Natural Language Searching in Onomasiological Dictionaries

Gerardo Sierra
Instituto de Ingeniería
Universidad Nacional Autónoma de México
gsierram@ii.unam.mx

Abstract

When consulting a dictionary, people can find the meaning of a word via the definition, which usually contains the relevant information to fulfil their requirement. Lexicographers produce dictionaries and their work consists in presenting information essential for grasping the meaning of words. However, when people need to find a word it is likely that they do not obtain the information they are looking for. There is a gap between dictionary definitions and the information being available in peoples' mind. This paper attempts to present the conceptualisation people engage in, in order to arrive at a word from its meaning. The insights of an experiment conducted show us the differences between the knowledge available in peoples' minds and in dictionary definitions.

1 Introduction

Many lexicographers recognise users need dictionaries to look for a word that has escaped their memory although they remember the concept. From a semantic point of view, Baldinger (1980) takes user needs into account and thus distinguishes dictionaries that serve as aids in encoding from those that help with decoding. The best known dictionaries of this type allow users to find the meaning of a word they already know. Such dictionaries are semasiological: they associate meanings with expressions/words, i.e. within entries we move from word to meaning. The second kind of dictionary helps those users who have an idea to convey and want to find a word to designate it. Such dictionaries are onomasiological: they connect names to concepts, i.e. within entries we move from meaning or concept to name or word.

Sierra (2000) confirmed the well known observation that the organisation of the world varies from author to author, by contrasting some recognized onomasiological dictionaries, such as Roget's Thesaurus of English Words and Phrases (1852), Bernstein's Reverse Dictionary (1975), and WordNet (Miller et al., 2008).

In order to build a system that maps natural language descriptions of concepts to the terms corresponding to those concepts, Sierra and McNaught (2000) outlined the design of an Onomasiological Search System. They described the principles of the system, whereas the architecture and its components are presented as part of the design. This also includes an idealised user interface, with a discussion of the organisation of the probable terms and additional information that can help the user to identify precisely the term he is looking for.

As cognitive issues for the design of such system, this paper attempts to present the conceptualisation people engage in, in order to arrive at a word from its meaning. In this sense, it breaks the traditional lexicographic assumption that one should utilise a semasiological approach to provide formal representations to describe the meaning of a word. In contrast, in the onomasiological approach, the user can formulate a concept in several ways and use a variety of words in order to find a particular word.

Our starting point is to understand what conceptualisation is (section 2), and the process of designation (section 3). To validate our approach in practice, section 4 presents the results of an experiment, which is compared with other stud-

Coling 2008: Proceedings of the workshop on Cognitive Aspects of the Lexicon (COGALEX 2008), pages 32–38
Manchester, August 2008

ies on conceptual analysis. Finally, the conclusions are stated.

2 Conceptualisation

The concept is a mental representation of an object which is formed in the mind of individuals through a process of abstraction. We call this process of constructing an internal representation of external things a *conceptualisation*. Conceptualisation is a mental activity of grouping the data of common properties according to external factors, and then concepts are internalised and form part of each individual's knowledge.

2.1 Properties

The terms *property*, *characteristic* or *feature* have been used for the knowledge necessary to describe and classify a concept. The identification of properties is crucial to concept analysis since it helps to define concepts and identify their interrelationships.

According to Sager (1990) some properties are necessary and sufficient to distinguish a concept from any other, and these properties reflect the essential characteristics of a concept. Conversely, other properties are inessential, merely observable in an individual thing, so that they are accidental, may change with time and may not even be necessarily true in a scientific sense (Petöfi, 1982).

The dichotomy of these two opposing types of properties has been discussed from a psycholinguistic point of view. Aitchison (1994) does not consider that it is obvious that experts and ordinary people distinguish between essential and inessential characteristics. Despite the fact that experts might be able to specify the true nature of things, they sometimes provide information which is irrelevant to the mental lexicon. Conversely, ordinary people disagree and change their minds.

As the comparative analysis in section 4 shows, the essential characteristics are not necessarily present in the mental lexicon of a person; each one describes different properties. Nevertheless, even a description of the inessential characteristics, given together, provides enough information to identify the term (Wierzbicka, 1985).

