, D., and Maillard, J. (2016). Black holes and white rabbits: Metaphor identification with visual features. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (San Diego, CA), 160–170.

Shutova, E., Sun, L., Korhonen, A. (2010). Metaphor Identification Using Verb and Noun Clustering. *Proceedings of the 23rd International Conference on Computational Linguistics* (Coling 2010), pages 1002–1010, Beijing, China.

Sikos, J. & Padó, S. (2019). Frame Identification as Categorization: Exemplars vs Prototypes in Embeddingland. *Proceedings of the 13th International Conference on Computational Semantics*. Gothenburg, Sweden.

Slobin, D. I. (1996). Two ways to travel: Verbs of motion in English and Spanish. In M. Shibatani & S. Thompson (Eds.), *Grammatical Constructions* (pp. 195–219). Clarendon Press.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3), 1558–1568. https://doi.org/10.1016/j.cognition.2007.06.010

Spitsyna, G., Warren, J. E., Scott, S. K., Turkheimer, F. E., & Wise, R. J. S. (2006). Converging Language Streams in the Human Temporal Lobe. *Journal of Neuroscience*, 26(28), 7328–7336. https://doi.org/10.1523/JNEUROSCI.0559-06.2006

Spivey, M. (2006). *The Continuity of Mind*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195170788.001.0001

Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A., Krennmayr, T., & Pasma, T. (2010). *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. https://doi.org/10.1075/celcr.14

Swingley, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1536), 3617–3632. https://doi.org/10.1098/rstb.2009.0107

Talmy, L. (1991). Path to Realization: A Typology of Event Conflation. *Annual Meeting of the Berkeley Linguistics Society*, 17(1), 480. https://doi.org/10.3765/bls.v17i0.1620

Talmy, L. (2000). *Typology and process in concept structuring* (1. MIT Press paperback ed). Cambridge, Mass.: MIT Press. https://doi.org/10.7551/mitpress/6848.001.0001

Tannen, D. (1993). What's in a frame? Surface evidence for underlying expectations. In D. Tannen (Ed.), *Framing in discourse* (pp. 14–56). New York: Oxford University Press.

Thaler, S., Simperl, E., Siorpaes, K., & Hofer, C. (2011). A survey on games for Knowledge acquisition. STI Technical Report, May 2011, 19.

Thomas, D. G., Campos, J. J., Shucard, D. W., Ramsay, D. S., & Shucard, J. (1981). Semantic Comprehension in Infancy: A Signal Detection Analysis. *Child Development*, 52(3), 798. https://doi.org/10.2307/1129079

Thornbury, S. (2013). *How to teach vocabulary* (13. impr). Harlow: Longman, Pearson Education.

Tomasello, M. (2003). *The cultural origins of human cognition* (4. print). Cambridge, Mass.: Harvard Univ. Press.

Tomasello, M., & Barton, M. E. (1994). Learning words in nonostensive contexts. *Developmental Psychology*, 30(5), 639–650. https://doi.org/10.1037/0012-1649.30.5.639

Tomasello, M., Strosberg, R., & Akhtar, N. (1996). Eighteen-month-old children learn words in non-ostensive contexts. *Journal of Child Language*, 23(1), 157–176. https://doi.org/10.1017/S0305000900010138

Trask, A., Michalak, M., & Liu, J. (2015). Sense2vec – a fast and accurate method for word sense disambiguation in neural word embeddings. *ArXiv e-prints*.

Traxler, M. J., Morris, R. K., & Seely, R. E. (2002). Processing Subject and Object Relative Clauses: Evidence from Eye Movements. *Journal of Memory and Language*, 47(1), 69–90. https://doi.org/10.1006/jmla.2001.2836

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, 66(1), 126–156. https://doi.org/10.1016/j.cogpsych.2012.10.001

Turney, P. D. (2001). Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In L. De Raedt & P. Flach (Eds.), *Machine Learning: ECML 2001* (Vol. 2167, pp. 491–502). doi: https://doi.org/10.1007/3-540-44795-4_42

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481), 453–458. https://doi.org/10.1126/science.7455683

Utsumi, A. (2011). Computational Exploration of Metaphor Comprehension Processes Using a Semantic Space Model. *Cognitive Science*, 35(2), 251–296. https://doi.org/10.1111/j.1551-6709.2010.01144.x

Van den Bos, G. R. (Ed.). (2007). *APA dictionary of psychology* (1st ed). Washington, DC: American Psychological Association.

