1

Formal Study of Natural Language

Summary

This chapter introduces computational semantics as the art and science of computing meanings for the expressions of a language. The chapter starts with a general overview of the formal study of language. The broad areas of syntax, semantics and pragmatics are distinguished, and the concept of meaning is discussed. Since the book will focus on language as a tool for describing states of affairs and for conveying information, and since logicians have designed special purpose tools for just these tasks, logic will be important to us. The chapter emphasizes the similarities between natural languages and the formal languages that have been designed by logicians and computer scientists. The chapter ends with a discussion of the usefulness of (functional) programming for computational semantics, and with an overview of the rest of the book.

1.1 The Study of Natural Language

Language is one of the most remarkable capacities of human beings and one of the distinctive features that set us apart from other inhabitants of this planet. Human languages are sophisticated systems for manipulating information-encoding symbols, and for composing sounds into structured expressions such as words, phrases, and sentences. These expressions can then serve in numerous ways for communicative actions like information exchange, persuading and deceiving, expressing thoughts, reasoning, and much more.

Linguistics is the scientific study of human language. To make linguistics into a science it is necessary to specify the object of study. In his book *Syntactic Structures* (1957) the linguist Noam Chomsky made the influential proposal to identify a human language with the set of all correct (or: grammatical) sentences of that language. This idealization abstracts from cognitive limitations of language users. Chomsky called the ability of language users to recognize the members of this set,

at least in principle, *competence*, and he distinguished this from *performance*, the actual abilities of language users, affected by conditions like memory limitations, distraction, and errors. In this book we will make the same distinction. We will focus on studying competence and therefore assume that a particular language is given to us as a set of sentences.

What exactly makes us competent speakers of our language? What is linguistic knowledge? It is certainly something we have only limited access to. For most speakers of a language the rules and regularities of their native tongue are implicit. They know how to apply them correctly, but when asked to state them, they usually are at a loss. Explicit descriptions of the rules and regularities of a language are called *grammars*. Grammars can be viewed as models of our language competence.

How grammars are represented as a cognitive device in the brain of the language user, we have no way of knowing. As a matter of fact, grammars can be represented in very different ways. In everyday life we encounter them, for example in dictionaries, in online language references, and in textbooks for learning foreign languages. Their design differs with their purpose. Grammars for language learners usually follow pedagogical guidelines and may oversimplify rules. Grammars used in linguistics usually aim to be as explicit and complete as possible. They may start with specifying basic elements such as sounds, words, and meanings. They then state rules for how to combine these into more complex structured expressions. Usually there are various representation levels, in accordance with the subdisciplines of grammar theory:

- *Phonology* explores what the smallest meaning-distinguishing units (sounds) are and how they are combined into the smallest meaning-carrying units (morphemes).
- *Morphology* is concerned with how morphemes are combined into words.
- *Syntax* studies how words are combined into phrases and sentences.
- *Semantics* investigates meanings of basic expressions and how meaning is assigned to complex expressions based on the meaning of simpler expressions and syntactic structure.

In this book, we will not be concerned at all with phonology, and we will also ignore most aspects of morphology. We will concentrate on linguistic form at the level of phrases and sentences. That means we will start with words as basic building blocks. The main task then will be to develop a grammar that provides us with notions of syntactic well-formedness and syntactic structure, and allows us to develop the notion of meaning for well-formed structures. Thus, our grammars

- should be capable of building exactly those expressions that are well-formed in the language of our choice,

- should determine the constituents of complex linguistics expressions, as well as their internal structure,
- and should allow us to assign appropriate meanings to syntactically well-formed expressions, on the basis of their structure.

In other words, the notion of syntactic well-formedness allows us to determine whether a particular expression is indeed a well-formed expression of a given category, and to determine its internal structure. The structure in turn will help us to relate the linguistic forms to the extralinguistic world (what they are about).

