1

From Usage to Meaning

The Foundations of Distributional Semantics

Distributional semantics is the study of how distributional information can be used to model semantic facts. Its theoretical foundation has become known as the **Distributional Hypothesis**:

> Lexemes with similar linguistic contexts have similar meanings.

This chapter presents the epistemological principles of distributional semantics. In Section 1.1, we explore the historical roots of the Distributional Hypothesis, tracing them in several different theoretical traditions, including European and American structuralism, the later philosophy of Ludwig Wittgenstein, corpus linguistics, and psychology. Section 1.2 discusses the place of distributional semantics in theoretical and computational linguistics.

1.1 The Distributional Hypothesis

Distributional semantics was born in the early 1960s within the emerging field of computational linguistics. One of the first explicit mentions of this term is in Garvin (1962), who defines it as follows:

> Distributional semantics is predicated on the assumption that linguistic units with certain semantic similarities also share certain similarities in the relevant environments. [. . .] it may be possible to group automatically all those linguistic units which occur in similarly definable environments, and it is assumed that these automatically produced groupings will be of semantic interest. (Garvin, 1962, p. 388)

This definition already contains all the essential ingredients of distributional semantics, in particular its grounding assumption:

https://doi.org/10.1017/9780511783692.002 Published online by Cambridge University Press

Distributional Hypothesis
The semantic similarity between two lexemes is a function of the similarity of their linguistic contexts.

DISTRIBUTIONAL HYPOTHESIS

The Distributional Hypothesis is a tree with many branches and multifarious roots. There are at least three different theoretical soils from which the Distributional Hypothesis has sprung: American structuralism (Section 1.1.1), the writings of later Wittgenstein (Section 1.1.2), and corpus linguistics (Section 1.1.3). These are very different theoretical traditions, but they share a descriptive perspective on language, and they all emphasize the importance of language use as the primary datum in linguistic theory. In Section 1.1.4, we illustrate the influence of distributionalism in psychology and the interpretation of the Distributional Hypothesis as a cognitive principle.

1.1.1 The Distributional Methodology in Structural Linguistics

The history of the Distributional Hypothesis predates computational linguistics, and originates outside the realm of traditional semantics. Its main root lies in the **distributionalism** advocated by American structuralists as the central method for making linguistics an empirical science. Prominent figures in this tradition include Bernard Bloch, Archibald Hill, Charles Hockett, Martin Joos, and George Trager, in addition to **ZELLIG S. HARRIS** (1909–1992), who is widely recognized as one of the fathers of mathematical approaches to linguistics, and the most influential theoretician of distributionalism. Harris' distributional program is delineated in his *Methods in Structural Linguistics* (1951), and is a consistent topic throughout works such as *Distributional Structure* (1954), *Mathematical Structures of Language* (1968), and *A Theory of Language and Information: A Mathematical Approach* (1991).

DISTRIBUTIONALISM

According to Harris, the basic elements of language can be identified in terms of their relative distributions:

> To be relevant these elements must be set up on a distributional basis: x and y are included in the same element A if the distribution of x relative to the other elements B, C, etc. is in some sense the same as the distribution of y. (Harris, 1951, p. 7)

The essence of the distributional methodology, as defined by Harris in the quote above, is thus quite clear and simple: The basic building blocks of language can be identified by their relative distributions in (samples of) language. This means that if we have two elements x and y with identical distributions, then they are functionally equivalent and should be regarded as the same

https://doi.org/10.1017/9780511783692.002 Published online by Cambridge University Press

distributional element. Of course, everything depends on what we mean by "distribution." Harris clarifies his use of the term in the following words:

> The ENVIRONMENT or position of an element consists of the neighborhood, within an utterance, of elements which have been set up on the basis of the same fundamental procedures which were used in setting up the element in question. [...] The DISTRIBUTION of an element is the total of all environments in which it occurs, i.e. the sum of the (different) positions (or occurrences) of an element relative to the occurrence of other elements. (Harris, 1951, pp. 15–16)

The term "environment" refers here to the **linguistic context** of an element and is formed by its neighboring elements. Harris' statement above thus constitutes a very clear and concise formulation of the distributional methodology:

LINGUISTIC CONTEXT

DISTRIBUTION

> Linguistic elements are identified by their **distributions**, defined as the sum of the contexts in which they occur.

In the same way, categories of elements can be identified by the distributional similarity of their constituent elements. For Harris (and other proponents of the distributional methodology), the entirety of language – phonology, morphology, grammar – could be described according to distributional criteria.

As explicitly acknowledged by Harris, distributionalism originates in the pioneering works of **EDWARD SAPIR**[1] (1884 – 1939) and **LEONARD BLOOMFIELD** (1887 – 1949), whose program for a **structural** and **descriptive** linguistics is founded on three main tenets:

STRUCTURAL LINGUISTICS

1. every language has a structure of its own, and there are no universal linguistic categories;
2. the study of language must be primarily synchronic;
3. linguistics must be autonomous with respect to other disciplines, especially psychology.

The distributional method is considered instrumental in achieving these goals, in particular to provide linguistics with an independent and methodologically sound foundation: The only data that are scientifically valid for Bloomfield are observable linguistic phenomena in the form of distributional patterns.

[1] Quoting a letter by Morris Swadesh, Nevin (1993) suggests that the origin of the use of the term distribution in linguistics was Sapir, who employed it as a geographical metaphor: "It was an application of the usage represented by 'geographic distribution', an expression which was much used by Sapir as by other anthropologists and linguists."

https://doi.org/10.1017/9780511783692.002 Published online by Cambridge University Press

Distributionalism and Meaning

It is a common conception that semantics is in the periphery, if not completely ignored, in the structural linguistic tradition in general, and in the works of Zellig Harris in particular.[2] It is certainly true that Harris himself does not offer a distributional semantics and that his distributional project is primarily occupied with phonology and morphology, but the reason for this seemingly agnostic stance toward semantics is less commonly understood.

