Chapter 2

Natural Language Understanding and World Knowledge

2.1 What is Natural Language Understanding?

In artificial intelligence and computational linguistics, *natural language understanding* (NLU) is a subfield of natural language processing that deals with machine reading comprehension. The goal of an NLU system is to *interpret* an input text fragment. The process of interpretation can be viewed as a translation of the text from a natural language to a representation in an unambiguous formal language. This representation, supposed to express the text's content, is further used for performing concrete tasks implied by a user request.

One of the earliest known attempts in natural language understanding is the program *STUDENT* developed by Bobrow (1964) for his dissertation project. This was a program, which could solve high-school algebra problems posed in a natural language form. For example, *STUDENT* was able to answer the following query: "If the number of customers Tom gets is twice the square of 20% of the number of advertisements he runs, and the number of advertisements is 45, then what is the numbers of customers Tom gets?" *STUDENT* was processing queries by doing simple pattern-matching in order to translate the queries into a format suitable for an equation-solving program.

A couple of years later, Weizenbaum (1966) created *ELIZA*, an interactive computer program able to carry on a dialog on any topic. *ELIZA* was equipped with a simple parser; its "semantic" analysis was based on transformation of key words into canned patterns. Although *ELIZA* was just a toy project, it has gained an amazing popularity, especially in the area of psychotherapy. The popularization of *ELIZA* and similar programs has led to the appearance of the term "ELIZA effect"[1] describing the tendency of people to attribute human-like intelligence to chat robots. Indeed, a typical dialog with *ELIZA* may look quite impressive:

[1] See preface in Hofstadter (1996) for a description of ELIZA effect.

```
 USER: Men are all alike
ELIZA: In what way?
 USER: They're always bugging us about something or other
ELIZA: Can you think of a specific example?
 USER: My boyfriend made me come here
ELIZA: Your boyfriend made you come here
 USER: He says I'm depressed much of the time
ELIZA: I'm sorry to hear you are depressed
...
```

STUDENT and *ELIZ*A were among the first computer programs to approach human intelligence in language comprehension and they were followed by many others. But can we actually say that these programs were able to *understand* natural language? How can we judge it?

The "Holy Grail" for most AI applications is a human-like performance. In order to get a clearer picture of what we expect from the computer systems, let us consider a human language understanding scenario. The man on Fig. 2.1[2] asks the woman to give him a pen for writing a letter. She analyses his request applying available knowledge and tries to guess what kind of meaning he has in mind. Finally, she concludes that he is asking for a writing instrument (rather than, for example, for a portable enclosure, in which babies may be left to play) and passes it over to him. Can we say that the woman in this scenario understands the message? She correctly complies the speaker's request, but it could be done by chance. What would probably convince us in her understanding is a comparison of the interpretation of the message, which she has created with the meaning intended by the speaker. Unfortunately, for humans such a comparison is impossible, because human conceptualizations are not explicitly available. There are only indirect ways to evaluate understanding, namely, by comparing the interpreter's behavior with the predicted behavior.

A similar scenario is applicable to computer programs. If a program is doing what we expect it to do then we can say that it "understands" the input. For example, if a search system outputs information that we have been searching, we can say that the system "understands" our query. If a summarization system summarizes a document like a human would do it then we can say that the system "understands" the document.

[2]The pictures of the pens on Fig. 2.1 are provided by DesignContest (*http://www.designcontest.com*) under Creative Common licence vers. 3 (*http://creativecommons.org/licenses/by/3.0/*).

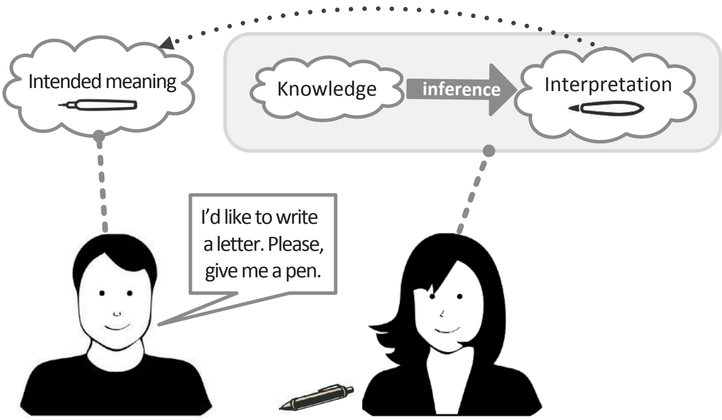

Fig. 2.1 Human natural language understanding.

For computer programs, there also exists another criterion of understanding. Working with computer programs, in contrast to human agents, we can look inside of processing and see what representation of the message content was created by the program. This automatically created representation can then be evaluated against human intuition about the content of the processed text fragment. Probably neither two speakers will arrive to the same conceptualization of the message content, because of the differences in individual experience. Thus, there will be as many conceptualizations as the number of readers comprehending the text fragment. However, all these conceptualizations will probably have a common part implied by the shared linguistic and conceptual knowledge of a language community rather than by individual experience. This shared part of the conceptualization is what we want our NLU system to grasp. The more information occurring in the shared conceptualization it can represent, the better it "understands" the text fragment, see Fig. 2.2 for an illustration.