1.2 Syntax, Semantics, and Pragmatics

A basic trichotomy in the study of language is that between syntax, semantics, and pragmatics. Roughly speaking, syntax concerns the form aspect of language, semantics its meaning aspect, and pragmatics its use. Developing insight about the semantics of a language presupposes knowledge of the syntax of that language; study of language use presupposes knowledge about both syntax and semantics. It is not surprising, then, that for lots of languages we know more about their syntax and semantics than about pragmatic aspects of their use. Here are approximate definitions of the three notions:

Syntax is the study of strings and the structure imposed on them by grammars generating them.

Semantics is the study of the relation between strings and their meanings, i.e. their relation with the extralinguistic structure they are about.

Pragmatics is the study of the use of meaningful strings to communicate about extralinguistic structure in an interaction process between users of the language.

In a slogan: syntax studies Form, semantics studies Form + Content, and pragmatics studies Form + Content + Use. In this chapter we will examine what can be said about this trichotomy in concrete cases.

It is a matter of stipulation what is to be taken as 'form' in natural language. In the following we will concentrate on written language, more specifically on well-formed sentences. Thus, we will look at languages as sets of strings of symbols taken from some alphabet.

This choice makes it particularly easy to bridge the gap between formal and natural languages, for formal languages can also be viewed as sets of strings. The difference between natural and formal languages lies in the manner in which the sets of strings are given. In the case of formal languages, the language is given by stipulative definition: a string belongs to the language if it is produced by a

grammar for that language, or recognized by a parsing algorithm for that language. A parsing algorithm for C, say, can be wrong in the sense that it does not comply with the International Standard for the C programming language, but ultimately there is no right or wrong here: the way in which formal languages are given is a matter of definition.

In the case of natural languages, the situation is quite different. A proposed grammar formalism for a natural language (or for a fragment of a natural language) can be wrong in the sense that it does not agree with the intuitions of the native speakers of the language. Whether a given string is a well-formed sentence of some language is a matter to be decided on the basis of the linguistic intuitions of the native speakers of that language.

Once a grammar for a language is given, we can associate structures with sentences. For natural languages, native speakers have intuitions about which constituents of a sentence belong together, and a natural language grammar will have to comply with these judgements. This might constitute an argument for considering constituent trees rather than strings as the forms that are given.

Still other choices are possible, depending on one's aim. If one considers the problem of natural language recognition, one might consider strings of phonemes (smallest distinctive units of sound of a language) as the basic forms. If one wants to allow for uncertainty in language recognition, the basic forms become lists of words with the possibility of choice, so-called word lattices (with *full proof* and *fool proof* listed on a par as alternatives), or phoneme lattices (strings of phonemes involving the possibility of choice). Maybe we can attach probabilities to the choices. And so on. And so on.

An alternative road to meaning is pointing, or direct demonstration, arguably the starting point of basic concept learning. One's views of the principles of concept learning tend to be coloured by philosophical bias, so let us not get carried away by speculation. Instead, here is an account of first-hand experience. This is how Helen Keller, born deaf-mute and blind, learnt the meaning of the word *water* from her teacher, Miss Sullivan:

> We walked down the path to the well-house, attracted by the fragrance of the honeysuckle with which it was covered. Some one was drawing water and my teacher placed my hand under the spout. As the cool stream gushed over one hand she spelled into the other the word *water*, first slowly, then rapidly. I stood still, my whole attention fixed upon the motions of her fingers. Suddenly I felt a misty consciousness as of something forgotten – a thrill of returning thought; and somehow the mystery of language was revealed to me. I knew then that "w-a-t-e-r" meant the wonderful cool something that was flowing over my hand. That living word awakened my soul, gave it light, hope, joy, set it free!
>
> [Kel02, p. 34]

Explanations of meaning fall in two broad classes: meaning as *knowing how* and meaning as *knowing that*. If we say that Johnny doesn't know the meaning of a good spanking, we refer to operational meaning. If Helen Keller writes that *water* means the wonderful cool something that was flowing over her hand, then she refers to meaning as reference, or denotational meaning.