2.2 Social conceptualisation

People acquire knowledge about things on the basis of cultural, geographical and social factors. The environment conditions the conceptualisation of reality and the use of language. In order to communicate effectively, people will try to use language in a similar way to that of the collective view of the community, in agreement with the social norm. In fact, because of the social norm, there is an idealised knowledge structure which makes it possible to use the same names for the same things. In the contrary case, when the designation of a concept is outside that norm, people assume that the individual's knowledge is wrong. However, as we will see in the final analysis, we must accept that an individual's knowledge cannot cover the whole knowledge of the community norm.

2.3 Individual conceptualisation

Reality goes beyond the perception of individuals. Our knowledge about reality has increased throughout human history. No one – human, computer or even the biggest library– possesses the whole knowledge about reality.

The knowledge structure of things varies from one person to the other, so that their description of concepts will be different. As Fugman (1993) states, since the number of properties is virtually unlimited, people concentrate on those characteristics which appear essential, according to their personal or professional view. As an example, he points out the different essential properties for the concept "benzene" given by a physicist, a biologist, an engineer, a fire-fighter and a chemist.

Even the same person can demonstrate different conceptualisations of simple things, such as "dog" or "apple", depending on the contextual situation. For example, a dog seen in different domains, such as a conference, a zoology lecture, a road or a house, may be described as canine, mammal, animal or pet.

3 Designation

The process of designation is the opposite of signification. Signification is the identification of the meaning of a word, and the result of finding a meaning is a definition. Designation is the identification of a term for a concept and the result is a word.

To retrieve a term, one can use a terminological definition, which provides the information necessary to identify and differentiate a concept within a system of concepts, so that it sometimes comprises encyclopaedic information, not usually necessary in a lexicographic definition.

3.1 Properties

Just as a word may have many meanings (semasiological approach), a concept, which is described by a set of properties, may be designated by more than one word (onomasiological approach). Within the onomasiological approach, all the properties together provide the necessary and sufficient information to identify the concept. However, since the description of concepts in natural language does not incorporate all knowledge or ideas associated with each concept, it can happen that the projection of a concept, i.e., the query formulation of the user, will retrieve a set of terms. For example, the concept "strong winds", consisting of two properties, can retrieve, in the domain of weather terminology, a variety of terms, such as: "gale", "tornado", "hurricane", "typhoon", etc.

The concept a user has in mind when looking for a target word is expressed by a sentence. When a person hears this sentence, he translates each word into his own language and easily identifies the context. A person may understand the expression "that which determines air pressure" and get a mental representation of "that" for "thing" and then for "instrument"; or that the speaker might have said "measures" instead of "determines". From the context, at the same time, he may differentiate whether the word "air" refers to the atmosphere or the air of a tyre.

Equally, either the lack or change of any one property may result in the identification of a different concept. For example, take the following definition for "barometer":

- A device to measure air pressure.

Each of the four keywords yields a property. Then, we can change one property at a time and get a different concept. If we change

- "Device" to "unit", the result is the concepts "inch" or "millibar".
- "Measure" to "provide", the result is "air scoop".
- "Air" to "blood", the result is "sphygmomanometer".
- "Pressure" to "humidity", the result is the concept "hygrometer".

4 Comparative analysis on conceptualization

In order to verify some of the ideas presented above, an experiment was carried out with a small group of twenty undergraduate students. Although a small group is unrepresentative for any generalisation to be made from a statistical point of view, it has been sufficient for our purpose to demonstrate that the conceptualisation used by a random set of students is far from the definitions found in a dictionary.

From two sets of five words, each student was asked to take a set and write on a blank sheet of paper, similar to an onomasiological search, a concept, a definition or the ideas suggested to them by each word. After interchanging the sheets, the other students participating in the experiment wrote the word or words designating the concepts identified or written on the blank sheets by the previous student.

The sets of word given contained three general language words and two terms.

- Set A: water, squirrel, bench, euthanasia, hurricane.
- Set B: lemon, bucket, clothing, monopoly, barometer.

The general words were chosen because they permit us, as can be observed from the following sections, to compare the results with the words analysed by other researchers as well.

We will next introduce our definitions by comparing with four studies on conceptualisation.