In order to fully appreciate Harris' perspective of the relationship between distributional properties and meaning, it is useful to first flesh out the ultimate conclusion of Harris' concerns about the scientific status of linguistic methodology. At the core of these concerns is the realization of the peculiar position of linguistics as a science, since it does not have recourse to a metalanguage that is external to its object of study. On the contrary, *language contains its own metalanguage*. We cannot describe language in something other than language, and any use of symbols needs to be defined ultimately in language:

> There is no way to define or describe the language and its occurrences except in such statements said in that same language or in another natural language. Even if the grammar of a language is stated largely in symbols, those symbols will have to be defined ultimately in a natural language. (Harris, 1991, p. 274)

This quandary has an important consequence: If the information in language can only be described in language, then it follows that this information cannot be an encoding of some prior representation of the information (e.g., a mental representation). This is a strong argument against mentalism, and for the scientific viability of the distributional approach. For Harris, this means that the science of language cannot deal with anything other than the elements of language, and their relationships to one another, that is, their distribution.

Following this line of reasoning, Harris argues – just like Bloomfield – that the form of a lexeme is not something different from the meaning conveyed by it. In the words of Bloomfield (1943): "in language, forms cannot be separated from their meanings." This is a point in the descriptive tradition that is often misconstrued. The insistence on the unity of form and meaning should not be understood as a denial of the existence of extralinguistic meaning. On the contrary, Harris – and even more explicitly Bloomfield – vigorously argues that meaning in all its social manifestations is most decidedly outside the scope of linguistic theory. As Bloomfield states: "the statement of meanings is therefore the weak point in language study, and will remain so until human knowledge advances far beyond its present state" (Bloomfield, 1933, p. 140). The best we

[2] Some commentators even label the whole American structuralist tradition "anti-semantic" (e.g., Murphy, 2003).

https://doi.org/10.1017/9780511783692.002 Published online by Cambridge University Press

can do within the descriptive linguistic project is to describe the observable manifestations of meaning, and indeed, *any* meaning (regardless of what it is and where it comes from) that can be conveyed in language *must* have a formal manifestation, since otherwise it would not be expressible in language: "a language can convey only such meanings as are attached to some formal feature: the speakers can signal only by means of signals" (Bloomfield, 1933, p. 168). Therefore, Bloomfield concludes, "in all study of language we must start off from forms and not from meanings" (Bloomfield, 1943, p. 402). The proper interpretation of this claim is that semantic considerations cannot enter into the definition of linguistic elements (e.g., "nouns denote things"), which must instead be defined in distributional terms (Goldsmith and Huck, 1991).[3]

Likewise, according to Harris, linguistic analysis cannot be founded on "some independently discoverable structure of meaning" (Harris, 1954, p. 152). It is meaning that must be studied as a function of linguistic distributions:

MEANING AND DISTRIBUTION

> if we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with difference of distribution. (Harris, 1954, p. 156)

Harris' words echo those of another structuralist, Martin Joos, who claims that "the linguist's meaning of a morpheme [...] is by definition the set of conditional probabilities of its occurrence in context with all other morphemes" (Joos, 1950, p. 708). The point is that a difference in meaning between two lexemes will be reflected by a difference in distribution, and this difference in distribution will be observable through distributional analysis:

> If A and B have almost identical environments except chiefly for sentences which contain both, we say they are synonyms: *oculist* and *eye-doctor*. If A and B have some environments in common and some not (e.g. *oculist* and *lawyer*) we say that they have different meanings, the amount of meaning difference corresponding roughly to the amount of difference in their environments. (Harris, 1954, p. 157)

In his later works, Harris characterizes linguistic environments in terms of syntactic dependencies involving relations between a word acting as **operator** and a word acting as its **argument**. The "selection" (i.e., the distribution) of a word

OPERATOR AND ARGUMENT

[3] Bloomfield's semantic skepticism concerns any approach to meaning, including the behaviorist and physicalist ones that he favored, since statements about meaning lie well beyond the limits of linguistic science: "There is nothing in the structure of morphemes like *wolf*, *fox*, and *dog* to tell us the relation between their meanings. This is problem for the zoölogist. The zoölogist's definition of these meanings is welcome to us as a practical help, but it cannot be confirmed or rejected on the basis of our science" (Bloomfield, 1933, p. 162).

https://doi.org/10.1017/9780511783692.002 Published online by Cambridge University Press

is the set of operators and arguments with which it co-occurs with a statistically significant frequency, and is strongly correlated with its meaning:

> It is thus that selection can be considered an indicator, and indeed a measure, of meaning. Its approximate conformity to meaning is seen in that we can expect that for any three words, if two of them are closer in meaning to each other than they are to the third, they will also be closer in their selection of operators and arguments. (Harris, 1991, p. 329)

Meaning "is a concept of no clear definition" (Harris, 1991, p. 321), but distributional analysis can turn it into a measurable and scientific notion:

> Selection is objectively investigable and explicitly statable and subdividable in a way that is not possible for meanings – whether as extension and referents or as sense and definition. (Harris, 1991, p. 329)

The goal of Harris' distributional program is therefore not to exclude meaning from the study of language (Harris, 1991, pp. 42–43), but rather to provide a scientific foundation for its investigation. Even if Harris has never explicitly formulated the Distributional Hypothesis, he argues that if we are to deal with meaning in language, *we can only do so through distributional analysis*. It is this idea of a correlation between meaning differences and distributional properties that lies at the heart of distributional semantics.

Syntagms and Paradigms: Distributionalism in Europe

Distributionalism is a direct product of American structuralism but is also strongly indebted to European structuralists like Ferdinand de Saussure, Louis Hjelmslev, and the Prague School, most notably represented by Nikolai Trubetzkoy and Roman Jakobson. According to Harris, the (semantic) relation between two words or morphemes is defined differentially, based on their distributional behavior within the language system, without recourse to an external world. This view recalls the words of the father of structuralism, **FERDINAND DE SAUSSURE** (1857–1913): "dans la langue il n'y a que des differences" ("in language there are only differences"; Saussure, 1916, p. 166).

VALUE AND SYSTEM

In structuralist theory, as it emanates from Saussure's posthumously published seminal work *Cours de linguistique générale* (1916), the term **valeur** "value" is used to define the function of a lexeme within the **language system**. A lexeme has a value only by virtue of being *different* with respect to the other lexemes. Such a differential view on the functional distinctiveness of linguistic elements highlights the importance of the system as a whole, since differences cannot exist in isolation from the system itself. A single isolated lexeme cannot enter into difference relations, since there are no other lexemes to differ from (and no system to define it functionally). In this view, the language system

https://doi.org/10.1017/9780511783692.002 Published online by Cambridge University Press

Table 1.1 Syntagmatic and paradigmatic relations

	Paradigmatic relations Selections: "x or y"			
Syntagmatic relations Combinations: "x and y"	she	adores	green	paint
	he	likes	blue	dye
	they	love	red	colour

becomes an interplay of functional differences, which can be divided into two kinds: **syntagmatic** and **paradigmatic relations**.[4]

SYNTAGMATIC RELATIONS

Syntagmatic relations hold between lexemes that co-occur in sequential combinations. A **syntagm** is such an ordered combination of lexemes.

