TABLE OF CONTENTS

1. OBJECTIVES

In this chapter we deal with words and concepts. More specifically, we shall learn some facts about differences, similarities and the overlapping areas among words, concepts and their respective relations with the codification of meaning.

We will also be introduced to how ontological material is represented for linguistic applications and the differences and similarities between general linguistic representations and specific representations for linguistic applications proper.

2. INTRODUCTION

In this chapter some basic semantic concepts are refreshed, with an emphasis on those which have a particular impact on the development of both ontologies and dictionaries.

The contribution of the different levels of linguistic analysis to the construction of meaning focuses on the different aspects of language: how the units of human-produced sounds are organized in words so as to be meaningful is something studied in phonetics; how words are made up and further combined in higher meaningful units is studied by both morphology and syntax; semantics basically studies how meaning is codified and how it does so from these different linguistic perspectives.

Because dictionaries focus on the meaning of words, lexical and morphosyntactic perspectives in linguistic analysis are important. The understanding of a word meaning by the users of a certain language, and their capability to explain it by putting it into other words, are preconditions for the creation of dictionaries.

Because ontologies focus on how concepts are captured and how they are codified in words, semantic analysis is also a kind of precondition for the further construction of applications such as programs called "ontologies".

The working perspective taken here is basically one that presupposes an online use. This means that both products, dictionaries and ontologies, are seen as digital products and therefore subject to computational treatment. Ontological and lexical representations in language applications are introduced to highlight their coincidences and overlapping areas. Then, the points where they coincide and overlap are noted.

Focusing on lexicography, Hanks (2003) provides a brief review of its basic aspects by linking them to their historical background. This is useful in helping us connect the present developments of language technologies with their roots and also in identifyng their most important varieties and initial developments. As Hanks (2003: 48) states, lexicographical compilations are inventories of words that have multiple applications and are compiled out of many different sources (manually and computationally):

> An inventory of words is an essential component of programs for a wide variety of natural language processing applications, including information retrieval, machine translation, speech recognition, speech synthesis, and message understanding. Some of these inventories contain information about syntactic patterns and complementation associated with individual lexical items; some index the inflected forms of a lemma to the base form; some include definitions; some provide semantic links to ontologies and hierarchies between the various lexical items. Finally, some are derived from existing human user dictionaries.

However, he concludes that none of them are completely comprehensive and none of them are perfect.

As with most aspects of our everyday life, dictionary compilation such as the craft of lexicography has been revolutionized by the introduction of computer technologies. On the other hand, new insights have been obtained by computational analysis of language in use, providing new theoretical perspectives.

Concepts and words are related, because for concepts to be transmitted we, as humans, need words, and this is why it is important to differentiate between concepts and words. Understanding the similarities, differences and interrelations between them in the present situation of massive use of the internet, where we interact with machines such as computers all the time, becomes more and more important. And this is why we will try to differentiate between those applications that are more heavily dependent on language and those which, being of a more abstract nature, can be "described" as language independent.

This *concept-word* relationship concerns the process of conceptualization. As Prevot, *et al.* (2010: 5) explain:

> The nature of a conceptualization greatly depends on how it emerged or how it was created. Conceptualization is the process that leads to the extraction and generalization of relevant information from one's experience. Conceptualization is the relevant information itself. A conceptualization is independent from specific situations or representational languages, since it is not about representation yet. In the context of this book, we consider that conceptualization is accessible after a specification step; more cognitive oriented studies, however, attempt at characterizing conceptualizations directly by themselves (Schalley and Zaefferer 2006).

Precodification of entities or relations that usually lead to the lexicalization of nouns and verbs is a specification step. This is the marking of either an entity or a relation in the notation of an ontology. Let us illustrate this: For example, the verb *run* as in "She runs the Brussels Marathon" is precodified as a 'predicate' and thus as a 'relation', and the nouns *she*, *Brussels* and *marathon* are precodified as entities. On the other hand, a possible example of direct cognitive type of conceptualization in the sense of Schalley and Zaefferer (2006) could be the famous one of asking for food in the context of a restaurant. In fact, both types of conceptualization are compatible.