4.1 Putnam

Putnam (1975) proposes the representation of the meaning of a word as a finite sequence of at least four properties:

The *syntactic markers*, which are the category-indicators used in a host of contexts to classify words.

The *semantic markers*, which are the most central properties, form part of a widely used classification system and very may be affected by any change in the knowledge about the thing.

A description of the features of the *stereotype*, which is a conventional idea of what the object looks like or acts like or is, regardless whether it is true or not for all the objects. For example, "yellow" is a stereotype of "gold", even when gold is white by nature.

A description of the *extension*, i.e., the set of things of which a term is true. The extension is determined socially depending upon the nature of the particular things, rather than on the concept of the individual speaker.

The first three properties belong to the individual competence of speakers. The extension

does not necessarily have to be known to every member of a linguistic community.

According to Petöfi (1982), the semantic markers and the stereotype may be compared with Ullman's concept of meaning. From the perspective of the lexicographic definition, they resemble genus and differentia, respectively.

The description of the meaning of "water", as a particular natural kind, following these components, is given in table 1.

Syntactic markers	**Semantic markers**	**Stereotype**	**Extension**
mass noun	natural kind	colourless	H_2O
concrete	liquid	transparent	(give or take
		tasteless	impurities)
		thirst-quenching	

Table 1. Properties for the natural kind "water"

In order to permit comparison of the definitions given in our experiment with his meaning of "water", the same four kinds of properties are used. Our definitions, as shown in Table 2, include the property "fluid", which easily can be classified as a semantic marker.

n.	***Concept***
1	It's a clear liquid that you get from a tap
2	The colourless transparent liquid occurring on rivers
3	A clear, neutral liquid that surrounds us everywhere
4	Liquid, clear, drinkable – constituents are hydrogen and oxygen
5	Liquid, clear, H_2O
6	Liquid form, scientific term H_2O
7	Liquid, freezes at 0°C
8	Liquid, clear, boils at 100°C, freezes at 0°C
9	Fluid, clear, tasteless, colourless
10	Wash with it; drink it; used for dilution; H_2O; found in springs, rivers, lakes, seas, oceans

Table 2. Conceptualisation of water

The properties referring to the boiling and freezing points of water, given in definitions 7 and 8 in our experiment, may be considered as part of the concept's extension, since these properties depend upon the nature of the water.

Therefore, definitions one to nine include the semantic marker "liquid", beside the particular features of water, the stereotypes, and/or the description by extension. The definition ten, which does not include the semantic marker, describes water by extension.

4.2 Wiegand

Wiegand (1984) tries to identify the properties of a definition by means of a *scale of usability* obtained statistically from a questionnaire to 100 students. He suggests 21 properties and asks the students to judge which of them describe a lemon. Each property is evaluated in three categories according to the sum of ticks it received as good, not so good and not good (Table 3).

GOOD	**NOT SO GOOD**	**NOT GOOD**
oval	yellow	tapers at both ends
juicy pulp	thick rind	oblong
sour pulp	citrus fruit	thin rind
yellow rind	green rind	used to make pectin
fruit of the lemon tree		pulp containing approx. 3.5-8% citric acid
		pulp rich in vitamin C
		variable protuberant tip
		pulp rich in vitamins
		many uses in cooking
		used to make drinks
		used to make citric acid

Table 3. Properties of lemon using a scale of usability

Even although a test with ten students is not a representative sample from which one can generalise the *scale of usability* of the properties of a concept, our experiment, as shows in table 4, allows us to challenge the values identified by Wiegand.

n.	***Concept***
1	It's a yellow fruit, like limes. Citrus. Used in cooking for sharpness
2	A yellow citrus fruit. Sour tasting. Often used as an accompaniment to drinks
3	a yellow citrus fruit with a bitter taste often sliced and put in drinks
4	It's a citrus fruit, yellow, used with sugar on pancakes
5	It's a yellow citrus fruit. Tastes bitter. Oval shaped
6	A yellow sour fruit

n.	*Concept*
7	A yellow citrus fruit
8	Yellow, citrus, fruit
9	Citrus fruit which is yellow
10	Yellow citrus fruit

Table 4. Conceptualisation of lemon

As observed in table 5, there is no match between the values for the seven properties extracted by Wiegand and our own experiment from the definitions.