Thus, there are two criteria for judging how well an NLU system "understands" a text fragment. The first criterion is performance-based. This evaluation strategy is realized in such series of test challenges as, for example, Text Analysis Conference (TAC[3]). The organizers of this challenge provide a large test collections for different NLP tasks and common evaluation procedures, which enable comparison of the output of NLP systems against hu-

[3] *http://www.nist.gov/tac/*

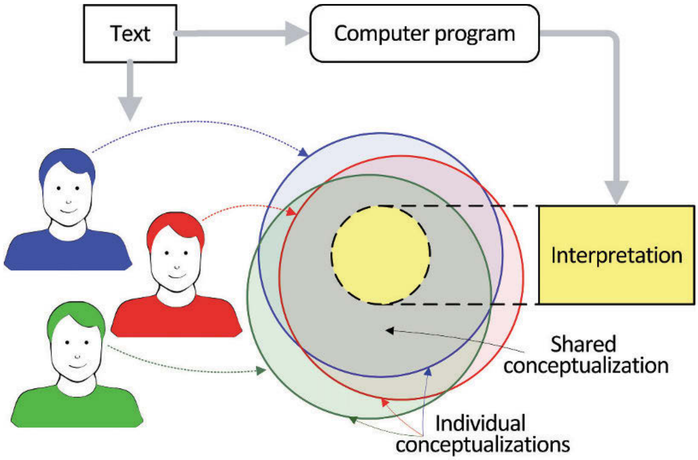

Fig. 2.2 Computational natural language understanding.

man solutions for the same tasks. The traditional TAC evaluation tasks are knowledge base population, recognizing textual entailment, and summarization.

The second evaluation criterion takes into account the intern representation of the text content created by the system. For testing NLP programs according to this criterion, automatically generated representations of a text are compared with human annotations of the same text. This evaluation strategy is realized, for example, in the serious of competitions called Semantic Evaluation (SemEval[4]) which mostly focus on such NLP tasks as word sense disambiguation, semantic role labeling, coreference resolution, subcategorization acquisition, etc.

2.2 Representation of Meaning

Talking about representations of the text content, one should regard the question about what an adequate representation of linguistic meaning should look like, which information it should carry, and how it can be constructed. There is no unique answer to these ques-

[4] *http://aclweb.org/aclwiki/index.php?title=SemEval_Portal*

tions; it depends on the underlying theory of linguistic meaning. Only abstract claims can be made about meaning representations in general. For example, the assumptions concerning representation of linguistic meaning made by Schank (1972) and Allen (1987) can be summarized as follows.

(1) If two sentences have the same/closely related meaning, they should be represented the same.
(2) The representation must be precise and unambiguous enough to enable us to express every distinct reading of a sentence as a distinct formula.
(3) Information implicitly stated in the sentence should be represented explicitly.

In this section, we review prominent approaches to natural language semantics in the fields of linguistics and artificial intelligence. Our goal is to discuss **what representations** were considered as suitable for expressing linguistic meaning and knowledge required for language understanding and **what information** was considered to be a part of meaning of natural language expressions. Special attention is payed to the **role of world knowledge** in different theories of linguistic meaning.

Before starting the discussion about theories of linguistic meaning, one needs to make a remark on terminology. Many of the basic terms in natural language semantics are highly ambiguous and presuppose a diverse understanding in different frameworks. Probably the most ambiguous term is "meaning" itself. One of the most influential definitions of this term was suggested by Frege who claimed that meaning is a relation between a sign, an object in the world, to which the sign refers, and a corresponding concept in someone's mind. Nowadays, not all researchers working on natural language semantics accept this definition. Especially in computational linguistics, the term "meaning" is often understood in its most general sense, as a relation between a linguistic sign and some non-linguistic entity, to which the sign refers; it can be a concept in the mind, an object in the world, a logical formula, etc. Since the work presented in this book does not rely on one single theoretical framework, but rather adopts ideas originating in totally different research paradigms, we adopt this fairly general understanding of "meaning".

2.2.1 *Meaning Representation in Linguistic Theories*

2.2.1.1 *Formal Semantics*

In the 1970's, natural language semantics came under the influence of Montague's work (Montague, 1973) proposing *formal semantics*, which represents the first formal "implementable" approach to linguistic meaning. In this framework, the focus has been set on the logical properties of natural language. Formal semantics mainly concentrates on the syntax-semantics interface, defining rules, which allow us to translate surface structures into logical representations in a compositional way. This approach is also called *model-theoretic semantics*, because it accounts for linguistic meaning in terms of truth conditions (hypothetical states of affairs, which the sentence correctly describes) and denotation (objects in the world words refer to).

In formal semantics, a sentence meaning is given by its logical representation, sometimes called *logical form*[5], which is a formal representation of its logical structure derived from the corresponding surface form. For example, the sentences *Shakespeare wrote a tragedy* and *A tragedy was written by Shakespeare* will be assigned the same logical representation, which abstracts from the surface form of the sentences: $\exists t, s, e(tragedy(t) \wedge Shakespeare(s) \wedge write(e, s, t))$.[6]

This approach mainly concentrates on linguistic means of expressing logical features of a natural language expression, such as, for example, quantification, logical connectors, or modality. The meaning of the non-logical predicates (e.g., *Shakespeare*, *write*, *tragedy*) expressed by content words, as opposed to function words (e.g., *and*, *if*, *a*), is irrelevant in the context of formal semantics. It is defined in a referential way, i.e. the meaning of *tragedy* is given by the set of all entities, which can be referred to as "tragedy". Thus, the sentences *a cat eats a rat* and *a mat facilitates a nap*, which have the same syntactic structure, will be assigned the logical representations equal to $\exists x, y, e(P(x) \wedge Q(y) \wedge R(e, x, y))$.[7] Distinguishing between these sentences is then a matter of an interpretation function mapping P, Q and R to different sets.

Discourse semantics has extended the Montague's approach in order to go beyond sentence boundaries. In this framework, it is possible to represent the semantic structure of a sequence of sentences describing different eventualities. For example, Discourse Representation Theory (Kamp and Reyle, 1993) and Dynamic Predicate Logic (Groenendijk and Stokhof, 1991) consider intersentential anaphora and relations between the eventualities of

[5]The term *logical form* is rather controversial. In traditional linguistics, it is strongly associated with the framework of generative grammar (see, for example, Chomsky, 1976; May, 1977; Heim, 1982), which is not related to formal semantics and not concerned with inferences following from natural language sentences. In computational semantics and artificial intelligence, the term is understood in a wide sense; it refers to *any* formal representation of the text content given as a logical formula (see, for example, Hobbs *et al.*, 1993; Bos, 2008; McCord, 2010). In the following chapters of this book, we use the term *logical form* in the latter sense.