PARADIGMATIC RELATIONS

Paradigmatic relations hold between lexemes that do *not* themselves co-occur, but that co-occur with the same *other* lexemes. Paradigmatically related lexemes can be substituted for one another in the same context. Such a set of substitutable lexemes constitutes a **paradigm**.

The term "syntagm" corresponds to what we have called "context." The only difference is that the former term implies an ordered set of neighboring lexemes, while the latter term does not. Syntagmatic and paradigmatic relations can be depicted as orthogonal axes in a two-dimensional grid. In the example in Table 1.1, the paradigms correspond to morphosyntactic classes, like adjectives and verbs, but they also define semantic categories, such as color terms.

In his essay *On linguistic aspects of translation*, **ROMAN JAKOBSON** (1896–1982) argues that meaning is a linguistic phenomenon:

> The meaning of the words "cheese," "apple," "nectar," "acquaintance," "but," "mere," and of any word or phrase whatsoever is definitely a linguistic – or to be more precise and less narrow – a semiotic fact [...]. There is no *signatum* without *signum*. The meaning of the word "cheese" cannot be inferred from a nonlinguistic acquaintance with cheddar or with camembert without the assistance of the verbal code. (Jakobson, 1959, p. 232)

Like for De Saussure, words have meaning only within a linguistic system, in which they are used and entertain various relations with other expressions. It is

[4] Saussure uses the term *associative* relation rather than paradigmatic relation. It was Hjelmslev who introduced the term "paradigmatic" relation.

https://doi.org/10.1017/9780511783692.002 Published online by Cambridge University Press

LINGUISTIC AND DIRECT ACQUAINTANCE

the knowledge of such relations that Jakobson calls **linguistic acquaintance**, whose importance supersedes the role of the **direct acquaintance** with the entities words refer to. The latter may lack (e.g., we can use *ambrosia* correctly even without direct experience of its referent), while linguistic acquaintance is essential to understand the meaning of any lexeme (cf. Sections 8.7–8.8).

STRUCTURAL SEMANTICS

Compared to its American counterpart, European structuralism attributes considerable importance to word meaning analysis. **Structural semantics**, represented by Jost Trier, Adrienne Lerher, Eugenio Coseriu, Algirdas Greimas, John Lyons, Alan Cruse, among many others, is a family of theories focusing on the paradigmatic organization of the lexicon (Murphy, 2003; Geeraerts, 2010). The theoretical apparatus of structural semantics includes:

SEMANTIC FIELDS

1. **lexical** or **semantic fields**, sets of mutually related lexemes defining the conceptual structure of a certain domain, such as the color domain;

SEMANTIC FEATURES

2. **semantic components** or **features**, inspired by structuralist phonology and used to describe meaning in terms of basic oppositions (e.g., +/– ANIMATE); and

SEMANTIC RELATIONS

3. **paradigmatic semantic relations** between lexemes, such as **synonymy** (sameness in meaning; *sofa* – *couch*), **antonymy** (opposition in meaning, *good* – *bad*), **hypernymy** (a taxonomic relation where a **hypernym** is a more general term than its **hyponym**, *animal* – *dog*), and **co-hyponymy** (lexemes that share the same hypernym; *dog* – *cat*).

Structural semantics is autonomous from distributionalism, but the latter is often adopted as a method to define semantic paradigms in terms of syntagmatic relations. The Distributional Hypothesis can indeed be reformulated in structuralist terms (Sahlgren, 2006):

> Words sharing syntagmatic contexts have similar paradigmatic properties.

For instance, Apresjan (1966) refers to Harris' distributional methodology as a way to provide more objectivity to the investigation of semantic fields by grounding it on linguistic evidence. Apresjan carries out a distributional analysis of adjectives in terms of their frequency of co-occurrence with various syntactic contexts. The interplay between syntagmatic and paradigmatic dimensions is also central for Cruse (1986): The greater the paradigmatic "affinity" of lexical items, the more congruent their patterns of syntagmatic relations.

1.1.2 Meaning as Use: The Echoes of Wittgenstein

The central principle of structuralism and the distributional methodology – that we should let data decide what our models of language encompass – echoes in

https://doi.org/10.1017/9780511783692.002 Published online by Cambridge University Press

Wittgenstein's insistence that we should "look and see" (Wittgenstein, 1953, §66) rather than presume. The intellectual work of **LUDWIG WITTGENSTEIN** (1889–1951) can be divided into two distinct periods: the early period of the *Tractatus Logico-Philosophicus* (1922), where he professes a logic-centered view on language, and the later period of *Philosophical Investigations* (1953), in which he explicitly rejects his earlier ideas about the nature of language. In this work, Wittgenstein is openly polemic against the view that we need a logical representation to obliterate the vagueness and incompleteness of natural language. Wittgenstein urges us not to *assume* a general and fixed meaning of words. Instead, we should look at *how* the words are being used:

> For a large class of cases – though not for all – in which we employ the word "meaning" it can be defined thus: the meaning of a word is its use in the language. (Wittgenstein, 1953, §43)

MEANING AS USE

This has sometimes been called a **usage-based theory** of meaning, but Wittgenstein is not so much offering a theory of meaning in his later works as pointing out a misconception regarding the nature of meaning. The misconception, according to him, consists in construing meaning as primarily a naming relation, in such a way that meaning is something (like a mental or physical object) that a word (or phrase or sentence) names. Such a *nomenclaturist* view on meaning has been both widespread and withstanding in the history of linguistics and the philosophy of language (although Wittgenstein's aim was primarily to attack his own earlier views), and the *anti-nomenclaturist* stance is a position the later Wittgenstein shares with the structuralist movement, and in particular with the contemporary Saussure (Harris, 1988).