Veale, T. (2012). *Exploding the creativity myth: The computational foundations of linguistic creativity*. London; New York: Continuum International Pub. Group.

Veale, T., Shutova, E., & Klebanov, B. B. (2016). Metaphor: A Computational Perspective. *Synthesis Lectures on Human Language Technologies*, 9(1), 1–160. https://doi.org/10.2200/S00694ED1V01Y201601HLT031

Vigliocco, G., Filipovic Kleiner, L. (2004). From mind in the mouth to language in the mind: Language in Mind. *Trends in Cognitive Sciences*, 8 (1), 5–7. https://doi.org/10.1016/j.tics.2003.10.019

Vigliocco, G., & Vinson, D. P. (2007). *Semantic representation*. https://doi.org/10.1093/oxfordhb/9780198568971.013.0012

Vigliocco, G., Ponari, M., & Norbury, C. (2018). Learning and Processing Abstract Words and Concepts: Insights From Typical and Atypical Development. *Topics in Cognitive Science* 10(3): 533–549. https://doi.org/10.1111/tops.12347

Vivas, L., Montefinese, M., Bolognesi, M., and Vivas, G. (2020). Core features: measures and characterization for different languages. *Cognitive Processing*. https://doi.org/10.1007/s10339-020-00969-5

von Ahn, L., & Dabbish, L. (2004). Labeling images with a computer game. *Proceedings of the 2004 Conference on Human Factors in Computing Systems – CHI '04*, 319–326. https://doi.org/10.1145/985692.985733

Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition*, 107(2), 729–742. https://doi.org/10.1016/j.cognition.2007.08.007

Wang, J., Conder, J. A., Blitzer, D. N., & Shinkareva, S. V. (2010). Neural representation of abstract and concrete concepts: A meta-analysis of neuroimaging studies. *Human Brain Mapping*, 31(10), 1459–1468. https://doi.org/10.1002/hbm.20950

Warrington, E. K., & Cipolotti, L. (1996). Word comprehension: The distinction between refractory and storage impairments. *Brain*, 119(2), 611–625. https://doi.org/10.1093/brain/119.2.611

Warrington, E. K., & Shallice, T. (1979). Semantic access dyslexia. *Brain*, 102(1), 43–63. https://doi.org/10.1093/brain/102.1.43

Waxman, S. R., & Markow, D. B. (1995). Words as Invitations to Form Categories: Evidence from 12- to 13-Month-Old Infants. *Cognitive Psychology*, 29(3), 257–302. https://doi.org/10.1006/cogp.1995.1016

Weaver, W. (1949). Translation. In Locke, W. N.; Booth, A. D. (eds.). *Machine Translation of Languages: Fourteen Essays*. Cambridge, MA: MIT Press.

Webb, S. (2008). The Effects of Context on Incidental Vocabulary Learning. *Reading in a Foreign Language*, 20(2), 232–245.

Weber, I., Robertson, S., & Vojnovic, M. (2009). Rethinking the ESP Game. *Conference on Human Factors in Computing Systems*, 11. Retrieved from https://www.microsoft.com/en-us/research/publication/rethinking-the-esp-game/

Weiland, H., Bambini, V., & Schumacher, P. B. (2014). The role of literal meaning in figurative language comprehension: Evidence from masked priming ERP. *Frontiers in Human Neuroscience*, 8. https://doi.org/10.3389/fnhum.2014.00583

Werkmann Horvat, A., Bolognesi, M., and Lahiri, A. (forth.). Processing of literal and metaphorical meanings in polysemous verbs: An experiment and its methodological importance.

Werkmann Horvat, A., Bolognesi, M., and Kohl, K. (forth.). Demolishing walls and myths: On the status of conventional metaphorical meaning in the L2 lexicon.