[6]This representation corresponds to a Davidsonian approach (Davidson, 1967). For simplicity, tense is disregarded in this and the following examples.

[7]This example is taken from Vieu (2009).

different sentences. This approach has been developed in Segmented Discourse Representation Theory (SDRT) (Asher, 1993; Lascarides and Asher, 1993; Asher and Lascarides, 2003). In order to link discourse segments, SDRT uses temporal, causal, mereological, or argumentative *discourse relations*.[8] The SDRT research naturally focuses on explicit discourse connectors, e.g., *but*, *because*. In addition, SDRT considers how world knowledge affects discourse structure. For example, knowledge like "when something is pushed, normally, it moves" enables us to establish the temporal and causal relations between such discourse segments as *John fell. Max pushed him*.[9]

Formal semantics has gained widespread popularity both among linguists and computer scientists, because it has opened new ways of computing natural language meaning. But it is definitely not the end of the story for computational semantics, because many natural language phenomena require more knowledge for their resolution than just logical structure.[10] For example, the logical representation does not help us to decide whether the word *tragedy* in the example above refers to a dramatic event or to a work of art. Being able to resolve this ambiguity presupposes looking deeper into the intern meaning of the content words. This was done by *lexical semantics* in its different versions.

2.2.1.2 *Lexical Semantics*

Lexical semantics developing in the framework of generative grammar and structuralism considered lexical meaning to be a starting point for a semantic theory (Katz and Fodor, 1963; Jackendoff, 1972). The main paradigms involved decomposing lexical meaning into *semantic markers* – atomic units of meaning and conceptualization. For example, Katz and Fodor (1963) proposed to capture different senses of the noun *bachelor* in terms of such semantic primitives as *human/animal*, *male*, *young*, *who has never been married*, *who has the first or lowest academic degree*, etc. This theory distinguishes between dictionary (definitional) and encyclopedic knowledge. While the former is considered to be a part of the lexical meaning, the latter is not. For example, the attribute *being a vehicle* is a part of the meaning of *car*, while *moving on a road* is not, because the latter attribute is prototypical rather than necessary. In practice, it turned out to be difficult to find "definitions" and a finite set of semantic markers for the largest part of the lexicon. Shortly after it appeared, the Katz-Fodor theory was subjected to diverse criticisms (see, for example, Bolinger, 1965; Vermazen, 1967; Putnam, 1975).

[8] The notion of discourse relations is borrowed from Rhetorical Structure Theory (Mann and Thompson, 1988).
[9] This example is taken from Asher and Lascarides (2003).
[10] See Sec. 2.3.2 for detailed examples.

Probably, the most successful application of the decomposition approach concerns verb meanings as decomposed into thematic roles. In this approach, the lexical meaning of a verb includes a specification of the types of arguments associated with this verb. For example, the verb *put* has an associated "putter" (agent role), the thing that is put (theme role), and the place where it is put (location role). This approach was taken, for example, by Fillmore (1968), Jackendoff (1987), and Dowty (1991). Nowadays, decomposition of verb meanings into thematic roles seems to be a standard solution for verb semantics. However, this approach still has a fundamental problem concerning fixing a universal inventory of roles and the ambiguity in assigning roles (see Riemer, 2010, for an overview).

Another aspect of verb meaning elaborated in the framework of generative grammar concerns *selectional preferences*, which are the semantic constraints that a word imposes on the syntactic environment. For example, the verb *to drink* usually takes a beverage as its object. This knowledge can help to disambiguate sentences like *Mary drank burgundy*, while *burgundy* can be interpreted as either a color or a beverage.[11] Chomsky (1965) has incorporated selectional preferences into his syntactic theory, whereas other researchers considered them to be predictable from lexical meanings (e.g., McCawley, 1973).

Instead of a definition-based model of lexical meaning, Rosch (1978) has proposed Prototype theory considering a category as consisting of different elements, which have unequal status. For example, a robin is a more prototypical member of the class of birds than a penguin and *being able to fly* is a more prototypical attribute of birds than *eating worms*. Prototype theory is based on Rosch's psycholinguistic research on internal structure of categories (Rosch, 1975). The conclusion followed from the experiments involving response times, priming, and exemplar naming was that some members/attributes of a category are more privileged than others. Prototype theory has quickly attracted a lot of attention, but also a lot of criticism. The critical issues concern problems of identifying attributes for classes, accounting for category boundaries, treating abstract non-visual categories, and compositionality issues (see Riemer, 2010, for an overview).

In spite of the problematic issues, prototype models of categorization have had a significant influence on the research in semantics. Many of the insights of prototype research are accounted for in cognitive approaches to semantics, which aim at developing a comprehensive theory of mental representation. The term "cognitive semantics" covers a variety of different approaches, which share several common points. Cognitive semantics proposes a holistic view on language. Cognitivists like Langacker (1987) and Lakoff (1987)

[11] This example is provided by Resnik (1997).

reject the modular approach promoted by Chomsky (1965) assuming that language is one of a number of independent modules or faculties within cognition. Consequently, most researchers in cognitive semantics reject the dictionary–encyclopedia distinction (see, for example, Jackendoff, 1983; Langacker, 1987). As a result, semantic knowledge like ”bachelors are unmarried males” is considered to be not distinct from encyclopedic knowledge. In cognitive semantics, lexical meaning has a conceptual nature; it does not necessarily concern a reference to an entity in the world. Instead, meaning corresponds to a concept held in the mind, which is related to other concepts such that without knowledge of all of them, one does not have complete knowledge of any one.

Different cognitive approaches to semantics propose different models of the structure of concepts underlying lexical meaning. Fillmore's ideas have developed into Frame semantics (Fillmore, 1968), which considers lexical meanings to be related to prototypical situations captured by *frames* – structures of related concepts. Lakoff (1987) has introduced *idealized cognitive models*, which are theories of particular domains reflected in language. Langacker (1987) has developed Cognitive Grammar modeling semantic aspects as *image schemes*. Talmy has published a number of influential works on linguistic imaging systems (e.g., Talmy, 1983, 2000). Being quite popular among researchers working in the overlap area of linguistics and psycholinguistics, cognitive semantics has been mainly criticized for informality and the speculative character of cognitivist theories not really grounded on psychological experiments (see Riemer, 2010, for an overview of the critics).