Property	**Our experiment**	**Wiegand**
yellow	good	not so good
citrus	good	not so good
oval	not good	good
sour pulp	not good	good
many uses in cooking and drinks	not good	not good
variable protuberant tip	not good	not good
similar to limes	not good	---

Table 5. Comparative analysis of the properties of lemon

The reasons why these values differ are not obvious. Probably this comparison means statistical methods from a group of students are not reliable to assess the properties of concepts.

4.3 Ayto

In order to define the meaning of words, Ayto (1983) adapts the componential analysis introduced by Pottier to semantic fields. He also identifies the semantic features that characterise various sorts of seats, but analyses these characteristics to compose an analytical definition. The definition of a word is determined by the semantic features that differentiate it from other words rather than by the sum of the individual characteristics. The genus for each word in the semantic field is "seat", as it presents the only common characteristic for the rest of the set.

The differentia is determined by comparing the other characteristics and checking those which are different. The characteristics are: For several people, not upholstered, for outdoors, functional.

Then the definition for "bench" is, for example, "a seat for two or more people that has a back, is typically used outdoors, and may be fixed in position".

For a comparative analysis, it is possible to find the semantic features of the definitions in our experiment (Table 6) and try to match them with those given by Ayto.

n.	*Concept*
1	You can sit on it in the street or a park and they are made of wood
2	A long hard seat for several persons on which the players on a sport team sit
3	An object for sitting on, usually long which can seat many people
4	Sit on it (a few people can) in parks, made of wood or iron
5	Object used for sitting on. Often found in public places such as parks and gardens. Used to seat 1 or more people at a time
6	Something you seat on, is longer than a chair, usually made of wood
7	Long platform for sitting on (fit many people on one)
8	Apparatus for sitting on, designed for more than one person, often found in parks
9	A kind of seat found in parks, made of wood
10	A type of chair

Table 6. Conceptualisation of bench

For this purpose, we should assume that:

1. For several people = longer than a chair.
2. Not upholstered = made of wood or iron, hard seat, platform.
3. Outdoors = street, park.
4. Functional = for a sports team.

Table 7 presents the four semantic features used in our experiment to define "bench".

	Char. 1	**Char. 2**	**Char. 3**	**Char. 4**
Bench 1		+	+	
Bench 2	+	+		+
Bench 3	+			
Bench 4	+	+	+	
Bench 5	+		+	
Bench 6	+	+		

Bench 7	+	+	
Bench 8	+		+
Bench 9		+	+
Bench 10			

Table 7. Semantic features of bench

In the light of this contrastive analysis, it is clear that each semantic feature (for example "outdoor") can be expressed by a set of equivalent alternatives (for example, "public places", "parks").

4.4 Wierzbicka

Wierzbicka (1985) considers that a good lexicographic definition must be exhaustive, i.e., providing all the properties of the concept. Her view of a definition differs from an encyclopaedic definition because the latter conveys knowledge about the object, while the lexicographic definition does not include specialised knowledge, unless it is part of the concept. Her demand for exhaustiveness is contrary to traditional semasiological lexicography, where, through a genus and the differentia, the definition provides the essential properties to identify a concept and distinguish it from others. However, when there is a full description, we may be sure that a user will retrieve the word in an onomasiological search.

Wierzbicka uses five general properties to reach a definition of animals, e.g. squirrels, namely: habitat, size, appearance, behaviour and relation to people. Table 8 presents an example of a description for each general property.

General property	**Description**
Habitat	They live in places where there are many trees.
Size	They are not too big for a person to be able to hold one easily in both hands.
Appearance	They have a big bushy tail. Their fur is reddish or greyish.
Behaviour	They collect and eat small hard things which grow on trees of certain kinds.
Relation to people	People think of them as nice and amusing little creatures.