This is not the only point of contact between the ideas of Wittgenstein and structuralist linguistics, and more specifically with the ideas of Saussure.[5] The former's view on meaning as founded in the use of language has striking similarities to the latter's concept of *valeur*: It is the *role* of the lexeme in language that constitutes its meaning. Wittgenstein even expresses himself in terms that could as well have been Saussure's: "the sign (the sentence) gets its significance from the system of signs, from the language to which it belongs" (Wittgenstein, 1958, p. 5). For both Wittgenstein and Saussure, meaning can be likened with the role or function of a word within language; indeed, they make heavy use of the game-metaphor – and in particular chess – for describing the holistic functional character of the language system.

Wittgenstein also stresses the importance of the social aspect of meaning and language use, just as the prominent figures of the structuralist tradition

[5] Despite the similarities between the ideas of Wittgenstein and Saussure, there is no evidence that they were influenced by (or even aware of) each other's works (Harris, 1988).

https://doi.org/10.1017/9780511783692.002 Published online by Cambridge University Press

had done. Wittgenstein even argues that language cannot exist in isolation from a language community. This is the essence of the so-called "private language argument," in which we are invited to imagine someone inventing a private language for naming a private sensation. One of the problems with such a private language is that there would be no criterion of correctness for using the private name: How could the private language user tell whether she uses the name in the correct way or not? Wittgenstein's point is that in order for there to be a criterion of correctness of use, there must be other language users that agree on this criterion, since correctness in language is established by convention. Wittgenstein coined the term **language game** to emphasize that language use is a *social activity* that requires other users, and for which there are *rules* that are established by convention between the language users, and which determine correctness of usage. The point is that the rules of the language game define the language system, and so are a prerequisite for meaning: Lexemes simply cannot have a meaning if there is no language game to be played. Therefore, the meaning of lexemes can be understood only by observing how they are used in language games.

LANGUAGE GAMES

In summary, although there is no concrete evidence that Wittgenstein had any direct influence on the development of distributionalism as formulated by the American structuralist tradition, there are some striking similarities between his usage-based view of meaning and the Distributional Hypothesis. Both adopt a descriptive perspective on language and emphasize the importance of usage data as the primary source of information for semantic analysis. Of course the term "use" in Wittgenstein does not refer only to linguistic distributions but to the more general usage in communicative situations, which include but are not limited to linguistic contexts. However, Wittgenstein's view of meaning strongly resonates with the one grounding distributional semantics.

1.1.3 Distributionalism and Corpus Linguistics

If Wittgenstein's influence on structuralism, and in particular the American distributionalists, remains obscure, it is much more explicit when it comes to European corpus linguistics. The idea that language use and distributional analysis is the key to understanding word meaning has flourished within the linguistic tradition stemming from the British linguist **JOHN R. FIRTH** (1890–1960). In fact, **corpus linguistics** represents another important root of distributional semantics. Firth laments the lack of interest in meaning by American structuralists, but he shares with them the idea that linguistics should address meaning in its own terms, the question being "not how much meaning can be *excluded*, but how much meaning can legitimately be included" (Firth,

CORPUS LINGUISTICS

https://doi.org/10.1017/9780511783692.002 Published online by Cambridge University Press

1955, p. 102). Differently from structural semanticists, Firth privileges the analysis of syntagmatic relations between lexical items over the paradigmatic ones.

Firth's **contextual view of meaning** is based on the assumption that meaning is a very complex and multifaceted reality, inherently related to language use in contexts (e.g., social setting, discourse, etc.). One of the key "modes" of meaning of a word is what he calls "meaning by collocation" (Firth, 1951), determined by the context of surrounding words:

CONTEXTUAL VIEW OF MEANING

> As Wittgenstein says, "the meaning of words lies in their use." The day to day practice of playing language games recognizes customs and rules. It follows that a text in such established usage may contain sentences such as "Don't be such an ass!," "you silly ass!," "What an ass he is!" In these examples, the word *ass* is in familiar and habitual company, commonly collocated with *you silly –*, *he is a silly –*, *don't be such an –*. You shall know a word by the company it keeps! (Firth, 1957, p. 11)

Collocations are lexical items that tend to co-occur in the same linguistic context: "Collocations of a given word are statements of the habitual or customary places of that word in collocational order [. . .] it is an order of *mutual expectancy*. The words are mutually expectant and mutually prehended" (Firth, 1957, p. 12). The meaning of a lexeme is thus defined by its **collocates**, other lexemes that have a syntagmatic relation with it: "One of the meanings of *night* is its collocability with *dark*, and of *dark*, of course, collocation with *night*" (Firth, 1951, p. 196). The analogy with Harris' claim on the relationship between meaning and linguistic distributions is patently very strong.

COLLOCATIONS

COLLOCATE

JOHN M. SINCLAIR (1933–2007), one of the fathers of corpus linguistics and of modern computational lexicography, deeply elaborates Firth's idea of the centrality of collocations to describe lexical meaning:

> the formal meaning of an item A is that it has a strong tendency to occur nearby items B, C, D, less strong with items E, F, slight with G, H, I, and none at all with any other item. (Sinclair, 1966, p. 417)

Like Harris, Sinclair uses the term "environments" to refer to the linguistic contexts of lexemes. Semantic analysis must start from the collection of the "environments" of a lexeme in a corpus. Since not all of them are equally important to characterize word meanings, significant collocations are distinguished from "casual" ones with statistical tests applied to the frequency distributions of collocates (Jones and Sinclair, 1974). Sinclair thus pioneered the use of computational techniques to extract collocations from corpora. The theoretical concept of collocation was introduced by Firth, but it is Sinclair

https://doi.org/10.1017/9780511783692.002 Published online by Cambridge University Press

who turned it into a quantitative method for semantic analysis. The study of collocations has grown as an independent line of research, but its theoretical assumptions and methods are deeply intertwined with distributional semantics (cf. Chapter 2).

1.1.4 The Distributional Hypothesis in Psychology

As argued by Goldsmith (2005), Harris' structuralist program is perfectly compatible with the interpretation of distributional analysis as a cognitive process:

> Harris was interested in determining what procedures IN PRINCIPLE could lead to a deep understanding of a natural language system, so it shouldn't be surprising that the one existing system that actually acquires a natural language should display a set of behaviors that resemble in interesting ways a Harrisian system. (Goldsmith, 2005, p. 729)

Indeed, Harris' distributional methodology had a significant impact on psychology. At the beginning, it was mainly regarded as a way to explain the strength of **word associations** produced by subjects. The word association technique is a common method in psychology, and consists in asking a subject to respond to a stimulus word (e.g., *dog*) with the first word that occurs to him or her (e.g., *cat*). Association strength is then measured by counting the number of subjects that have produced a given word in response to a stimulus.