As an alternative to classical decomposition theory of meaning and cognitive semantics, a relational approach to lexical meaning has been developed. Instead of defining lexical meaning in terms of semantic primitives, the meaning is represented as a network of relationships between word senses called *lexical-semantic relations*.[12] For example, one sense of bachelor is related to *unmarried* and another sense is related to *academic degree*. This approach has been described in detail by Cruse (1986). In computational linguistics, it has been implemented in electronic network-like dictionaries, the most famous of which is currently WordNet (Miller *et al.*, 1990; Miller and Fellbaum, 1991; Fellbaum, 1998b). Similar to decomposition theories, semantic networks describe lexical meaning in a definitional way. For example, WordNet relates *airplane* to *vehicle*, but not to *sky*. In contrast to decomposition theories, in this approach words are defined by other words rather than by semantic primitives. In a relational framework, representation of such complex definitions as “who has the first or lowest academic degree” is impossible, because

[12] See Sec. 3.1 for more details on lexical-semantic relations.

each lexical-semantic relation is just a two-place predicate relating word senses. However, this representation simplicity makes this approach implementable and extremely useful in practical NLP.

A relatively recent theory of linguistic meaning called Generative Lexicon (GL) was proposed by Pustejovsky (1991, 1995). Pustejovsky criticized the standard view of the lexicon, on which each lexeme is associated with a fixed number of word senses. For example, the adjective *fast* implies three different senses in the phrases *a fast typist* (one who types quickly), *a fast car* (one, which can move quickly), and *a fast waltz* (one with a fast tempo).[13] Pustejovsky argues that just listing these senses does not help to account for creative language use. For example, the use of *fast* in *fast motorway* cannot be accounted on the basis of the senses mentioned above. In order to cope with this problem, Pustejovsky focuses on additional non-definitional aspects of lexical meaning. He introduces semantic structures, which he calls *qualia structures* of words. A qualia structure includes facts about the constituent parts of an entity (Constitutive role), its place in a larger domain (Formal role), its purpose and function (Telic role), and the factors involved in its origin (Agentive role). For example, the qualia structure of *school* includes *an educational institution* as its Telic aspect and *building* as its Formal aspect. This knowledge enables us to generate different senses of *school* in such sentences like *The school was painted white* and *John has learned it at school*. The Generative Lexicon theory represents an important step towards linking lexical meaning to world knowledge. However, the theory has weak points concerning the speculative character of the qualia roles and the difficulty of assigning these roles to concepts associated with a target lexeme, for experimental studies revealing weak points in GL; see Kilgarriff (2001); Cimiano and Wenderoth (2007) .

Formal and lexical semantics refer to quite orthogonal aspects of linguistic meaning. Formal semantics accounts for logical features of languages, pays particular attention to compositionality, and focuses mainly on functional words, while content words are represented as atomic predicate names having referential meaning. In contrast, lexical semantics mostly ignores logical aspects, does not propose any adequate theory of compositionality, and concentrates on the specification of the lexical meaning of content words. It seems to be natural that both approaches could perfectly supplement each other in an integrative approach enabling a fuller understanding of natural language meaning. However, up to the

[13] This example is provided by Pustejovsky (1995).

present time not so many researchers have been working in the both frameworks; formal and lexical semantics seem to a large part to ignore each other.[14]

2.2.1.3 *Distributional Semantics*

With the development of machine learning techniques, distributional approaches to lexical meaning have become extremely popular in computational linguistics and practical NLP. These approaches are based on the idea captured in the famous quotation from Firth (1957): "You shall know a word by the company it keeps". This idea is often referred to as the *distributional hypothesis*, because it presupposes deriving lexical meaning from the distributional properties of words: "words which are similar in meaning occur in similar contexts" (Rubenstein and Goodenough, 1965).

The distributional hypothesis is often considered to originate in the works of Harris (1954, 1968). In this approach, linguistic meaning is inherently differential, and not referential; differences of meaning correlate with differences of distribution. Distributional semantics defines lexical meaning of a word w as a vector of values of similarity between w and other words in the corpus. There are different approaches to calculating co-occurrence-based semantic similarity between two words w_1 and w_2.

One approach is based on pointwise mutual information (PMI) defined as:

$$log_2 \frac{freq(w_1, w_2)}{freq(w_1) \cdot freq(w_2)},$$

where $freq(w_1, w_2)$ is the frequency of co-occurrence of w_1 and w_2 and $freq(w_i)$ is the frequency of occurrence of w_i. Pointwise mutual information was introduced as a lexical association norm by Church and Hanks (1989). The authors showed that word pairs with a high PMI are often semantically or associatively related.

Another approach to semantic similarity is based on the *vector space models*. In these models, a word w is represented by a vector of word co-occurrence frequencies. Each vector dimension k shows how many times w co-occurs with another word w_k in the same context. The context can be defined as a sequence of n words, entire document, a fixed pattern (e.g., *X is a part of Y*), or a syntactic structure. The similarity of two words is captured as the distance of the corresponding vectors, which can be calculated by one of the usual vector distance measures (e.g., Euclidean distance, cosine).[15] The comparison of co-occurrence vectors is also referred to as *second order co-occurrence* (cf. Grefenstette,

[14] But see integrative approaches developed by Pustejovsky (1995), Partee and Borschev (1998), Asher and Lascarides (2003).

[15] An overview on different distance measures is given by Salton and McGill (1986).

1994). In this approach, similarity is established because two words occur with similar words rather than with each other. Thus, two words can prove to be semantically related even if they never co-occur.

