Are ontologies involved in natural language processing ?

Maryvonne ABRAHAM

Institut TELECOM TELECOM-Bretagne
Université Européenne de Bretagne
Laboratoire LaLICC
TELECOM Bretagne CS 83818 F29238 Brest cedex
Maryvonne.Abraham@telecom-bretagne.eu

Abstract

For certain disable persons unable to communicate, we present a palliative aid which consist of a virtual pictographic keyboard associated to a text processing from a pictographic scripture. Words and the grammar are given as pictograms. The pictographic lexicon must be organized following the mental lexicon of the user to propose the pictograms of grammar in order to facilitate his task of writing. We discuss the utility of ontologies in the organization of lexicons and in the building of texts.

Ontologies and Natural language

The problem

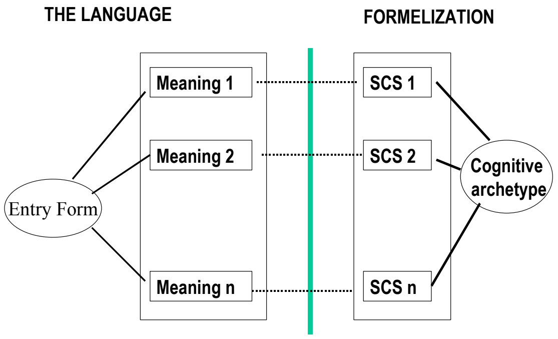
In computer science, ontologies are aimed to structure the concepts of a domain. In the language field, ontology will structure the concepts of a language.

The languages are systems that describe the world. Therefore, the question of the relationship between the language structures and ontologies structure have to be rise. There are several ways to structure a language: i) following two sets of units: the vocabulary and the grammar; ii) following an applicative and cognitive structure, where operations apply to operands; in the latter case, we must see how the language define operations and operands.

In both cases, the questions of the organization of the lexicon and the description of words arise. The role of grammar, that enables to build sentences, is defined in the first case by rules of grammar. In the second case, we consider semantic grammars, expressed by abstract operations and performed at the level of observable by a syntax, morphological changes, and certain words of the lexicon. The grammar of a language contains complex operations peculiar to that language, which can be recognized as the result of combinations of elementary operations[1]. Here is an overview of these operations for the French language.

Lexicon and ontologies

The lexicon of a language is a collection of words designating certain entities in the world. Our access to reality requires a construction of representations and interpretations from our perceptions. Everyone can easily notice that an entity or an idea can receive a name in one language and not in another one, although this entity or that idea seem to fall under the same concept.



To a same entry in the lexicon, several meanings are associated; they are described by SCS's.

Figure 1: Polysemy

It is in concepts level that ontologies are described. But in computer processing, we must describe these concepts. A concept refers to several words; Words are polysemous and therefore refer to several concept. This is not using words that we can describe ontologies, except if we reduce them to specific domains. This is known as domain ontologies. However, we expect that the systems of languages can describe all domains. The polysemy of words lead us to describe concepts by combinations of primitive designating most basic concepts. In order to describe concepts, we propose to use cognitive primitives that we use to describe the different meanings of words in the lexicon [Desclés, 1985,1990]. Ontologies can be linked with the language. through primitive. The words are polysemic, but one meaning of a word belongs to a semantic field, which itself can be described by using primitive. Within a semantic

[1] See the work of the team LaLIC, Paris IV Sorbonne, in particular [Desclés, 1990].

field, close meanings can be linked by networks of synonymy (figure 2). Several meanings are associated to a given word; these meanings are described by arrangements of structured primitives. These patterns are called semantico-cognitive scheme (SCS, figure 1).

We have defined different categories of primitives able to describe the meanings of words: primitives structuring, that contribute to organize the grammatical patterns, and empirical primitives, which are closer to perception, the manipulation and categorizations made by human beings.

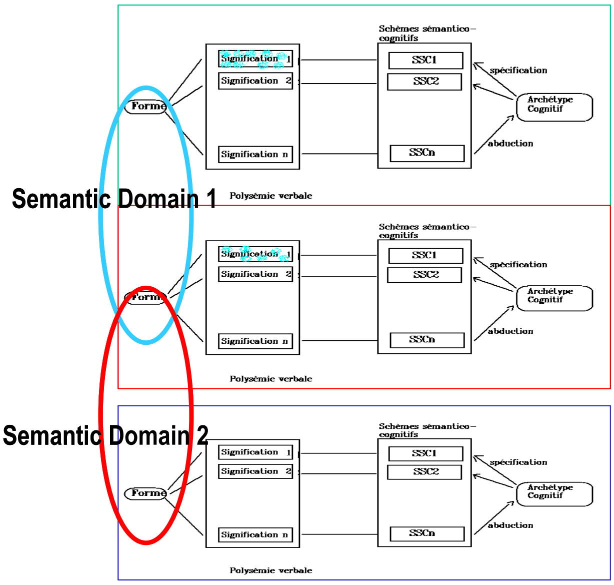


Figure 2: Polysemy an semantic domains.