WORD ASSOCIATIONS

BEHAVIORISM

The analysis of word associations plays a central role in **behaviorist psychology**, which pursues an associationist view of meaning, based on the idea that simple association or co-occurrence of stimuli is the primary basis of thought and learning. Jenkins (1954) suggests that word associations can be interpreted as a result of the statistical distribution of stimuli and responses in language. In particular, the association strength would depend both on **paradigmatic** and **syntagmatic similarity** between stimulus and response. Jenkins defines these notions in structuralist terms, explicitly referring to Harris:

> The similarity between any two words can be conceived linguistically as the degree of similarity in distribution. [. . .] this similarity may be profitably divided into two classes, *paradigmatic* and *syntagmatic*. Two words are considered paradigmatically similar to the extent that they are substitutable in the identical frame (this corresponds rather closely to Zellig Harris' use of the term "selection") and syntagmatic to the extent that they follow one another in utterances. (Jenkins, 1954, p. 11)

ASSOCIATIVE MEANING

Analogously, Deese claims that the **associative meaning** of a word, defined in terms of the responses it evokes on a word association test, depends on its distributional properties: "the extent to which words share associative

https://doi.org/10.1017/9780511783692.002 Published online by Cambridge University Press

distributions is determined by the extent to which they share contexts in ordinary discourse" (Deese, 1965, p. 128).

Vector-based representations of meaning, like those later adopted in distributional semantics, were pioneered in psychology by Charles Osgood, who is one of the first to refer to the semantic system as a **semantic space**. Osgood (1952) and Osgood et al. (1957) represent concepts in terms of n-dimensional feature vectors. However, the dimensions of Osgood's semantic spaces are not distributional,[6] but are built with the **semantic differential** method: Subjects are asked to locate the meaning of a word along different scales between two polar adjectives (e.g., *happy – sad*, *slow – fast*, *hard – soft*, etc.), and their ratings determine its position in the semantic space, which mainly captures connotative aspects of meaning. Such feature vectors are then used to measure the psychological distance between words.

SEMANTIC SPACE

SEMANTIC DIFFERENTIAL

The interest of psychologists in distributionalism survives the crisis of the behaviorist paradigm, and the consequent downfall of associationism. A strenuous supporter of the importance of linguistic distributions in shaping semantic representations is **GEORGE A. MILLER** (1920–2012), one of the fathers of **cognitive psychology**. Miller is deeply acquainted with contemporary structuralist linguistic theories (Miller, 1954) and considers Harris' distributional analysis as a method to provide an empirical foundation to the notion of **semantic similarity** (Miller, 1967). Judgments of semantic similarity between words (e.g., *dog* is semantically more similar to *cat* than to *car*) play a key role in the exploration of the mental lexicon, and they are routinely used as an explanatory factor in psychological experiments.

COGNITIVE PSYCHOLOGY

SEMANTIC SIMILARITY

Rubenstein and Goodenough (1965) carry out computational experiments showing that semantic similarity judgments on 65 noun pairs strongly correlate with the overlap of the linguistic contexts of the two words. The contexts are collected from sentences produced by a different group of subjects for each noun in the test pairs. Miller (1967) also distinguishes syntagmatic distributional similarity from paradigmatic one, and mentions the ongoing research in computational linguistics to measure semantic similarity with distributional data automatically extracted from corpora (cf. Section 1.2.1).

A distributional definition of semantic similarity is theorized by Miller and Charles (1991), who conceive it as a "function of the contexts in which words are used" (p. 3). Like Firth, they advocate a contextual view of meaning:

> What people know when they know a word is not how to recite its dictionary definition – they know how to use it (when to produce it and how to understand it) in

[6] Osgood et al. (1957) are actually quite critical of Harris' distributional methodology for semantic analysis.

https://doi.org/10.1017/9780511783692.002 Published online by Cambridge University Press

everyday discourse [. . .] And because words are used together in phrases and sentences, this starting assumption directs attention immediately to the importance of context. (Miller and Charles, 1991, p. 4)

Even if context in a broad sense must also include the extra-linguistic information about the communicative and social setting, Miller and Charles claim that speakers are able to acquire many new words only using distributional information (cf. Section 8.8). The repeated observations of a word in linguistic contexts lead to the formation of its **contextual representation**:

CONTEXTUAL REPRESENTATION

The *contextual representation* of a word is knowledge of how that word is used. [. . .] That is to say, a word's contextual representation [. . .] is an abstract cognitive structure that accumulates from encounters with the word in various (linguistic) contexts. (Miller and Charles, 1991, p. 5)

Contextual representations correspond to the distributional representations we introduce in Chapter 2. Judging the semantic similarity of two words thus consists in comparing the similarity of their contextual representations. This is what Miller and Charles (1991) call **contextual hypothesis**, stating that "two words are semantically similar to the extent that their contextual representations are similar" (p. 5). The contextual hypothesis is related to Harris' distributional methodology, a debt that Miller and Charles explicitly acknowledge.

CONTEXTUAL HYPOTHESIS

1.2 Distributional Semantics in Language Research

Because of its different roots, distributional semantics is a manifold program for semantic analysis, which is pursued in different disciplines, like computational linguistics and cognitive science. This contributes to its being a framework with multiple souls and goals, which, however, share common methods. In fact, we can identify two major views of distributional semantics that also correspond to alternative ways to interpret the Distributional Hypothesis.

WEAK DISTRIBUTIONAL HYPOTHESIS

First of all, distributional semantics is an empirical methodology for semantic analysis. This view is based on a **Weak Distributional Hypothesis** (Lenci, 2008) that postulates a correlation between semantic content and linguistic distributions, and exploits such correlation as an "observable" of meaning. The distribution of words in contexts is determined by their meaning (*whatever this might be*) and the semantic properties of lexical items act as constraints governing their syntagmatic behavior. Consequently, by inspecting a relevant number of distributional contexts, we can identify those aspects of meaning

https://doi.org/10.1017/9780511783692.002 Published online by Cambridge University Press

that are shared by lexemes with similar linguistic distributions. The Weak Distributional Hypothesis does not entail that word distributions are themselves constitutive of the semantic properties of lexical items at a cognitive level, but rather that meaning is a kind of "latent variable" responsible for the distributions we observe, which we try to uncover by analyzing such distributions.