Different models for computing vector-based semantic similarity have been developed. The most prominent approaches include Hyperspace Analogue to Language (Lund and Burgess, 1996), Latent Semantic Analysis (Landauer, 2007), Topic-based vector space model (Kuropka and Becker, 2003), Generalized vector space model (Tsatsaronis and Panagiotopoulou, 2009).

Naturally, distributional semantics does not make any distinction between lexical and world knowledge. All possible associative links, which can be mined out of corpora are considered to be a part of lexical meaning. In this approach, the word *airplane* can be related to *vehicle*, *fly*, *pilot*, *plane crash*, *Aerobus*, and *Wright brothers*.

Distributional semantics provides an account of compositionality by assessing the acceptability of verb-noun, adjective-noun, noun-noun combinations. For example, the higher similarity between *boil* and *potato* as compared to the pair *boil* and *idea* can be used to predict that the combination *boil a potato* is more acceptable than the combination *boil an idea*. Based on this idea, semantic similarity has been used for modeling selectional preferences (Resnik, 1997; Erk *et al.*, 2010; Schulte im Walde, 2010) and learning qualia structures as defined in the Generative Lexicon theory (Lapata and Lascarides, 2003b).

In psycholinguistics, the distributional hypothesis has been used to explain various aspects of human language processing, such as lexical priming (Lund *et al.*, 1995), synonym selection (Landauer and Dumais, 1997), and semantic similarity judgments (Mcdonald and Ramscar, 2001). It has also been employed for a wide range of NLP tasks, such as disambiguation, information retrieval, anaphora resolution, identification of translation equivalents, word prediction and many others.[16]

2.2.2 *Linguistic Meaning in Artificial Intelligence*

The earliest natural language understanding programs, e.g., *STUDENT* and *ELIZA* described in Sec. 2.1, were able to process specific predefined domains only. Input sentences were restricted to simple declarative forms and were scanned by the programs for predefined key words and patterns.

Some of the systems developed during the mid-1960s (see, for example, Raphael, 1964; Craig *et al.*, 1966; Collins and Quillian, 1969), were able to store text representations and

[16]See Manning and Schtze (1999) for an overview.

draw simple inferences. Given the sentences *Sokrates is a man* and *All men are mortal*, they could answer queries like “Is Sokrates mortal?” These systems were using formal representations for storing information in a database and employed simple semantic processing for translating input sentences into this representation. Some systems, for example, could use simple logical relations like “if *A* is a part of *B* and *B* is a part of *C*, then *A* is a part of *C*”, but this relationship had to be stored in the program, so that they could only handle relationships they were designed for.[17]

Most of the natural language understanding programs developed in the 70s and later might be called *knowledge-based* systems; their development is closely related to the AI research on knowledge representation. These programs use world knowledge about the domain of interest, which is required for text interpretation. Knowledge-based systems can be roughly classified according to representation schemes and reasoning mechanisms which they employ to access world knowledge.

2.2.2.1 *Procedural Semantics*

In the framework of procedural semantics, knowledge is represented as an executable program in a computer language. Both meaning of sentences and knowledge about the world are represented in this way. The execution of these programs corresponds to reasoning from the meanings and knowledge. Thus, in procedural semantics, meaning is embodied in abstract procedures for determining referents, verifying facts, computing values, and carrying out actions. These procedures are built on computational operators and can include sensing and acting in the world. In the early 70s, two systems were developed, which were employing procedural semantics. Winograd’s *SHRDLU* was verbally controlled by a user and simulated a robot manipulating simple objects (Winograd, 1972). Woods’s *LUNAR* system answered queries about the samples of rock brought back from the moon (Woods *et al.*, 1972). For example, *LUNAR* represented the query “What is the average concentration of Aluminum in each breccia?” as a little program:

```
(FOR EVERY X5 / (SEQ TYPECS) : T ;
  (PRINTOUT (AVGCOMP X5
    (QUOTE OVERALL) (QUOTE AL203))))
```

[17] A good overview of the early NLU systems in given by Winograd (1972).

2.2.2.2 *Semantic Networks*

Quillian (1968) proposed a knowledge representation framework named *semantic networks*, which quickly became popular and has been employed by a variety of knowledge-based NLU systems (see Sowa, 1987, for an overview). Semantic networks, being a model for human associative memory, represented word and sentence meanings as a set of nodes linked in a graph. Networks were used to represent both meaning of text fragments and world knowledge. Quillian has developed simple operations on semantic networks that corresponded to drawing inferences. Compared to formal logical deduction, this sort of reasoning appeared to be more simple and efficient.

Inspired by semantic networks and the dependency theory of Hays (1964), Schank (1972) developed *conceptual dependency* theory. Figure 2.3 represents a conceptual dependency graph for the sentence *John gave Mary a book*. Schank used different kinds of arrows for different relations, such as ⇔ for the agent-verb relation. The distinction between the semantic network theory and the conceptual dependency theory lies in their focus. Semantic networks are about how knowledge should be organized and how to interpret a semantic net structure. This approach says nothing about what should be represented. In contrast, conceptual dependency theory aims at enumerating the types of nodes and arcs used to build meaning representations. This theory specifies content rather than structure. The conceptual dependency representation was used by Schank and his colleagues in several NLU systems (Schank, 1975; Schank and Abelson, 1977).

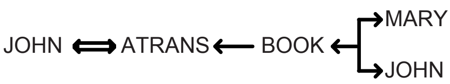

Fig. 2.3 Meaning of *John gave Mary a book* as a conceptual dependency graph.

2.2.2.3 *Frames*

Minsky (1975) proposed a knowledge representation format based on *frames*:

> When one encounters a new situation (or makes a substantial change in one's view of the present problem) one selects from memory a structure called a Frame. This is a remembered framework to be adapted to fit reality by changing details as necessary.
>
> A frame is a data-structure for representing a stereotyped situation, like being in a certain kind of living room, or going to a child's birthday party. Attached to each frame are several kinds of information. Some of this information is about how to use the frame. Some is

about what one can expect to happen next. Some is about what to do if these expectations are not confirmed.