Ontologies role in the production process of language by humans

An assistive application of TAL in speech disability

How can this description of semantics (words, concepts) find applications and evidence of cognitive matches in the NLP?

In a project aimed to help people without speech, without alphabetic writing and very severely disabled communication is established through pictograms. The idea came very quickly to build sentences from an ordered series of pictograms. This project raises a number of issues which we have responded in some articles [Abraham]: i) What is the status of pictograms if these icons are used to rebuild sentences?; ii) how can the lexicon be organized to be quickly accessible? If the only means of communication is established through the pictograms, several questions arise: How can they be arranged on a virtual keyboard so that the organization of this virtual keyboard match at best the mental lexicon of users? For a user, the writing process from pictograms is divided into several steps:

1. Find the words represented as pictograms on the screen
2. Organize them in the order they appear in the sentence
3. Apply grammatical operations which assign them a role in the sentence.

The three steps are usually carried out simultaneously in our minds. We have difficulties to advance evidence that our thoughts proceed in that order. What we can see, is the learning situation accompanied by a speech therapist. It is difficult to say whether we believe the situation in its time frame, and how words come to us to put a situation into words, knowing that the construction of the sentence has its own rules and that we fully integrated into our language. The problem is tantamount to best simulate our writing by giving organized lexicon to the user, and allowing him (her) to indicate the grammatical operations which are carried out by : i) places of words in the sentence; ii) morphological changes, results from grammatical operations that we believe semantic. The places of words can be handled smoothly by the user, but grammatical operations which are carried out by morphological changes must be indicated. Pictograms of grammar should be offered to the user in a separate category.

Theories of semantic applications based on ontologies

The theory of applicative and cognitive grammar (ACG, [Desclés, 1990]) analyzes the language as an operator / operand structure: different types of operators waits operand types: operators on names are distinct from operators of verbs or of adjectives. In a pictographic writing system of words, whatever the organization of the lexicon, lexical pictograms must bear the marking of their syntactic category since this category involves a series of possible operations on this icon.

The lexicon must also be organized in semantic fields. The pictograms are carriers of figurative representations of semes, as well as their syntactic categories. Two organizational approaches are possible, depending on the practice that we made with this new typewriter. At the first level, either we find several words in different syntactic categories dealing with the same semantic fields, or, toward a broader range, a syntactical level; under this syntactical level, lower levels are organized into micro-semantic fields. In both cases, the problem of polysemy arises: a polysemous word can belong several semantic categories, with different graphical representations, since images visually represent a designated entity. Such set of representations is not economical because it multiplies figurations of a single word. Moreover, problems arise if the word has abstract representations. One single representation of a good representative of the word should be a more economical solution, and perhaps easier to manage (in presentation and research processes) in a very large dictionary.

Domain ontologies

It must be noticed that the semantic categories does not cross syntactical categories, so, it makes more sense to divide the first level into syntactical categories; then, each syntactical category is divided into semantic domains, organized according to ontologies which rank the images of the world following human point of views. At the first level, broad areas of the world are found. In each sub-category, semantic entities share contextual relations of belonging to the same subdomain.

So, the lexicon is structured into ontology domains, in which each semantic subdomain (SDS) is linked to its higher domain by a relationship of inclusion.

For example, at the first level, we find the SDS names, adjectives, verbs, grammatical operations and sentences of emergency.

Then, in the SDS names is structured into :people, animals, artifacts, ... In a given SDS, the pictograms represent entities linked by a semantic relationship. The organization of these SDS is more or less empirical, constrained by the readability of icons on the screen: the more numerous ae icons on the same page, the more they are small; in this case, a lower level containing pictograms collected by a new criterion semantics is created. So the depth of trees depends on the number of icons placed in each SDS.

The vocabulary is built on entities of the world; they are represented figuratively. These entities are represent words that are polysemic, and can therefore designate entities belonging to other semantic subdomains. The lexicon is not presented in its entirety following an ontology for several reasons:

The image found in a semantic category is that of an object that represents a word. This image gets a double status, and depends on the processing it will receive: for the person who sees and selects it, this image represents a word, identified by the entity; Then, once selected, the picture becomes scripture of the word, and does not necessarily represent the entity which helped to find the word. The user who chooses his words switches from the world organization to that of language, from an ontological organization to a writing system. We must take account of the writing process of the user: it must find its words quickly: it looks for it as an entity of the world, kept in the category where it is the best representative of the entity designated by the word. He thinks the word, to find an image even if this image does not refer to the entity that identifies it. Once the image is found, it is only the scripture of the word, and corresponds often to another entity than that contained in the lexicon.

The lexicon contains only grammatical information associated with words, without semantic features. The semantic indications are given by the images of words to find the words. It is only at this level that ontologies are usefull.

The solution of including all entities, therefore deploying the polysemy, would have the following characteristics: all entities should be able to be represented figuratively in each of the semantic fields to which they belong. But a figurative scripture can not represent the whole lexicon, a symbolic system must be used. For example, BLISS describes concepts from a base of 26 elements and a law concatenation. If the lexicon is semantically structured, it is expected that the semantic indications should be used. Such information may be used as helps to word prediction in building of the sentence. But what is the support of these predictions? Several methods are available:

- A frequency of occurrence of words from one or more previous words (n-grams).
- A calculation from semantic compatibility of weighted semes, that is, use of ontologies.