> Distributional semantics is a theoretical and computational framework to **learn and study the semantic properties of lexemes from their distribution in linguistic contexts** collected from text corpora.

The investigation of lexical meaning typically relies on two kinds of evidence: (i) native speakers' intuitions about the semantic properties of linguistic expressions and (ii) the description of meanings in lexical resources, like dictionaries. Distributional semantics adds a third type of empirical evidence: *The computational analysis of lexeme distributions in linguistic contexts*.

Datasets with human semantic judgments (e.g., similarity ratings) are commonly used in cognitive science, but collecting them is a complex and time-consuming task (e.g., rating 100 words for their pairwise semantic similarity amounts to collecting $4,950$ judgments per participant, which raise to $19,900$ for 200 words). Although crowdsourcing methods facilitate the elicitation process, collecting speakers' semantic intuitions can hardly scale up to cover large lexical samples. Computational lexicons like WordNet (Fellbaum, 1998) provide important information about word senses and their organization. A major limit of such resources is that they contain "second-hand" evidence, as the organization of the semantic space heavily depends on the lexicographers' choice. Moreover, they have a limited coverage, are hand-built, and therefore hard to maintain and extend to new domains. In fact, the lexicon is a dynamic entity, with new items and new senses constantly appearing. Lexical meanings are highly context sensitive and undergo continuous modification and modulation in contexts (cf. Section 9.1.1).

Given the ever-increasing availability of digital texts, distributional semantics can rely on huge amounts of empirical evidence to characterize the semantic properties of lexemes. Building distributional models for large samples of the lexicon is fast and cheap, at least comparatively to other methods of collecting semantic information, and can be performed for any language or domain, as long as we have enough textual data. From a computational linguistic perspective, distributional semantics is an efficient and effective method to build corpus-based lexical resources, and to learn semantic representations for NLP and AI systems. Moreover, since distributional models are

https://doi.org/10.1017/9780511783692.002 Published online by Cambridge University Press

grounded in language usage, they are more suitable to capture its variability and dynamicity, thereby offering new perspectives to analyze the complex interplay between meaning and context. Obviously, distributional data do not replace other types of semantic evidence but rather complement them. For instance, speakers' judgments provide benchmarks for the evaluation of distributional models (cf. Chapter 7), which can in turn be employed to expand lexical resources.

According to a second view, distributional semantics is a methodology to investigate and model how meanings are acquired and represented in the mental lexicon. This conception directly stems from the psychological research we have reviewed in Section 1.1.4 and is grounded on the assumption that the linguistic contexts of a word have a causal role in creating and shaping its neurocognitive representation (cf. the Contextual Hypothesis by Miller and Charles, 1991). We call this the **Strong Distributional Hypothesis** (Lenci, 2008), since it regards the distributional behavior of a lexeme as an explanatory factor of its cognitive properties.

STRONG DISTRIBUTIONAL HYPOTHESIS

> Distributional semantics is a theoretical and computational framework to **build models of semantic memory** based on the hypothesis that the distribution of words in the linguistic input contributes to determine their **conceptual representations**.

Semantic memory stores concepts and general world knowledge as mental representations that allow us to recognize entities in the world, interact with them, and interpret language (McRae and Jones, 2013). Semantic memory is shaped by our *experience*, which includes both **sensory-motor experiences** of perceiving and acting in the world, and **linguistic experiences** of using and being exposed to language. Distributional models of semantic memory "hypothesize a formal cognitive mechanism to learn semantics from repeated episodic experience in the linguistic environment (typically a text corpus)" (Jones et al., 2015, p. 239). The contribution and role of linguistic experience vis-à-vis other kinds of extralinguistic inputs in building concepts is an empirical question that is widely debated in cognitive science (Vigliocco et al., 2009; Dove, 2014). Distributional semantics is a scientific framework to investigate the structure and origin of semantic representations in mind and brain (cf. Section 8.8).

Distributional semantics is both a method to represent meaning, and a family of computational models to learn such representations from linguistic data. Therefore, it allows us to explore a wide range of issues related to meaning dynamics, including its acquisition, change, and use. The actual descriptive and explanatory adequacy of distributional semantics is of course an empirical

https://doi.org/10.1017/9780511783692.002 Published online by Cambridge University Press

matter. It is one of the main purposes of this book to investigate this issue. In the following sections, we briefly review the past and present of distributional semantics in computational linguistics and semantic theory.

1.2.1 Computational Linguistics

Today, distributional semantics is a mainstream research paradigm in computational linguistics. However, this is just the last step of a long process whose beginnings date back to the early 1960s and were influenced by the cultural and scientific environment we have analyzed in Section 1.1. The first experiments in the distributional analysis of meaning were aimed at building thesauri for machine translation and information retrieval (Sparck Jones, 2005). A **thesaurus** is a lexical resource in which words are grouped together according to paradigmatic relations, like synonymy and hypernymy, or because they belong to the same semantic field. Thesauri were considered extremely useful to provide machine translation systems with semantic information about lexical items, but the experiments with existing resources like *Roget's Thesaurus* did not prove wholly satisfactory. Hand-made thesauri turned out to be extremely laborious to produce and suffered from limited coverage. Hence the idea of exploiting distributional information for the automatic identification of synonyms and the semantic classification of lexical items.

THESAURUS

Hays (1960) and Garvin (1962) use the term distributional semantics to refer to a research program in machine translation inspired by Harris' idea that similarity of meaning depends on similarity of linguistic contexts (cf. the quotation at the beginning of Section 1.1). Harper (1961, 1965) provides experimental tests of such program, by measuring the similarity between 40 Russian nouns in terms of their syntactic dependencies extracted from a small corpus. Parallelly, Sparck Jones (1961, 1964) carries out experiments to identify synonyms with distributional information extracted from a machine-readable dictionary.

An essential contribution to the development of distributional semantics has come from the **Vector Space Model** in information retrieval (Salton et al., 1975), which was pioneered in the SMART system (Salton, 1964, 1971a). The core idea of the vector space model is to represent a collection of documents (i.e., texts to be retrieved) with a **term-document matrix**: The row vectors correspond to terms (i.e., lexical items), the column vectors correspond to documents, and each matrix entry records the occurrences of a term in a document (cf. Chapter 2). Similarity between documents is computed by measuring the similarity between their column vectors. Queries are treated as "pseudo-documents" and represented with column vectors in the same matrix, thus the

VECTOR SPACE MODEL AND INFORMATION RETRIEVAL

https://doi.org/10.1017/9780511783692.002 Published online by Cambridge University Press

relevance of a document to a query is computed by measuring the similarity of the document vector to the query vector.