For example, the concept of *Pacific island* can be represented by the following frame: `[is-a: island, located: Pacific_ocean, belongs_to: country, name: island_name]`, where `is-a`, `located`, `belongs_to`, and `name` are predefined slots for characterizing islands. In this framework, linguistic meaning is given by mapping of linguistic constituents into corresponding frame slots. Reasoning over frames is based on a unification procedure, which Minsky (1975) defines as follows: "given two frames *A* and *B*, [...] *A* can be viewed as a kind of *B* given a "mapping" or frame-transformation *C* that expresses (perhaps in terms of other mappings) how *A*'s terminals can be viewed in terms of *B*'s terminals".[18]

Building upon this framework, Schank and Abelson (1977) introduced the concepts of scripts, plans, and themes to handle story-level understanding. The classical example of Schank's theory is the restaurant script, which has the following characteristics:

Scene 1: Entering
S PTRANS S into restaurant, S ATTEND eyes to tables, S MBUILD where to sit, S PTRANS S to table, S MOVE S to sitting position

Scene 2: Ordering
S PTRANS menu to S (menu already on table), S MBUILD choice of food, S MTRANS signal to waiter, waiter PTRANS to table, S MTRANS 'I want food' to waiter, waiter PTRANS to cook

Scene 3: Eating
Cook ATRANS food to waiter, waiter PTRANS food to S, S INGEST food

Scene 4: Exiting
waiter MOVE write check, waiter PTRANS to S, waiter ATRANS check to S, S ATRANS money to waiter, S PTRANS out of restaurant

A variety of computer programs have been developed to implement the theory. Schank (1991) applied his theoretical framework to story telling and the development of intelligent tutors. Schank and Cleary (1995) described an application of these ideas to educational software.

In the late 70s and 80s, frame-based knowledge representation was one of the most active area of AI research in natural language understanding; see Barr (1980) for an overview of the early systems. The Ontological Semantics framework (Nirenburg and Raskin, 2004) is an example of a more recent large long-term project employing frame-like structures for

[18]For an elaborated approach to unification, see Shieber (1986).

knowledge representation. This approach represents an attempt on combining linguistic analyses (syntax, semantics-pragmatics pipeline) with background knowledge.

2.2.2.4 *Logical Formulas*

The idea of representation of linguistic meaning by logical formulas and using automated deduction for natural language understanding originated in the context of automated question answering. Green and Raphael (1968) developed a system that offered the full expressiveness of the first-order predicate calculus for the representation of natural language meaning. The deductive procedure of this system was based on an automatic theorem-proving algorithm by Robinson (1965). In this framework, both linguistic meaning and the knowledge base were represented as a set of logical axioms.

At present, there exist two main development directions for the approaches to NLU, which employ theorem proving. The first one follows the initial ideas of using automated deduction, see Sec. 4.2 for more details. In the second line of research, abduction rather than deduction is employed as the principle reasoning mechanism, see Sec. 4.3.

Some of the recent deduction-based approaches are using the full expressiveness of first order logic (see, for example, Dahlgren *et al.*, 1989; Bos and Markert, 2006). Others employ a decidable fragment of FOL, such as Description Logics, see Franconi (2003) for an overview of the applications of Description Logics to natural language processing.

2.3 Shared Word Knowledge for Natural Language Understanding

In the previous section of this chapter, we considered how representation of the linguistic meaning has been approached in linguistics and artificial intelligence. We have seen that many researchers working on computational NLU have come to the conclusion that world knowledge associated with content words plays an important role in understanding natural language. A natural question arises about what the term “world knowledge” actually means in the context of NLU and whether/how world knowledge differs from linguistic knowledge. Section 2.3.1 concerns this issue. Section 2.3.2 gives an overview of the typical natural language phenomena requiring world knowledge for their resolution, which represent the main challenges for knowledge-intensive NLP.

One important remark should be made at this point. Obviously, any kind of non-linguistic knowledge may be useful for the goals of natural language understanding: knowledge about a specific domain, which is considered in text, about laws of nature (e.g., physical laws), about the state of the art in the world, about the specifics of the text producer and his current state and so on. In this book, we focus on knowledge that is **shared** by humans belonging to the same linguistic and cultural community and do not consider situational and individual aspects of discourse.

2.3.1 *Linguistic vs. World Knowledge*

In order to discuss possible differences between knowledge of language and knowledge about the world, we illustrate different levels of knowledge relevant for natural language understanding with the following examples.[19]

(2.1) If *NP* is a noun phrase and *V* is an intransitive verb, then the concatenation *NP V* is a clause.
(2.2) The phrase *x wrote y* corresponds to the proposition $write(x,y)$.
(2.3) The proposition $write(x,y)$ refers to a "text creation" event, such that *x* plays the role of author and *y* plays a role of text in this event.
(2.4) If something is a play then it is a dramatic composition.
(2.5) The main function of a playwright is writing plays.
(2.6) If x creates y at time t then y is an artifact and it has not existed before t.
(2.7) "Romeo and Juliet" was written by Shakespeare.

Example (2.1) represents a typical syntactic rule. In Ex. (2.2), a surface realization of the predicate *write* is mapped to a logical form. In Ex. (2.3), the predicate and its arguments are mapped to an event frame and corresponding roles. Example (2.4) describes the *type-of* relation. Example (2.5) refers to the common sense knowledge about playwrights. Features of artifacts are defined in Ex. (2.6). Example (2.7) contains a specific fact about the world.

Syntactic rules like (2.1) are included in the grammar and are language-dependent. Mappings from surface realizations to logical forms like (2.2) are often a part of the lexicon, also language-dependent. Rules like (2.3) and (2.4) can be included into a lexical-semantic dictionary like FrameNet and WordNet.[20] Knowledge like (2.5) can occur in a definition provided by a thesaurus (e.g., WordNet), in a common sense ontology (e.g., OpenCyc), or in a lexicon (e.g., a lexicon based on the Generative Lexicon theory). Statements like (2.6) can be included into an abstract ontological theory of artifacts or into event calculus

[19] These examples are for a large part inspired by similar examples provided by Hobbs (2009). The reason for us constructing new examples instead of directly citing Hobbs is that we intend to link them to the theories and resources described elsewhere in this book.
[20] All the mentioned resources are discussed in Chap. 3

semantics. Facts like (2.7) can be a part of a factual ontology containing knowledge about instances rather than classes (e.g., YAGO).

