Part I
The Lexicon in Linguistic Theory

1 Introduction

1.1 Overview

In this introductory chapter we outline the basic concepts covered and terms used in this book. We first introduce the notion of *mental lexicon* and explain how it is different from the word listings familiar from most dictionaries. The next issue we address is the definition of *word* as a basic unit of language: we present the different criteria of wordhood, and lay out a preliminary approach to the notion of *lexical entry* and its structure (dealt with in more detail in Chapter 6). In order to explain how different kinds of lexical data should be treated and how they can be used to validate theoretical claims, we offer a general typology of lexical data coupled with the research methodology usually adopted in each case. Finally, we present a general vision of the concept of *natural language grammar* and establish the role of the lexicon as one of its core modules.

1.2 What Is the Lexicon?

Loosely speaking, the **lexicon** is a collection of words capturing the knowledge that speakers and hearers have about basic lexical expressions in a language. There is no question that our repertoire of words (what we will call our *mental lexicon*) is large, containing up to 250,000 lexical entries. This is based on fairly conservative estimates of speaker competence with *active* vocabulary (do I know how to use it?) and *passive* vocabulary (do I understand it?). For example, an average speaker's lexicon might contain at least 5,000 different verbs, 20,000 distinct nouns, and over 5,000 adjectives. Combine this with an additional 20,000 compound forms and at least 200,000 distinct proper names, and the lexicon grows fairly large indeed. This certainly indicates that the organization of our mental lexicon has the capacity for storing large amounts of information.

Furthermore, words are retrieved from the mental lexicon surprisingly fast: when reading, it takes us less than half a second to decide whether the sequence of letters we see is a real word (e.g., *parrot*) or a non-word (e.g., *varrot*), a problem known as a *lexical decision task* in psycholinguistics. This would be impossible if words were just randomly heaped up in our mind. Rather, this suggests that the lexicon is a highly organized and very complex system.

The notion of a lexicon as a reference list of words is familiar to anyone who has used a dictionary or studied a foreign language. However, there are a number of important differences between the mental lexicon and conventional dictionaries, designed by people for specific purposes. In the words of Hanks (2000), "checklist [numbered list] theories in their current form are at best superficial and at worst misleading. If word meanings do exist, they do not exist as a checklist." Hanks believes that the very structure of dictionary definitions has given people a false impression of how language is used, and certainly how it is organized. Most book dictionaries emulate a *sense-enumerative lexicon* strategy, which involves essentially lists of words, wherein each lexical entry represents a different word, and each lexical entry is structured as a list of senses. *Homonyms* (words having the same sound and spelling but different, unrelated meanings, cf. Section 6.5.1) are treated as different words (i.e., different lexical entries), and logically related meanings of the same acoustic and written form are treated as different senses of the same *polysemic* word within the same lexical entry. For example, the following entries from the Merriam-Webster Online Dictionary illustrate how both cases (polysemy and homonymy) are dealt with in conventional dictionaries:

(1) [1]**race** *noun*
1. the act of running
2. a strong or rapid current of water flowing through a narrow channel
...

[2]**race** *verb*
1. to compete in a race
2. to go, move, or function at top speed or out of control
...

[3]**race** *noun*
1. a breeding stock of animals
2. a family, tribe, people, or nation belonging to the same stock
...

As we can see, there are three lexical entries corresponding to *race*, two of which are nominal and one verbal. The 'act of running' sense and the 'current of water' sense are treated as related, they are included in the same lexical entry, and so are the different submeanings included in the other two entries. The 'act of running / water current' meaning cluster is treated as unrelated to the other nominal sense (referring to a group of common descent), or to the verbal meanings. Is this division justified? Most speakers would probably agree that the first and third lexical entries are not logically related. But what about the first and second entries? Can we state in good faith that the act of running and the running competition are totally unrelated? Hardly: both are obviously related, and the noun *race* is even used in the definition of the verb *race*.