Table 8. Examples of full description of "squirrel"

As observed in the definitions of our experiment (Table 9), the sum of properties in our definitions agrees with the description of the five kinds of properties of Wierzbicka, although she does not consider that squirrels build nests, as one of our definitions does.

n.	***Concept***
1	It's a little rodent and can be red or grey, it has a big bushy tail
2	A small rodent living in trees with a long bushy tail
3	A small rodent which lives in trees, collects nuts and has a bushy tail
4	Animal, grey/red, bushy tail, lives in trees, buries nuts
5	Small animal, lives in trees, eats acorns, has a bushy tail
6	Animal, bushy tail, eats nuts, builds nests in trees called dreys
7	Small funny animal with big, bushy tail, likes nuts, likes trees
8	Animal that lives in trees and collects acorns, has a long tail
9	A small-sized animal, habitat in trees
10	Small grey mammal, relative to the rodent, found in both countryside and town

Table 9 Conceptualisation of lemon

5 Conclusions

The distinction between semasiology and onomasiology permits us to consider a new perspective in lexicography. In the semasiological approach, the perspective is from the dictionary to the user. Dictionaries are a lexicographer's product and definitions provide the necessary and sufficient elements in order to know the meaning of a word.

Conversely, the onomasiological approach is from the user to the dictionary. The user should provide the concept, while the dictionary interprets that concept in order to find the most appropriate word. The user can formulate the concept by several methods and may use a variety of words that in a certain context are similar. According to the user's social, cultural and geographical background, the description of the concept may differ in multiple properties.

With regard to the preceding analysis, it is worthwhile to note that even the most complete description of a concept can lack "essential" properties from the point of view of a user. None

of the methods of componential analysis, even the most open ones, has been sufficient to foresee the properties used by a small set of students. That gap should be filled with the aid of a good onomasiological retrieval system.

This does not mean that it is unlikely that we shall be able to design a complete and efficient onomasiological dictionary. In our context, efficiency means that a dictionary has to satisfy the requirements of a particular kind of user, in a certain domain of a terminology with a specific background. Therefore, the design of an onomasiological dictionary must first foresee a multiplicity of properties for each concept and secondly the diversity of words that can be used to name them. Then, the task consists in the accurate interpretation of the description of the concept and providing the word or probable words the user is looking for.

The core of such onomasiological dictionary, as reported by Sierra and McNaught (2000), is the lexical knowledge base (LKB), which should provide all the necessary knowledge to be manipulated in order to enable onomasiological search. In principle, it must represent what a person knows about both concepts and their corresponding terms. Such LKB consists then of a set of terms, a set of definitions for each term, a set of keywords associated with the definitions and a set of lexical paradigms that group keywords with the same meaning. It not only includes the databases that constitute these sets of data, but the interrelationships among all the sets.

References

Aitchison, Jean. 1994. *Words in the mind: an introduction to the mental lexicon*. Blackwell Publishers, Oxford.

Ayto, John R. 1983. "On specifying meaning: semantic analysis and dictionary definitions". *Lexicography: principles and practice*. R.R.K. Hartmann (ed). Academic Press, London: 89-98.

Baldinger, Kurt. 1980. *Semantic theory: towards a modern semantics*. Basil Blackwell, Oxford.

Bernstein, Theodore M. 1975. *Bernstein's reverse dictionary*. Routledge & Kegan Paul, London.

Fugman, Robert. 1993. *Subject analysis and indexing: theoretical foundation and practical advice*. INDEKS Verlag, Frankfurt.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. "Introduction to WordNet: An on-line lexical database". *International Journal of Lexicography*, 3(4): 235-244.

Petöfi, János S. 1982. "Exploration in semantics: analysis and representation of concept systems". *The Cocta Conference*. F.W. Riggs (ed).

Putnam, Hilary. 1975. *Mind, language and reality: philosophical papers*, Volume 2. Cambridge University Press, New York.

Roget, Peter. 1852. *Thesaurus of English Words and Phrases*. Longman, London.

Sager, Juan C 1990. *A practical course in terminology processing*. John Benjamins, Amsterdam.

Sierra, Gerardo. 2000. "The onomasiological dictionary: a gap in lexicography". *Proceedings of the Ninth EuralexInternational Congress.* Stuttgart.

Sierra, Gerardo and John McNaught. 2000. "Design of an Onomasiological Search System: a Concept-Oriented Tool for Terminology". *Terminology* 6(1).

Wiegand, Herbert E. 1984. "On the structure and contents of a general theory of lexicography". *Proceedings of the International Conference on Lexicography*. M. Niemeyer, Tübingen, 13-30.

Wierzbicka, Anna 1985. *Lexicography and conceptual analysis*. Karoma Publishers.