Since its outset, the Vector Space Model was also applied to automatic thesaurus construction. In fact, one major problem that retrieval systems face is that the same information can be described with different terms.[7] This negatively impacts on the system recall (i.e., its ability to retrieve documents relevant with respect to the user's query), because the term in the query (e.g., *production of automobiles*) may not be the same one used to index a document (e.g., *manufacture of motor vehicles*). Various experiments were carried out to exploit co-occurrence statistics extracted from document collections to identify semantically associated words (Stiles, 1961; Stevens et al., 1964). Jones (1964) explicitly relates this research line to Harris' distributionalism on the one hand, and associationist psychology on the other (cf. Section 1.1.4). Giuliano and Jones (1962), Salton (1964), Dattola and Murray (1967), and Sparck Jones (1971) use the term-document matrices to compute term similarity with row vector similarity. Lewis et al. (1967) discriminate synonyms and antonyms from other kinds of related terms with distributional statistics. Instead of term frequency in a document, Hirschman et al. (1975) represent words with vectors recording their co-occurrences with particular grammatical relations in texts, and cluster the vectors to obtain semantic classes.

LINEAR ALGEBRA

The Vector Space Model in information retrieval has introduced **linear algebra** as the core mathematical framework for distributional semantics. Notions that have become standard in distributional models of meaning, like co-occurrence matrices and vector representations of lexical items, were already in place in the first researches in the 1960s. Unfortunately, the quality of the resulting thesauri and their effectiveness in applications were greatly hampered by the technological limitations of the time, in particular the small size of the corpora, and the computational cost required to build co-occurrence matrices and measure distributional similarity (Salton and Lesk, 1966; Salton, 1971b).

While distributional semantics continued to be pursued in information retrieval, it was virtually ignored in computational linguistics throughout this early period. Except for works on collocation analysis (Church and Hanks, 1989) and on the acquisition of lexical information from machine-readable dictionaries (Wilks et al., 1990), formal and logic methods dominated mainstream computational semantics, as proved by the fact that distributional semantic themes were practically absent from any major conference or journal. In the early 1990s, the new empiricist turn in computational linguistics

[7] According to Furnas et al. (1983), two people choose the same word to refer to the same object less than 20% of the time.

https://doi.org/10.1017/9780511783692.002 Published online by Cambridge University Press

and the emergence of statistical natural language processing, together with the availability of larger corpora and more powerful computers, favored a fast-growing interest in distributional semantics. Hindle (1990) is one of the first works of this new trend to explicitly mention Harris's distributional hypothesis. Hindle derives a distributional classification of words in English from their co-occurrences with syntactic relations automatically extracted from a parsed corpus.

LATENT SEMANTIC ANALYSIS

A major innovation was represented by **Latent Semantic Analysis** (LSA), also known as **Latent Semantic Indexing** (LSI), proposed by Deerwester et al. (1990). LSA extends the Vector Space Model of information retrieval by applying Singular Value Decomposition (SVD) to the term-document matrix. This linear-algebraic method allows a more sophisticated analysis of distributional data by projecting the co-occurrence matrix onto a new reduced one, with the purpose of finding higher-order associations between terms and documents and uncovering the "latent semantic structure" in the original matrix (cf. Section 2.5.1). Schütze (1992, 1997, 1998) and Schütze and Pedersen (1993) apply SVD to matrices recording co-occurrences between lexical items appearing within the same text window. While previous models almost exclusively aimed at the identification of similar terms for thesaurus construction, Schütze was one of the first to apply distributional methods to more advanced semantic problems, like word sense induction and disambiguation. Schütze's works, together with those by Gallant (1991), Ruge (1992), Pereira et al. (1993), Dagan et al. (1993), Grefenstette (1994), and Niwa and Nitta (1994) among many others, contributed to spreading distributional methods in computational semantics.

Research in distributional semantics has kept on growing steadily in the 1990s and the first decade of the new millennium. Most of the distributional models presented in Chapters 4 and 5 were created in this period, and the range of semantic tasks addressed with distributional methods has been increasing since then, and now includes topics like compositionality, inference, multimodality, semantic change, and several others (cf. Part III). Lately, a most significant breakthrough has occurred with the emergence and fast success of **deep learning** methods, which have dramatically changed computational linguistics and distributional semantics by developing a new generation of models based on **artificial neural networks** (cf. Chapter 6). Deep learning has also spread the use of the term **(word) embedding** for distributional vectors. The last years have witnessed a further significant novelty, with the appearance of so-called **contextual embeddings** generated by a new kind of deep neural models that represent each word token with a distinct, context-sensitive vector (cf. Chapter 3 and Section 9.6.3).

DEEP LEARNING

NEURAL NETWORKS

EMBEDDINGS

CONTEXTUAL EMBEDDINGS

https://doi.org/10.1017/9780511783692.002 Published online by Cambridge University Press

Deep learning has also radically modified the scope of distributional semantics itself, boosting an exponential growth of interest in this field. Neural networks represent words with vectors, and embeddings trained on larger corpora are nowadays routinely used in deep learning architectures to initialize their word representations. These **pretrained embeddings** allow neural networks to capture lexical semantic properties that are beneficial to carry out downstream supervised tasks. Pretrained vectors can be directly used as **features** in classification algorithms or **fine-tuned** to address specific tasks. The main novelty is that distributional semantics is no longer just a computational method to measure semantic similarity or to build lexical resources from corpora, but a general approach to provide NLP and AI applications with knowledge about the meaning of linguistic expressions.

PRETRAINED EMBEDDINGS

1.2.2 Semantic Theory

The research landscape in linguistics is characterized by two major semantic approaches that are not based on the distributional hypothesis: cognitive semantics and formal semantics. These theories present prima facie striking differences with distributional semantics, which, however, turn into important similarities or at least potential synergies at a closer and deeper look.