The lexical entries and word senses in such listed representations are *isolated* and *compartmentalized*: there is no straightforward way to establish a connection even between related word senses in the same lexical entry, let alone different

lexical entries. To a large extent, this is due to the fact that the lexical meanings are presented as *atomic*: even though in each case there is a specification of the more general class to which the defined entity, property, or event belong ('act', 'current', 'stock' in (1)), there is no systematic indication of how these classes and, consequently, the defined terms are arranged and related among themselves.

A further indication of why this is not how the mental lexicon is organized is the order of lexical entries. Alphabetical order is convenient for human search and use. However, it does not seem to reflect the characteristics or properties of everyday language use, such as frequency and semantic and morphological relatedness. For example, the proximity of one word to another in the aphabetical order seems to have little to do with how they are related to each other: *abdicate* and *accolade*, or *table* and *taboo*, are relatively close in a dictionary, but this is only because they share the first letter or two, and not because they are related semantically or syntactically (cf. Section 1.6).

Another important property of the mental lexicon is that we naturally assign words to categories and larger conceptual classes. While such groupings (referred to as *hypernyms*, see Section 10.5) are typically captured in print dictionaries, humans can easily distinguish instances of such categories – that is, typical representatives of a class – from those which are not as typical (cf. Section 5.5). Thus, we know that an *apartment* or a *house* are more typical kinds of *housing* than, say, a *tent* or a *barrack*. This has proven to be an important distinction, giving rise to *typicality effects* in experimental psycholinguistic studies. However, the closeness of the defined term to the prototype is not accounted for in conventional dictionaries: both *apartment* and *barrack* are defined as *housing*.

Typicality Effects in Dictionaries

Of course, there are exceptions to this characterization even among conventional dictionaries, for example the *Diccionario del español usual en Mèxico* (Lara Ramos 1996). Anna Wierzbicka also advocates enriching the lexicographical entries with the information related to stereotypes and other socially and culturally relevant information in her natural semantic metalanguage proposal (Wierzbicka 1985a, see Section 5.6).

The list of entries and senses registered in the dictionaries (even very large ones) is finite and thus incomplete for obvious reasons: only the *actual* uses of words in a language will ever appear in a dictionary, usually after they have been circulating for some time. The *potential* uses of a word (sense extensions, contextual modulations, and so on) are not reflected in a dictionary, or even predictable, although we continuously witness the spontaneous creation of novel senses of existing words or even new words, which is an important source of linguistic creativity.

Finally, the information on how words function in context is very limited in book dictionaries, and it does not allow relating the lexical features of the

described item to their syntactic projection. The only syntactically relevant feature provided in the example (1) is the syntactic category (verb or noun). It does not explain, for example, why the noun *race* can be the direct object of the verb *witness* or can be modified by the adjective *fast* (unlike many other nouns, such as *table*: we cannot *witness a table* and there is no such thing as *fast table*).

From our discussion above, it should be clear that conventional dictionaries are not an accurate reflection of the mental lexicon: print dictionaries provide an incomplete, static view of our knowledge of words and how they are used in language. By contrast, it appears that the mental lexicon is flexible, extensible, and much more richly structured.

1.3 What Is a Word?

The notion of word is deeply rooted in the Greco-Latin grammatical tradition. It is assumed to be a basic unit of language, which demarcates the border between morphology and syntax: morphology deals with the internal composition of words and syntax deals with the combination of words. Also, it seems easy to grasp intuitively: we know a word when we see one, so we think. When presented with the following list, a native speaker of English would: surely identify (2a) as a word; reject (2b) because she has never heard it before and because it doesn't look like an English word; claim that (2c) is an expression composed of four words, as is (2d) (probably with a little hesitation about the status of *a*, to which we will return later).

(2) a. orange
b. aeegf
c. Give it to me!
d. become a compulsive liar

DISCUSS

Do you agree with this? Whether or not your opinion is the same, try to justify your answer. Use the following as guiding questions:

- What criteria did you use to decide whether or not a given sequence of letters is a word? Did you consistently use the same criteria for all the cases?
- Should we define *orange* as one or two words, depending on whether it refers to the fruit or the color? Are both senses related?
- Would equivalent linguistic expressions in other languages qualify as words, too?
- Would your answer be the same if you perceived these expressions acoustically and not visually?