It is quite straightforward that rules like (2.1) and (2.2) are language-dependent and belong to linguistic knowledge, while (2.7) is not related to the linguistic competence, it is a part of knowledge about the world. Everything between (2.2) and (2.7) is more difficult to classify. Statements (2.3)-(2.6) concern lexical knowledge, i.e. knowledge about word meanings, which is both language-dependent and anchored to the world.

Different semantic theories consider different types of knowledge to be part of lexical meaning. For example, according to the lexical-semantic theories as presented by Cruse (1986) and Miller and Fellbaum (1991), the lexical meaning of *play* comprises (2.4). Frame semantics (Fillmore, 1968) concerns knowledge like (2.3) to be a part of the lexical meaning of *write*. In the framework of Generative Lexicon (Pustejovsky, 1991), knowledge like (2.5) is an integral part of the lexical meaning of *playwright*.

Drawing a line between lexical semantics and world knowledge is a difficult issue. However, some researchers believe that this distinction is important. As the reader has seen in the previous section, this view has been especially promoted by the researchers working in the framework of traditional structuralism and generative grammar. Recently, it is supported by computational linguists working on formal grammars. For example, Copestake (1992) claims that it would help to isolate linguistic theory from non-linguistic phenomena:

> [I]t is methodologically important to distinguish between linguistic and non-linguistic representation, even though the two have to be interrelated so that linguistic utterances can be interpreted as having some connection with the real world. We want to avoid the situation where linguistic representation is dependent on the scientific knowledge about the world [...] we wish to provide a testable constrained theory, and a formal representation language, and to avoid problems, which arise in knowledge representation which do not have a linguistic dimension.

In Copestake's view, in order to construct a lexicon one should "start from the null hypothesis that all that the lexicon contains [...] are pointers connecting the phonological or orthographic representation of the word with its real world denotation. We then have to establish criteria for providing further information about word meaning, which will ensure that the additions have linguistic motivation" (Copestake, 1992). It remains unclear, however, whether establishing such criteria is possible at all.

Other researchers do not consider the borderline between lexical and world knowledge to be crucial for an adequate theory of linguistic meaning. For example, researchers working in the framework of cognitive semantics reject the dictionary-encyclopedia distinction.

More recently, Hobbs (2009) claims that "lexical knowledge is just ordinary knowledge where the entities in question are words". In order to support this view, the author reviews some of the psycholinguistic studies (e.g., Hagoort *et al.*, 2004; Tanenhaus and Brown-Schmidt, 2008) suggesting that semantic interpretation cannot be separated from non-linguistic knowledge. Hobbs argues that "[t]he most common argument in linguistics and related fields for drawing a strict boundary between lexicon and world is a kind of despair that a scientific study of world knowledge is possible". As opposed to this despair, Hobbs suggests a scientific account of world knowledge and a framework, in which all levels of semantic interpretation can be equally implemented as inferences.

In line with the cognitive approach to semantics, in this book, we do not distinguish between lexical and world knowledge. Thus, we are concerned with world knowledge as exemplified by the statements like (2.3)-(2.7), i.e. everything which goes beyond syntax and mapping of surface predicate-argument constructions to predications. Although the integrative knowledge base proposed in this book stores different types of knowledge in separate modules (cf. Chap. 5), this happens due to various technical reasons and not because we believe that there is a cognitive motivation for such modularity.

2.3.2 *Natural Language Phenomena Requiring Word Knowledge to be Resolved*

Ambiguity

The potential ability of linguistic signs to have more than one meaning is one of the major problems in NLP. Ambiguity affects all linguistic levels: phonological, morphological, lexical, syntactic, and semantic. Text fragments lifted out of context can be highly ambiguous, whereas within a discourse ambiguity can be mostly successfully resolved with the help of context and background knowledge. Example (2.8) below is a classical example of the syntactic ambiguity: Did John use a telescope to see the man or was the man carrying a telescope? Interestingly, if "man" is replaced by "picture", as shown in (2.9), the ambiguity disappears and only one reading remains possible. Our world knowledge implies that seeing an object using a telescope is quite normal, while what a picture can do with a telescope is unclear.

An example of lexical ambiguity is given in (2.10): *bank* can refer either to a financial institution or to a wall of a river channel. However, the following context makes us prefer the first reading, because we know that it is hardly possible to open accounts just sitting on a bank, but it is possible to do it in a financial institution.

(2.8) *John saw the man with a telescope*.

(2.9) *John saw the picture with a telescope.*

(2.10) *John went to the bank to open an account.*

Bridging

Bridging[21], or connecting parts of a discourse, implies a wide range of natural language inferences. One of the most studied bridging phenomena is anaphora. It is well known that syntactic and semantic agreement and parallelism are very helpful for anaphora resolution (see Mitkov, 1999). For example, given the sentences (2.11), in order to bind the anaphoric expression *John* to its antecedent *The boy* it is enough to find out that both expressions agree in number and belong to the same semantic typ, e.g., *human*. In order to resolve anaphora in (2.12) more information has to be involved: we need to know that the predicate *to be hungry* normally prefers a living being as its argument. Relating *house* and *door* in (2.13) presupposes knowledge about doors being parts of houses.

(2.11) *John reads a book. The boy likes reading.*

(2.12) *We gave the bananas to the monkeys because they were hungry.*[22]

(2.13) *John approached the house. The door was open.*

Discourse Relations

A discourse relation describes how two segments of discourse are logically connected to one another.[23] The sentences in Ex. (2.14) and (2.15) discussed by Lascarides and Asher (1993) have the same syntactic structure, but the corresponding events stand in different temporal relations. This follows from the background knowledge relating falling and pushing in a causative way, which is not the case for standing up and greeting. Similarly, reading (2.16) we understand that the alarm breaking event happened before waking up late and was the reason for it. We infer it using our world knowledge about typical waking up scenarios involving alarm bells.