Cognitive linguistics is based on the work by Ronald Langacker, George Lakoff, Charles Fillmore, William Croft, Adele Goldberg, and many others, who argue for a **conceptualist** view of meaning. In **cognitive semantics**, the meaning of a lexical expression is a particular conceptualization of an entity or situation. Conceptual representations are conceived as inherently grounded in our physical embodiment: "The meaning of words in languages and how they can be used in combination depends on our perception and categorization of the world around us" (Ellis et al., 2016, p. 25). The central role of **grounding** and **embodiment** in cognitive semantics apparently contrasts with the main tenets of distributional semantics and its program of constructing meaning from linguistic co-occurrences. However, as we show in Sections 8.7 and 8.8, distributional semantics is not incompatible with grounded models of meaning.

COGNITIVE SEMANTICS

GROUNDING

An important commonality between distributional and cognitive semantics is the **usage-based** perspective. Many cognitive linguists advocate a **usage-based model** of language acquisition and change (Goldberg, 1995; Tomasello, 2003; Croft and Cruse, 2004; Goldberg, 2006; Bybee, 2010; Hoffman and Trousdale, 2013), according to which "use of language figures critically in determining the nature of cognitive representations of language, or put another way, usage events create linguistic structure" (Bybee, 2013, p. 68). Language is

USAGE-BASED MODELS

https://doi.org/10.1017/9780511783692.002 Published online by Cambridge University Press

viewed as a complex adaptive system whose structure is *emergent* from underlying, domain-general processes that operate in areas of human cognition other than language itself (Elman, 1998; MacWhinney, 1999; Beckner et al., 2009; MacWhinney and O'Grady, 2015). In distributional approaches, lexical representations emerge from co-occurrences with linguistic contexts (Ellis, 1998), and semantic spaces are built with domain-independent learning algorithms that record the distributional statistics in the linguistic input.

Moreover, cognitive linguists regard neural networks and connectionism as a computational paradigm implementing emergent and usage-based representations (Elman et al., 1996; Ellis, 1998; Bybee and McClelland, 2005; McClelland et al., 2010). The goal of **connectionism** and the **Parallel Distributed Processing** (PDP) approach (Rumelhart and McClelland, 1986) is to explain cognition with artificial neural networks (Jones et al., 2015), as domain-independent algorithms that learn representations from co-occurrence statistics across stimulus events in the environment (cf. Section 6.1). Connectionism is consistent with the distributional hypothesis, since linguistic co-occurrences are just a particular type of stimuli that can be used by neural networks. Landauer and Dumais (1997) and Schütze (1993) already give a connectionist interpretation of their models, and neural networks today are widely used in distributional semantics. A further element of convergence with cognitive semantics is its emphasis on linguistic categories characterized by gradience and prototype effects (Taylor, 1995), which can be modeled with continuous representations like distributional vectors (Acquaviva et al., 2020). In fact, distributional semantics nowadays has a growing number of applications in linguistics, to study polysemy, semantic change, productivity, selectional preferences, and so on (cf. Chapters 8 and 9).

CONNECTIONISM

If important "family resemblances" characterize distributional and cognitive semantics, the relationship with formal semantics is more complex and controversial. Stemming from the work by Gottlob Frege, Rudolf Carnap, Alfred Tarsky, Richard Montague, David Lewis, Hans Kamp, Barbara Partee, among many others, **formal (model-theoretic) semantics** is a rich family of models that share a **referential (denotational)** view of meaning (cf. Section 9.1). Its main assumption is that meaning is a relation between linguistic symbols and entities external to language, and that the goal of semantics is to characterize the **truth-conditions** of sentences as a function of the reference (denotation) of their parts. Lewis (1970) claims that "semantics with no treatment of truth conditions is not semantics" (p. 18), and Heim and Kratzer (1998) that "to know the meaning of a sentence is to know its truth conditions" (p. 1).

FORMAL SEMANTICS

MEANING AND TRUTH

The core notions of Frege's program for formal semantics – *truth*, *reference*, and *logical form* – are as different as possible from those of Harris' program

https://doi.org/10.1017/9780511783692.002 Published online by Cambridge University Press

for distributional semantics – *linguistic contexts*, *use*, and *co-occurrence statistics*. Formal and distributional semantics have indeed proceeded virtually ignoring each other, focusing on totally different semantic phenomena. As a matter of fact, a whole range of issues in the agenda of formal semantics, such as compositionality, quantification, inference, anaphora, modality, tense, and so on, have usually remained beyond the main horizon of distributional semantics, which has instead mostly concentrated on lexical meaning. Recently, the relationship between formal and distributional semantics has changed, and the barriers between these paradigms are now reducing. Distributional research has begun to explore the potential synergies with formal models of meaning and to address problems like compositionality, inference, and reference (Erk, 2013; Baroni et al., 2014b; McNally and Boleda, 2016; Chersoni et al., 2019; Boleda, 2020). The aim is to combine the effectiveness of distributional semantics in learning and representing word meaning with the capacity of formal models to account for compositional semantics and logical inference (cf. Chapter 9).

1.3 Summary

In this chapter, we have identified the origins of distributional semantics in structural linguistics, and in particular in the distributional methodology pioneered by Bloomfield and refined by Harris. We have also noted its close kinship with the philosophy of the later Wittgenstein and with Firthian corpus linguistics, and its impact on psychology. We have charted the course of distributional semantics in computational linguistics and cognitive science, and we have tried to articulate the position of distributional semantics in relation to current research in linguistics.

Our main findings from this journey through the history of distributional semantics can be summarized in the following way:

- its theoretical foundation is the **distributional hypothesis**;
- the distributional hypothesis is primarily a conjecture about **semantic similarity**, which is modeled as a function of **distributional similarity**. Semantic similarity is therefore the core notion of distributional semantics;
- the distributional hypothesis is primarily a conjecture about **lexical meaning**, so that the main focus of distributional semantics is on the **lexicon**; and
- distributional semantics is based on a **contextual** and **usage-based** view of meaning: The meaning of a lexeme is determined by the way it is used in linguistic contexts.

https://doi.org/10.1017/9780511783692.002 Published online by Cambridge University Press

1.4 Further Reading

- Distributional semantics and structural linguistics: Sahlgren (2006, 2008); Gastaldi (2021)
- Information retrieval and the Vector Space Model: Manning et al. (2008)
- General introductions to distributional semantics: Lenci (2008, 2018); Turney and Pantel (2010); Erk (2012); Clark (2015)
- Distributional and linguistic semantics: Geeraerts (2010); Acquaviva et al. (2020); Boleda (2020)

https://doi.org/10.1017/9780511783692.002 Published online by Cambridge University Press