What is the objective of an ontology? Basically, it is to conventionalize concepts in order to handle meaning and knowledge efficiently. As Prevot, *et al.* (ibidem) explain:

> Every conceptualization is bound to a single agent, namely, *it is a mental product which stands for the view of the world adopted by that agent;* it is by means of ontologies, which are language-specifications of those mental

> products, that heterogeneous agents (humans, artificial, or hybrid) can assess whether a given conceptualization is shared or not, and choose whether it is worthwhile to negotiate meaning or not. The exclusive entryway to concepts is by language; if the layperson normally uses natural languages, societies of hybrid agents composed by computers, robots and humans need a formal machine-understandable language.
>
> To be useful, a conceptualization has to be shared among agents, such as humans, even if their agreement is only implicit. In other words, the conceptualization that natural language represents is a collective process, not an individual one. The information content is defined by the collectivity of speakers.

There are two —opposed and complementary— ways to access the study of words and concepts: the onomasiological approach and the semasiological approach.

The first one, whose name comes from the Greek word ὀνομάζω (onomāzo), 'to name', which comes from ὄνομα, 'name', adopts the perspective of taking the concept as a starting point. Onomasiology tries to answer the question *how do you express x?* As a part of lexicology, it starts from a concept (an idea, an object, a quality, an activity etc.) and asks for its names. The opposite approach is the semasiological approach: here one starts with the word and asks what it means, or what concepts the word refers to. Thus, an onomasiological question is,*what is the name for medium-high objects with four legs that are used to eat or to write on them?* (Answer: *table*), while a semasiological question is, *what is the meaning of the word* table*?* (Answer: medium-high object with four legs that is used to eat or to write). The onomasiological approach is used in the building of ontologies, as we will see in depth in chapter 5, and the semasiological approach is adopted for the construction of terminologies, banks of terms, to be applied in different areas, as we will see in chapter 6.

3. THE USE OF A METALANGUAGE

A much debated issue in relation to these matters is the use of a metalanguage. Saeed (2003) defines semantics as *the study of meaning communicated*

through language. Since there are quite a number of languages and since meaning and knowledge are, to some extent, interchangeable terms, we can say that knowledge representation is fairly connected to the particular language on which the referred knowledge is expressed. Consequently, in his preliminary discussion of the problems of semantics this author suggests that the use of a metalanguage could be a possible solution to the problem of the circularity of the meaning of a word in a dictionary. Setting up a metalanguage might also help to solve the problem of relating semantic and encyclopedic knowledge, since designing meaning representations of words, involves arguing about which elements of knowledge should be included. But metalanguages also present problems for lexical representation. After all, most linguistic models of all kinds (generative, functional etc.) have designed a metalanguage of their own, more or less based on linguistic signs, to represent what the linguist in question considers to be the set of foundational concepts upon which their subsequent linguistic representations are built.

Generally, a metalanguage will be necessary to build up any ontology, especially if it is aimed to be applicable to more than one language. There are two kinds of components in an ontology that make use of a metalanguage: the represented categories and relations, and the represented procedures. Sometimes the represented procedures are just the relations themselves.

An example of a metalanguage combining meaning postulates and thematic frames for the event +ANSWER_00, as in Periñán Pascual and Mairal (2010: 20) is:

(1)

Thematic Frame: (x1: +HUMAN_00) Theme (x2) Referent (x3: +HUMAN_00) Goal

Meaning Postulate: PS: +(e1: +SAY_00 (x1) Theme (x2) Referent (x3) Goal (f1: (e2: +SAY_00 (x3) Theme (x4: +QUESTION_00) Referent (x1) Goal)) Scene)

The thematic frame of the event +ANSWER_00 belongs to the higher frame of the metaconcept #COMMUNICATION, to which the metaconceptual unit is assigned a prototypical thematic frame. In this case, the thematic frame of communicative situations, from which we obtain all other conceptual units related to this metaconcept of #COMMUNICATION is:

(2)

(x1) Theme (x2) Referent (x3) Goal

The Thematic Frame in (1) derives from this general one in (2), as well as the Meaning Postulate, which gives more detailed conceptual information about the specific event +ANSWER_00, which can be paraphrased as "someone (x1) **say** something (x2) to somebody (x3) related to a question (x4) that x3 said to x1". All the symbols used (+, #, x, numbers, etc.) are part of the metalanguage used in COREL[1] (which stands for *Conceptual Representation Language*) for the representation of concepts.