(2.14) *Max stood up. John greeted him.*

(2.15) *John fell. Max pushed him.*

(2.16) *John woke up late today. His alarm broke.*

[21] See (Clark, 1975) for a classification of bridging types. A more recent overview of approaches to bridging is given in Asher and Lascarides (1998).

[22] This is a Wikipedia example.

[23] Discourse relations have been studied in (Hobbs, 1985a; Grosz and Sidner, 1986; Mann and Thompson, 1988; Lascarides and Asher, 1993) among others.

Implicit Predicates

When a predicate is highly predictable from the context, it can be omitted from the discourse.[24] Example (2.17) cannot be interpreted uniquely. But even if we do not know anything about John we can guess that he is either an author, an owner, an editor, or, for example, a seller of the mentioned book. This inference is possible because of our knowledge about the situations, in which a book (being an information container, a physical object, etc.) can be involved. Example (2.18) also lacks an explicit predicate. Knowing that the main feature of a wine is its taste, we can interpret (2.18) as *This wine is tasty*.

Noun compounds (2.19), possessives (2.20), and prepositional phrases (2.21) can be also interpreted in terms of implicit predicates. A morning coffee is most probably a coffee drunk in the morning, while a morning newspaper is a paper read in the morning. A Shakespeare's tragedy is a tragedy written by Shakespeare, while Shakespeare's house is a house where Shakespeare lives. *John in the house* describes the location of John, while *John in anger* denotes a state of John.

(2.17) *John finished the book.*

(2.18) *This wine is very good.*

(2.19) *morning coffee* vs. *morning newspaper*

(2.20) *Shakespeare's tragedy* vs. *Shakespeare's house*

(2.21) *John in the house* vs. *John in anger*

Metaphor and Metonymy

Rhetorical figures such as metaphor and metonymy are discourse phenomena requiring extremely strong reasoning capacities for their resolution.[25] Metaphor, or direct comparison of seemingly unrelated domains, requires deep knowledge about these domains, which allows us to find commonalities between them. Consider the famous citation from Shakespeare: “All the world's a stage”. Later in the play the author explains which aspects of the concepts *world* and *stage* have to be compared (“And all the men and women merely players; They have their exits and their entrances”). Without this hint it would be difficult to come to a unique interpretation. Metaphors often do not presuppose a single reading leaving the reader with a spectrum of different associations (c.f. “Juliet is the sun” from

[24] See Pustejovsky (1991) for a detailed study of implicit predicates.

[25] See Fass (1997) for a detailed description of metaphor and metonymy.

Shakespeare). Because of the complexity and relatively low frequency of occurrence in texts, metaphors tend to be a rather marginal topic in NLP research.[26]

Metonymy, or using a word for a concept, which is associated with the concept originally denoted by the word, is usually easier to handle. In order to understand a metonymic expression one needs to know the associative link. For example, in the sentence (2.22) *White House* denotes not the building but the cabinet of the US president sitting in the building. Metonymy is closely connected to the notion of *regular polysemy*[27] which refers to a set of word senses that are related in systematic and predictable ways. Example (2.23) illustrates this phenomenon. In order to understand the proper relation between the referents of the word *school* in both sentences, we have to know that it can refer to different aspects of the concept *school*: building, institution, group of people, etc. This is a regular conceptual shift; the same set of senses is predictable for words denoting different types of organizations, e.g., firms, universities, ministries.

(2.22) *The White House supports the bill.*

(2.23) *John hurried up to the school. The school was going for an outing that day.*

2.4 Concluding Remarks

Computational natural language understanding implies automatically creating a formal representation of the text content. The more relevant information an NLU system manages to capture in this representation, the better it "understands" the text. The form and the content of the representation depend on the underlying theory of the linguistic meaning.

In linguistics, there are three main approaches to meaning: formal semantics, lexical semantics, and distributional semantics. These three frameworks, being for the most part orthogonal, consider different aspects of natural language semantics. Formal semantics is focused on the logical and compositional properties of language. Lexical semantics accounts for the organization of the lexical systems and semantic links between word senses. Distributional semantics regards properties of words as used in contexts. As for computational applications, formal semantic approaches mostly result in semantic parsers, while research in lexical and distributional semantics leads to construction of lexical-semantic databases (see Chap. 3).

[26] But see, for example, (Alonge and Castelli, 2003).
[27] This term was first introduced by Apresjan (1973).

In the study described in this book, we benefit from all the mentioned research directions. We use semantic parsers for producing logical representations of text fragments and enrich these representations on the basis of knowledge extracted from lexical-semantic databases. In addition, we use distributional information in order to recover those semantic relationships, which cannot be inferred with the help of lexical-semantic knowledge.

AI research on natural language understanding has been mostly focused on knowledge representation and reasoning techniques paying less attention to linguistic meaning. However, some of the recent AI approaches to NLU successfully employ sophisticated semantic parsers and lexical-semantic databases developed by computational linguists, see Chap. 4 for more details. Following this research direction, we use logical axioms for formalizing the developed integrative knowledge base and employ automated reasoning for drawing inferences relevant for natural language understanding.

This chapter briefly discusses the differences between linguistic knowledge and knowledge about the world. In this book, we do not make a distinction between lexical and world knowledge. We hope to have shown that a borderline between these two is difficult to draw, while non-linguistic knowledge about the world is crucial for interpretation of linguistic expressions.

http://www.springer.com/978-94-91216-52-7

Integration of World Knowledge for Natural Language Understanding
Ovchinnikova, E.
2012, XVII, 242 p. 16 illus., 2 illus. in color., Hardcover
ISBN: 978-94-91216-52-7
A product of Atlantis Press