As you probably realized at this point, a considerable amount of complexity is hiding behind this intuitively simple notion, especially if we try to find a more technical definition of "word". To a large extent, this complexity is due to the inherently intricate nature of the word, which is a highly structured object: it has internal constituents (studied in morphology) and it is used to build bigger linguistic expressions (studied in syntax); it conveys meaning (studied in semantics); and it has graphic and acoustic identity (as studied in orthography, and phonetics and phonology, respectively). Depending on which of these perspectives on wordhood we adopt, we will end up with the following definitions:

(3) **Orthographic word**
A sequence of letters written consecutively, with no blank spaces.

In the above examples, (2a) would certainly count as a word according to this definition, which seems to confirm our initial intuitions. But so would (2b), which certainly seems counterintuitive because we cannot assign any meaning to this sequence of letters (we take up the meaning component of wordhood shortly). (2c) and (2d), in turn, have four orthographic words, but we would probably have to answer differently when dealing with other languages: the Spanish equivalent of *give it to me* is written with no spaces (*dámelo*, lit. 'give-me-it'), and *become a compulsive liar* is expressible with a single verb in Russian: *izolgat'sja*. Does it make sense? As it turns out, there is much cross-linguistic variation in the definition of word.

(4) **Phonological word**
"A string of sounds that behaves as a unit for certain kinds of phonological processes, especially stress or accent" (Aronoff and Fudeman 2011: 40).

The stress criterion is relevant in English and in many other languages (although there are many other phonological criteria, depending on the language) where a phonological word must have a stress. But, again, we find a counterexample in (2c): under normal circumstances, *give it to me* is pronounced with just one main stress, on *give*, which makes the whole expression one phonological word and seems to indicate that *it*, *to*, and *me* are not treated like words by the phonology. These kinds of linguistic elements, which are syntactically (and sometimes orthographically, as in English) independent, but which cannot stand alone phonologically, are called *clitics*: they usually need to be incorporated into the prosodic structure of adjacent words.

🔍 *Prosodic* or *supersegmental features* are stress, rhythm, pitch, volume, and intonation. They characterize strings of sounds or phonemes rather than individual sounds.

Consider the ways that stress can convey different kinds of meaning: in (a) below the pronoun *it* is used *metalinguistically* (we talk about the word *it* instead of using it to refer to a real-world entity); the preposition *to* in (b) carries

contrastive stress, to highlight the opposition between *to* and *from*; finally, the personal pronoun *me* in (c) is used after a pause in an elliptical construction.

a. What does *it* stand for in this sentence?
b. I said I was returning *to* Madrid, not *from* Madrid.
c. What do you think about this book? – Who, *me*?

(5) **Content words** (lexical categories)
Content words encode a rich conceptual meaning, related to the real world.

(6) **Function words** (functional categories)
Function words denote much more abstract, language-internal meaning.

The lexical versus functional category distinction is a fundamental one in syntactic studies, and we are following the syntactic perspective on it in this section (although it is not the only one, as will be demonstrated).

Lexical categories are also called *content words* or *semantic words* because they carry meanings that denote or describe things in the world: they refer to real entities (*table, sun, book, crowd*), imaginary entities (*unicorn, horcrux, holodeck*), events (*blow, coronation; teach, explode*), and properties (*whiteness, wisdom; tired, pretty*). In most approaches, lexical categories include nouns, adjectives and verbs. The lexical categories largely overlap with the so-called "open-class lexical items", the latter term referring to the fact that new words can be introduced into the lexicon by speakers of the language, and that they can also fall into disuse. The acquisition of open-class words is a life-long process for every speaker. It is also usually a conscious one: even young children can explain what *table* or *milk* means, and can point to the objects in the real world denoted by these words.