Usually the two are intertwined. Suppose we ask you directions to the Opera House. You might say something like: 'Turn right at the next traffic light and you will see it in front of you.' Understanding the meaning of this involves being able to follow the directions (being able to work out which traffic light counts as the 'next' one, which direction is 'right', and being able to recognize the building in front). Phrased in terms of 'being able to work out ...', 'being able to recognize ...', this is meaning as knowing how. Being able to understand the meaning of the directions also involves being able to distinguish correct directions from wrong directions. In other words, the directions classify situations, with me being positioned at some location in some town, facing in a particular direction: in some situations the directions provide a description which is true, in other situations a description which is false. This is denotational meaning.

Denotational meaning can be formalized as knowledge of the conditions for truth in situations. Operational meaning can be formalized as algorithms for performing (cognitive) actions. The operational semantics of *plus* in the expression *seven plus five* is the operation of adding two natural numbers; this operation can be given as a set of calculating instructions, or as a description of the workings of a calculating machine. The distinction between denotational and operational semantics is basic in computer science, and often a lot of work is involved in showing that the two kinds match.

Often operational meaning is more fine-grained than denotational meaning. For instance, the expressions *seven plus five* and *two times six* both refer to the natural number twelve, so their denotational meaning is the same. But they have different operational meaning, for the recipe for adding seven and five is different from the recipe for multiplying two and six.

1.3 The Purposes of Communication

The most lofty aim of communication is to use language as a tool for collective truth finding. But language is also a tool for making your fellow citizens believe things, or for confusing your enemies. Even if two language users agree on non-deception, this does not exclude the use of irony.

A very important use of language, and one that we will be concerned with a lot in the following pages, is as a tool for describing states of affairs and as a reasoning tool. This is a use which formal languages and natural language have in common.

Formal languages are often designed as query tools or reasoning tools, or at least as tools for formal reconstructions of reasoning processes. Therefore, they are a natural focus point for the formal study of language.

One can look at it like this. Suppose we want to use language to communicate basic facts, like 'the sun is shining', 'it is cold', 'it is raining', and so on. If you want to deny such a fact, you need to be able to say a thing like 'it is not cold'. You might also wish to express your uncertainty about which of two facts is the case, so you might like to say 'either it is cold or it is raining'. Similarly, you might like to say 'it is cold and it rains', or: 'if it rains, then it is cold.' So the ingredients of just about the simplest kind of communication are: basic facts, negations, conjunctions, disjunctions, and implications. A fragment of natural language that has only these is already quite useful. In fact, the usefulness of this simple fragment has been evident to logicians for a long time. The study of what can be expressed in this fragment is called *propositional logic* or *Boolean logic*, after the British mathematician George Boole (1815–1864).

To see how propositional logic can be used to express what goes in simple communicative discourses, suppose we want to talk about basic facts a, b. Suppose you know nothing about whether these facts are true or not. Then for you there are four possibilities: both facts are true, both facts are false, the first fact is true and the second one is false, or the first fact is false and the second one is true. Now we tell you 'a or b'. If you believe us, this will allow you to rule out one of the four possibilities, the possibility where both a and b are false. Or if you take our statement in the exclusive sense, it will even allow you to rule out two of the four possibilities. Your knowledge has grown by elimination of possibilities.

Exercise 1.1 We have seen that (complete) ignorance about the truth or falsity of two facts can be modelled as uncertainty between four possibilities. Now suppose there are ten basic facts. How many possibilities would that give? How about the general case of n basic facts?

If you just want to study simple factual information exchange, it makes sense to focus on the fragment of natural language that can be translated into propositional logic. But suppose you want to be more explicit about expressing relations between things. If you wish to declare your love to someone, say, then stating a basic proposition like 'there is love' is maybe not articulate enough. You might wish to say something more daring, like 'I love you.' Talk like this expresses relations between subjects and objects. You can use pronouns like 'I' and 'you', but also proper names. You can still do the propositional stuff, like disjunctions, negations, but you can also express quantificational facts. You can use the power of quantification to add to your romantic declaration, strengthening it to 'I love no-one but you'. You are now in the realm that is called *predicate logic*.