o-o-o-o-o-o-o

Exercise 1:

Build up your own metalanguage: propose a small ontology -four concepts or so- for concepts related to a specific conceptual field (for example verbs of emotions, or verbs of movement) and select one or two of these concepts. Then, "invent" a series of symbols that you would use in order to represent the entities and relations involved in these concepts. You can use some parts of English as a pro-metalanguage (as Dik 1997 or Van Valin 2005 do), or you can suggest new symbols.

o-o-o-o-o-o-o

4. LANGUAGE DEPENDENT AND LANGUAGE INDEPENDENT KNOWLEDGE REPRESENTATION

The representation knowledge based on language or based on concepts differs in a number of aspects. It is a circular issue and it always affects the creation of ontologies and dictionaries.

The concept of prototypicality is a highly influential one in both lexicographic and ontological studies, and it works in both directions. As

[1] COREL, which stands for Conceptual Representation Language, is the language used within the Lexical Constructional Model of language and by the project group FunGramKB (Functional Grammar Knowledge Base) in order to build up a whole conceptual ontology. See: http://www.fungramkb.com/default.aspx

Geeraerts (2007: 161) notes, how prototypicality effects in the organization of the lexicon blur the distinction between semantic information and encyclopedic information. This does not mean that there is no distinction between dictionaries and encyclopedias but that the references to typical examples and characteristic features are a natural thing to expect in dictionaries. On the other hand, in the construction of ontologies prototypical examples of categories tend to be linked to higher categories or inclusive categories, usually taking the lexical form of hyperonyms.

4.1. Language dependent applications

Language dependent knowledge representation is based on the way in which a certain language codifies a certain category —for example a certain state of affairs— because it affects the representation of this particular piece of world knowledge. For example, a debated issue is how the different languages of the world and English in particular, lexicalize (with more or less detail) certain aspects of their external world that affect their speakers.

Within the general studies of semantics we see how there are certain linguistic categories that are highly dependent on the language in which they are identified. Remember, for example, the pronominal system in Arabic languages contrasting with that of English or Spanish, where the former codifies a dual pronoun whereas the latter two only codify singular and plural. As it is well known, there are more or less universal linguistic features that all languages codify with various syntactic, morphological or lexical devices to refer to the addressee. An example of language independent knowledge or conceptual representation could be the lexical terms for numbers, or the mathematical symbols and the symbols of logic (see Chomsky 1975, Dik 1997, Van Valin 2005, and so many others, for different examples).

Language is a conceptual phenomenon, as postulated by Lakoff (1987) and others. This means that different languages lexicalize with more or less detail those aspects (concepts) of the external world affecting their speakers in a particular way. For example, the different types of snow are lexicalized with different words in Eskimo languages, in the same way that the differ-

ent types of winds are given different names in many cultures, where the type and intensity of the wind directly affects people's everyday lives.

An example is the Spanish word, *chirimiri* or *sirimiri*, deriving from the Basque *(txirimiri)* used to refer to a kind of rain characterized by water drops that are very small and in abundance, so that you are anaware that you are getting wet but in fact you are. In www.wordreference.com it is translated as "fine drizzle", but there is not a single unique term in English that can represent this specific kind of rain that is typical of the Basque Country. Still another example is taken from our urban kind of life, where we have lexicalized two different terms for human beings depending on whether such human being is or is not driving (*driver /pedestrian*). In addition, the words *rider* or *caballero* lexicalize the fact that a human being is or is not riding a horse, because such a difference was important before the invention of the car. Nowadays, a rider is also someone riding a bike, also in opposition to a pedestrian, who is not using any other way of moving but his/her own legs.

Therefore, the way a certain language codifies a certain category —for example a certain state of affairs— affects the representation of this particular piece of world knowledge because it selects some elements instead of others. As already mentioned, the codification of a certain state of affairs, for instance, affects the types of constructions that a certain language may produce. For example, in English the codification of a resultative element in a certain state of affairs leads to constructions like the famous *wipe the slate clean* study in Levin (1991).

Knowledge representation in language applications

A particular kind of knowledge representation is based on lexical organization. This organization can take many forms. For example, a network is a group of words that are not so tightly organized. Sets can be defined as organized and bounded groups of words, while hierarchies are organized groups of words usually following a certain semantic relationship (e.g.: hyponymy). In inheritance systems the main link is a certain feature (semantic or syntactic) that can be identified as recurrent at different levels of a structure. Understanding how words are organized in a certain format is important for both dictionaries and ontologies.