Functional categories, on the other hand, represent a very restricted and synchronically stable (closed) repertoire: the speaker cannot coin a new determiner or auxiliary verb voluntarily. The following syntactic classes are usually considered functional categories: prepositions (e.g., *in, to, of*), conjunctions (e.g., *and, or, until*), complementizers (*that, if, whether*), pronouns (*she, him, themselves*), determiners and quantifiers (e.g., *a, the, her; most, more, several*), and auxiliary verbs (*do, be*).

The function words are acquired for life in early childhood and their content is much less accessible consciously, which is one of the reasons why adult second language learners usually have a hard time acquiring these categories even if their vocabulary is extensive in general: how would you explain to your Czech friend what *the* or *a* means (Czech has no indefinite/definite distinction)? The meaning of these words seems to be language-internal and very abstract: e.g., *a* introduces new entities belonging to a certain class into the discourse (8a), and *the* establishes definite reference: it appears in nominal groups referring to uniquely identifiable entities, which are known from previous contexts (8b):

(7) a. I saw {a/ the} cat.
b. Viděl jsem kočku.
(8) a. I just saw *a cat*. (in a context where there was no cat before)
b. I just saw *the cat*. (in a context where a cat had been mentioned or alluded to previously)

Some function words, however, do have a much more transparent meaning. The semantic content of spatial prepositions (*under, on, in, between*, etc.), for example, is much easier to grasp. While the distinction between lexical and functional words is not a clear one in many cases, it has far-reaching consequences for syntactic theory and the articulation of the relationship between syntax and the lexicon: although both kinds of elements are sound–meaning pairings, roughly, function words belong to the realm of syntax, since they are instrumental in combining different words in phrases and sentences, and lexical words belong to the realm of lexicon, understood as the repository of different kinds of idiosyncratic, not uniquely linguistic, information. If we compare syntactic structures with buildings (again, following the syntactic perspective on this opposition), the lexical words would be the bricks or building blocks, and the function words would be the cement or the glue that holds the bricks together. Without the function words, the syntactic structure would fall apart. For example, the sentence (2c) above would turn into the generally uninterpretable *Give!* if we remove the function words.

Considering how complex and typologically heterogeneous an entity a *word* is, we might ask whether it should still be considered a linguistic technical term at all. Researchers working in different subfields of linguistics seem to opt for a positive answer to this question. Independently of which criterion of wordhood (orthographic, phonological, or grammatical) prevails in any given language, there seems to be reliable psychological evidence that 'word' is an important *cognitive* unit. One of the proofs is the *word superiority effect* detected in psycholinguistic studies: letters are processed faster and more accurately when they are embedded in a word rather than presented alone or within a random sequence of letters, which indicates that words have some kind of access advantage as compared to non-words or single letters.

At this point, we are ready to introduce the definition of 'word' as adopted throughout the book. Unless otherwise noted, we assume that 'word' used in the context of linguistic theory has the following definition:

(9) **Word**
A "meaning–form" pairing (i.e., association of an acoustic or graphical form and meaning) used in forming a sentence in a language and intuitively recognized by native speakers as the basic unit of meaningful speech.

🔍 This definition comes from the classical Saussurean definition of *linguistic sign*, which has been used in linguistics since the beginning of the twentieth century: "the linguistic sign unites a concept [signified] and a sound–image [signifier]" (Saussure 1959).

1.4 Lexeme, Word Form, and Grammatical Word

After this detailed review of all the things that *word* stands for, you may still wonder whether there is a way of referring to a word in all its richness, with all these nuances? In fact, there is one. *Lexeme* is the term used to refer to the word as an abstract linguistic unit, which represents all the information we associate with a lexical item out of context, i.e., its inherent features. It is abstract because words are always used in context (except the *metalinguistic uses*), and because none of the words we actually hear or read are lexemes: they are (physical or acoustic) *word forms* or *grammatical words*.

We can also talk about lexemes as *types* and word forms as *tokens*. The type–token distinction is familiar to anyone who has dealt with multiple instances of the same kind of object: coins, apples, books, etc. If you buy two copies of the same book, you have two *tokens*, but just one *type*. In the actual use of linguistic utterances we make reference to linguistic types but each use of the type is a specific token. For example, in (10), the number of **tokens** of the **type** *the* is 3: i.e., # of types = 1, # of tokens = 3.