You can say interesting things with predicate logic, at least if you know how to use it, for predicate logic is very expressive. We will see below that predicate logic gets us a long way in expressing the meanings of natural language statements. Still more expressive is *typed higher-order logic*, which will also be used extensively in this book. In typed logic you can say more abstract (but still romantic) things like 'To love someone like you makes me very happy.' Finally, at the end of the book, we will have a look at the logic of knowledge, or *epistemic logic*. This will shed light on the meaning of statements like 'I am not completely sure whether I still love you.' It will also allow us to give an abstract picture of how communication by means of declarative statements leads to growth of knowledge, in subtle ways. We will study growth of audience knowledge about what the speaker knows, but also growth of speaker knowledge about audience knowledge about speaker knowledge, and so on.

Logic is a field that has made tremendous progress by focussing on well-defined formal languages, such as the example languages above, and studying their properties in depth. It is possible to view logical languages like the language of propositional logic or the language of predicate logic as fragments of natural language, by focussing at the specific set of sentences of natural language that can be translated into the logical language. This is going to be an important method in this book. This method of fragments, of taking things step by step, was first proposed for natural language analysis by logician and philosopher Richard Montague (1930–1971) in the 1970s, when he made the following famous statement:

> There is in my opinion no important theoretical difference between natural languages and the artificial languages of logicians; indeed, I consider it possible to comprehend the syntax and semantics of both kinds of languages within a single natural and mathematically precise theory.
>
> [Mon74b, p. 222]

Montague's pioneering work has shown how natural languages can be formally described using techniques borrowed from logic. It introduced tools to compute meanings in a systematic way and gave rise to a whole tradition of *formal semantics*, a very general term covering logical approaches to natural language semantics.

Our ultimate goal is to form an adequate model of parts of our language competence. Adequate means that the model has to be realistic in terms of complexity and learnability. We will not be so ambitious as to claim that our account mirrors real cognitive processes, but what we do claim is that our account imposes constraints on what the real cognitive processes can look like. In the rest of this chapter, we are going to explore the means that will help us fulfilling this goal.

1.4 Natural Languages and Formal Languages

English, Swedish, Russian, and Hindi are natural languages. The language of primary school arithmetic, the language of propositional logic, and the programming language Haskell are formal languages. But the distinction is not absolute, since we can find cases in between, like Esperanto. In order to see whether we can draw a dividing line, let us look at some crucial design features of human languages.

- *Duality of patterning* (or *double articulation*): The construction of linguistic content can be analyzed on two structural levels. One is the level containing the smallest meaningful units of language: morphemes or words, like *cat*. However, they are not minimal but on another level are made of a small set of speech sounds, the so-called phonemes. These do not carry a meaning themselves but only differentiate meaningful units, e.g. /k/ and /m/, that distinguish between *cat* and *mat*.
- *Recursion*: Structural patterns in sentences or phrases can repeat themselves. A common noun *bird* can be modified to a complex common noun *green bird*, this can be modified again, to form *small green bird*, and again, to form *beautiful small green bird*.
- *Contextuality*: What a phrase or sentence means is determined in part by the context in which the phrase is used.

Exercise 1.2 Pollard and Sag, in their textbook [PS94], give the following example of the use of recursion to extend sentences:

- Sentences can go on.
- Sentences can go on and on.
- Sentences can go on and on and on.
- Sentences can go on and on and on and on.
- ...

Can you give a concise description of this recursion pattern?

Exercise 1.3 Does it follow from the example in Exercise 1.2 that there are infinitely many English sentences? Or does it follow that English sentences can have infinite length? Or both?

The properties of duality of patterning, recursion, and contextuality set human languages apart from communicative systems of animals, such as bee dancing. They are responsible for the creative economy of language, because they allow us to build and understand infinitely many sentences using only finitely many sounds and rules – a precondition for enabling children to learn language as quickly and easily as they do.

Another key notion when talking about properties of human languages is *compositionality*. The so-called principle of compositionality will concern us a lot in what

follows. This principle is usually attributed to the German mathematician Gottlob Frege (1848–1925). What it says is that the meaning of a complex expression depends on the meanings of its parts and the way they are combined syntactically. This formulation is quite vague, however, because nothing is said yet about what meanings are, what counts as a part of an expression, and what kind of dependence we are talking about. In order to make sense, the principle has to be specified with respect to these matters. Moreover, it is meaningful only when understood against the background of further requirements on a semantic theory, for example the requirement that meaning assignment is systematic. Such a systematic account will capture the fact that once we know the meaning of *white unicorn* and *brown elk*, we also know what *brown unicorn* and *white elk* mean.