Westermann, G., & Mareschal, D. (2014). From perceptual to language-mediated categorization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1634), 20120391. https://doi.org/10.1098/rstb.2012.0391

Wilks, Y. (1975). A Preferential, Pattern-Seeking, Semantics for Natural Language Inference. *Artificial Intelligence* 6, 53–74.

Wilks, Y. (1978). Making preferences more active. *Artificial Intelligence* 11(3), 197–223.

Wheeler, D. J. (1952). The use of sub-routines in programmes. *ACM '52: Proceedings of the 1952 ACM national meeting*, 235–236. https://doi.org/10.1145/609784.609816

Wiemer-Hastings, K., & Xu, X. (2005). Content Differences for Abstract and Concrete Concepts. *Cognitive Science*, 29(5), 719–736. https://doi.org/10.1207/s15516709cog0000_33

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences*, 104(19), 7780–7785. https://doi.org/10.1073/pnas.0701644104

Wittgenstein, L. (1953). *Philosophical Investigations*. Blackwell Publishing.

Wolff, P., & Holmes, K. J. (2011). Linguistic relativity: Linguistic relativity. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(3), 253–265. https://doi.org/10.1002/wcs.104

Yu, C. (2008). A Statistical Associative Account of Vocabulary Growth in Early Word Learning. *Language Learning and Development*, 4(1), 32–62. https://doi.org/10.1080/15475440701739353

Yu, C., & Smith, L. B. (2007). Rapid Word Learning Under Uncertainty via Cross-Situational Statistics. *Psychological Science*, 18(5), 414–420. https://doi.org/10.1111/j.1467-9280.2007.01915.x

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition*, 125(2), 244–262. https://doi.org/10.1016/j.cognition.2012.06.016

Yu, C., Zhong, Y., & Fricker, D. (2012). Selective Attention in Cross-Situational Statistical Learning: Evidence From Eye Tracking. *Frontiers in Psychology*, 3. https://doi.org/10.3389/fpsyg.2012.00148

Yurovsky, D., Fricker, D. C., Yu, C., & Smith, L. B. (2014). The role of partial knowledge in statistical word learning. *Psychonomic Bulletin & Review*, 21(1), 1–22. https://doi.org/10.3758/s13423-013-0443-y

Zarcone, A., Lenci, A., Padó, S., & Utt, J. (2013). Fitting, Not Clashing! A Distributional Semantic Model of Logical Metonymy. *Proceedings of IWCS*, 404–410.

Zarcone, A., Padó, S., & Lenci, A. (2014). Logical Metonymy Resolution in a Words-as-Cues Framework: Evidence from Self-Paced Reading and Probe Recognition. *Cognitive Science*, 38(5), 973–996. https://doi.org/10.1111/cogs.12108

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley: Boston, MA, USA.

Zwaan, R. A. (2014). Embodiment and language comprehension: Reframing the discussion. *Trends in Cognitive Sciences*, 18(5), 229–234. https://doi.org/10.1016/j.tics.2014.02.008

Zwaan, R. A. (2016). Situation models, mental simulations, and abstract concepts in discourse comprehension. *Psychonomic Bulletin & Review*, 23(4), 1028–1034. https://doi.org/10.3758/s13423-015-0864-x

Index

Words are not just labels for conceptual categories. Words construct conceptual categories, frame situations and influence behavior. Where do they get their meaning?

This book describes how words acquire their meaning. The author argues that mechanisms based on associations, pattern detection, and feature matching processes explain how words acquire their meaning from experience and from language alike. Such mechanisms are summarized by the distributional hypothesis, a computational theory of meaning originally applied to word occurrences only, and hereby extended to extra-linguistic contexts.

By arguing in favor of the cognitive foundations of the distributional hypothesis, which suggests that words that appear in similar contexts have similar meaning, this book offers a theoretical account for word meaning construction and extension in first and second language that bridges empirical findings from cognitive and computer sciences. Plain language and illustrations accompany the text, making this book accessible to a multidisciplinary academic audience.

JOHN BENJAMINS PUBLISHING COMPANY