Meaning representation in language applications

Lexical representation is, after all, meaning representation. It must be noted, however, that lexical meaning is just one part of the whole meaning transmitted in verbal communication, where the total content of information transmitted is usually more than purely verbal communication. The point is that in order to facilitate transmission between speakers, language is organized using a limited number of formal structures of different kinds (lexical, syntactic, morphologic, phonetic), which are complemented with a whole repertoire of ontological and situational (sensory-perceptual) information simultaneously processed. Formal codification of purely linguistic input is inevitably limited because of processing constraints, but it is complemented with additional non-linguistic information that enters the processing human system via other non-linguistic means.

The difference between trying to represent meaning and trying to represent other linguistic levels is that it is easier to represent something with a more or less tangible side such as the lexicon, syntax, morphology, or phonetics of a certain language. Trying to represent something like meaning, highly dependent on contextual information, is not an easy task. Part of the difficulty is that it is precisely the lexicon, syntax, morphology, or phonetics of a certain language that we use to convey meaning. Only a small part of the more salient and socially engraved aspects of social behavior constitutes contextual information codified in human languages in many different ways, and which can be labeled with different kinds of pragmatic parameters.

In addition, other non-pragmatically codified information is missing in linguistic representation as such, and must enter the system through other non-linguistic-processing-input-systems. The paradox is that in order to other codify all that non-linguistic information we sometimes need a language. Whether this language is a conventional language, a metalanguage or another symbolic system is a different question to be addressed. Sometimes it is very difficult to think of fully language-independent representations. In what follows, we will deal with this topic in more detail.

Lexical representation in language applications

A typical example of lexical representations for language applications is a common parser, and the clearest example of a language dependent linguistic application is a dictionary of any kind.

4.2. Language independent applications

An easy example of a language independent representation could be the figures for numbers or the mathematical symbols that mathematicians of all languages use.

In this book, language independent applications will be studied at a very basic level and under the perspective that knowledge is partly organized independently of the language in which it is put into words and partly organized in a sort of network. A knowledge network is understood as a collection of concepts that structures perceived information and allows the user to organize it.

So an example of language independent application is mathematical notation. What is represented in a mathematical formula is simple: a series of mathematical concepts and a series of relations linking them:

(3)

$$(5 + 3).\ 5 = 45$$

Here we have two types of quantities: one grouped in two sub entities (5 and 3) and the other is the whole amount, just by itself. And the relations linking these quantities are shown below:

(4)

[( )] represents a set.

[+] represents the addition operation of natural numbers.

[.] represents the multiplication operation of natural numbers.

[=] represents the result of both operations.

Finally, the organization of concepts into hierarchies is also relevant. Each concept is studied as related to its super- and sub-concepts. The inheritance of defined attributes and relations from more general to more specific also affects complex conceptualizations.

4.3. Conclusions about knowledge representation

Knowledge representation has taken the form of printed and electronic (both on and off-line) dictionaries and ontologies. It is self-evident that dictionaries are one possible kind of language-dependent instrument of knowledge representation in the sense that dictionaries compile all or most parts of the lexical information of a particular language.

Ontologies, on the other hand, compile other than lexical information. However, it is also becoming evident, as said above, that in order to codify non-lexical information, lexical means are needed. Ontologies can be defined as knowledge networks. A knowledge network is a collection of concepts that structure information and allows the user to view it. In addition, concepts are organized into hierarchies where each concept is related to its super- and sub-concepts. All this forms the basis for inheriting defined attributes and relations from more general to more specific concepts. This is seen more in depth in chapter 5.

5. DIFFERENT APPROACHES TO MEANING INTERPRETATION

According to Prevot *et al.* (2010), the topic of the interface between ontologies and lexical resources is a re-examination of traditional issues of psycholinguistics, linguistics, artificial intelligence, and philosophy in the light of recent advances in these disciplines and in response to a renewed interest in this topic due to its relevance for the Semantic Web major applications.

These studies also recognize the importance of a multidisciplinary approach for lexical resources development and knowledge representation and the influential contributions to the field of Hobbs *et al.* (1987),