Compositionality does not occur in our list of the crucial properties of human languages listed above, because we assume that it is primarily a matter of methodology. The question is not whether natural languages satisfy the principle of compositionality, but rather whether we can and want to design meaning assembly in a way that this principle is respected. Setting up the representation of meaning in a compositional way has the merit of elegance, but it is not always straightforward. A recurring difficulty is the context-dependence of natural language meaning. To state the meaning of a pronoun, information is needed about the context in which the pronoun occurs. Something similar holds for presuppositions. In such cases, there are in general two kinds of reactions. The easy way out is to give up on compositionality. The other kind of reaction is to enrich and extend our semantic theory in ways that allow us to capture seemingly non-compositional phenomena in a compositional way. We will return to these matters in due time, when we take a closer look at pronoun resolution, in Chapter 12, and at presupposition, in Chapter 13.

Besides natural languages, there are other symbol-manipulating systems that also show some of the properties listed above, for example the language of predicate logic and high-level programming languages like C, LISP, and Haskell. But these apparently lack other properties of human languages. For example, they lack the whole pragmatic dimension that human languages employ: deception, irony, conveying information without explicitly stating it, and capabilities like creating and understanding metaphors. Formal languages also lack the flexibility of human languages induced by vagueness together with heavy use of context and background knowledge. These differences, in fact, are the reason why natural languages are well-suited for efficient inter-human communication, whereas formal languages excel for doing mathematics and for interacting with computers.

These important differences become visible if we focus on the way natural languages and formal languages are used. The differences disappear from view if we look at languages as sets of sentences. When we focus on fragments of natural

languages and describe their grammar in a formal way, we are, in fact, doing just the same as when describing formal languages.

1.5 What Formal Semantics is Not

A widespread prejudice against formal semantics for natural language is that it is just an exercise in typesetting. You explain the meaning of *and* in natural language by saying that the meaning of *Toto barked and Dorothy smiled* equals A **and** B, where A is the meaning of *Toto barked* and B is the meaning of *Dorothy smiled*. It looks like nothing is gained by explaining *and* as **and**.

The answer to this is that **and** refers to an operation one is assumed to have grasped already, namely the operation of taking a Boolean meet of two objects in a Boolean structure. Assuming that we know what a Boolean structure is, this is a real explanation, and not just a typesetting trick. On the other hand, if one is just learning about Boolean structures by being exposed to an account of the semantics of propositional logic, it may well seem that nothing happens in the semantic explanation of propositional conjunction.

A well known story about the linguist Barbara Partee has it that she once ended a course on Montague semantics with the invitation to her students to ask questions. As this was the end of the course, they could ask any question, however vaguely related to the subject matter. So a student asked: 'What is the meaning of life?' And Partee said: 'To answer that question we just have to see how Montague would treat the word *life*. He would translate it into his intensional logic as the constant *life'*, and he would use a cup operator to indicate that he was referring to the meaning, i.e. the extension in all possible worlds. So the answer to your question is: the meaning of *life* is ^*life'*. Any other questions?'

The core of the problem is that it is hard to avoid reference to meaning by means of symbols anyway. Compare the process of explaining the meaning of the word *bicycle* to someone who doesn't speak English. One way to explain the meaning of the word is by drawing a picture of a bicycle. If your pupil is familiar with bicycles, the meaning will get across once he or she grasps that the drawing is just another way to refer to the actual meaning. The drawing itself is just another symbol, for 🚲 is just another way to refer to bicycles. In the same way **and** is just another symbol for Boolean meet.

One of the things that make the study of language from a formal point of view so fascinating is that we can borrow insights from the formal sciences – mathematics, logic, and theoretical computer science – just like linguistics also borrows insights from psychology, philosophy, and so on. Formal insights and tools can be used for modelling language competence in a clear and precise framework, that finally allows us to implement natural language in a machine. Whether this is helpful