(10) **The** boy bought **the** pizza at **the** store.

In order to refer to a lexeme (e.g., in a dictionary), one of its forms is used conventionally, which is called the *lemma* or the *citation form*. For example, in most European languages the infinitive is used as the citation form for verbs (*go*, *break*), the singular masculine form is used for adjectives and participles (*green*, *broken*), and the singular is used for the nouns (*woman*, *house*).

Once the lexeme is inserted into the syntactic context, it becomes a *grammatical word* or *word form*. In the sentence, the lexeme acquires additional syntactic features. In the following example, there are two word forms corresponding to the lexeme *to be* (*were* and *was*) and two word forms corresponding to the lexeme *week* (*week* and *weeks*):

(11) **Were** you in Cape Cod last week? – Yes, I **was** there for the last two weeks.

In both cases, the verb acquires the past tense features and, in addition, it must agree in person and number with the subjects *you* and *I*, respectively. The noun, on the other hand, is realized in the singular and plural grammatical number forms (*week*, *weeks*, respectively).

1.5 What Is in a Lexical Entry?

In order to understand what information needs to be represented in a lexical entry, imagine for a second that you are given a list of new terms in English, which you have to memorize and learn how to use in context. What kind of data would you need to include for every lexical item, in order to achieve this goal? As we will see in more detail in Chapter 6, there are at least four

types of information in a lexical entry: *orthographic* (how do I write the word?); *phonological* (how do I pronounce the word?); *semantic* (what does the word mean?); and *syntactic* (what other words is this one compatible with in a phrase or a sentence?).

Let's take as an example the word *table*. We first need to specify its spelling (*table*) and pronunciation (/ˈteɪbəl/). As we have seen, semantic information is often provided through paraphrases called *definitions* in the dictionaries. The following is the definition of *table* taken from the Merriam-Webster Online Dictionary:

(12) **table**: a piece of furniture with a flat surface that is designed to be used for a particular purpose.

Like most dictionary definitions, this one is composed of two parts: the *definiendum*, which is the word being defined; and the *definiens*, which is how it is defined. The definiens makes reference to two features: (a) the larger class in which the word meaning is included (i.e., *furniture*); and (b) the set of distinctive features which differentiates this concept from the other members of the same class (i.e., *with a flat surface* and *designed to be used for a particular purpose*).

As shown at the beginning of this chapter, the syntactic information provided in the dictionaries is usually limited to the syntactic category: e.g., *table* is a noun. This is an important piece of data, from which we can deduce (with the help of syntactic rules) that *table* is projected in the syntax as the head of the noun phrase. As such, it is compatible with certain kinds of complements (adjectival and prepositional phrases: [$_{NP}$[$_{AP}$ red] table], [$_{NP}$ table [$_{PP}$ by the window]]), and it can be used as a complement of higher-level functional projections, for example within a determiner phrase (DP): [$_{DP}$ the/three [$_{NP}$ red tables]]. However, it says nothing about more fine-grained requirements this noun imposes on its complements. Recall the ungrammatical **fast table* example: if any kind of adjective was compatible with *table*, *fast* should be able to combine with it, just as well as the adjectives *red* or *nice*, but this is not the case. Also, it does not account for the interpretation of *useful table*, for example: the speakers know that it means 'useful for a particular purpose' (writing, cooking, etc.), but this meaning cannot be inferred from the dictionary definition.

Print dictionaries are a useful source of linguistic data, but as we will see in the following chapters the structure of the lexical entry in the mental lexicon requires a much richer and much more complex representation. Chapter 6 deals with this topic in detail and shows formally how the different kinds of information are connected in a lexical entry (in particular, how the word meaning components affect the syntactic behavior of the word).

1.6 The Role of Empirical Data in Lexicon Research

Determining the appropriate kind of data is, of course, critical for the development of a theory in any empirically oriented field of scientific research.