In a general use of writing, we are not persuaded that it is helpful to offer this assistance. The discussion comes on domain ontologies, where it can be useful, but in the case of a language that crosses the fields, associations of words in different semantic fields seem too uncontrollable. More, in this case, it becomes necessary to introduce controls in order that the writer can verify that what is assisted written is truly that the user wanted to write.

Promote the writing process in the case of a pictographic writing

The organization of the lexicon should at the best match organization of entities of the world in our mind. Therefore, the problem is to build the ontological categories that refer to the words of the lexicon, with the restriction that they do not represent the different meanings of a word.

Building sentences: grammar application

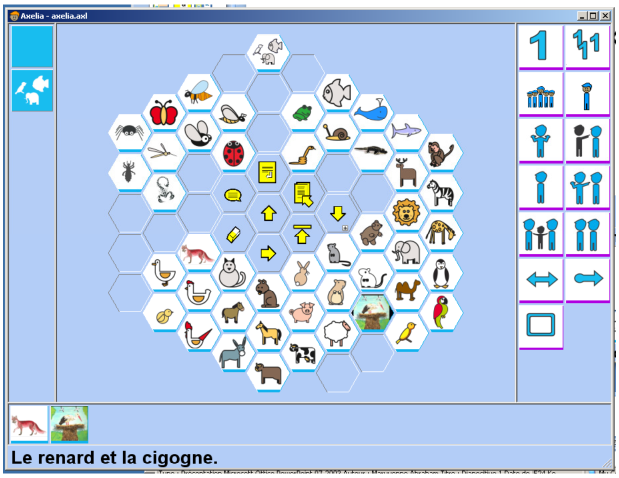
How grammatical operations on words can be represented? It is clear that these operations are not the same for all syntactical categories: a name cannot be conjugated, and so on ... Typing operators and operands obviously becomes necessary. One can wonder how far grammatical differentiation made by the syntactical categories denotes ontological differences. Syntactic categorizations are not universal and we can not find the same categories from one language to another. The large partition made by physics distinguishes what is stable and what is evolving. This distinction is found for example in the work of Wilkins that distinguishes sorts, which can classify the world, and particles that can be considered as operations on sorts.

Here, we will consider names, that roughly represent stable, and the verbs used to represent the evolving.

In the pictographic writing, the place of grammar symbols in the sentence arises: are they placed before or after the word on which they operate? That is to ask how we believe the operation on the name: prefixed or suffixed? It seems increasingly that we think about situations and that we know how to say them directly with the language, but for now the question remains.

When a user comes to write otherwise, everything depends on his ability to use the GUI. To save changes to windows, operations for a syntactic category are placed on the same page as pictograms category: passing on the pictogram of operation gives the result of the sentence in the lower window. For example, passing on the pictogram <PLURIEL> placed in a page of names, the name will be morphologically affected of a plural mark, preceded by an article indicating plural. Then this so pre-viewed operation can be selected.

The operations that include the names



Le renard et la cigogne.

Figure 3: the names and their operations

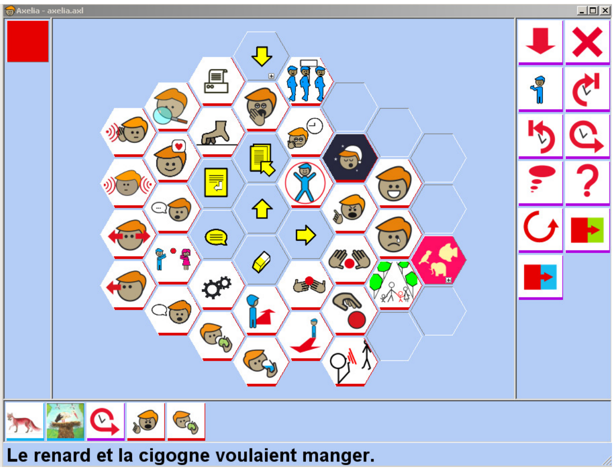
We give here a summary of information contained in the lexicon, which can link a word to an image. Important information for the construction of sentences are:
the word, its syntactic category, a semantic element (style) for grammatical purpose, and the image that allows the selection of the word.

```
<element mot="moi" type="PRONOM" style="LOCUTEUR" image="je.gif"/>
<element mot="toi" type="PRONOM" image="tu.gif"/>
<element mot="petite_fille" type="NOM" style="FEMININ" image="petit_fille.gif"/>
<element mot="Mamie" type="NOM" style="FEMININ|SANSARTICLE"/>
<element mot="moustique" type="NOM" image="moustique.gif"/>
<element mot="mouche" type="NOM" style="FEMININ" image="mouche.gif"/>
```

The operations which focus on verbs

The lexicon of verbs is organized in the same way as the lexicon of names: the word, its syntactic category, and the image that allows the selection of the word.

```
<element mot="exécuter" type="VERBE"  image="executer.gif"/>
```



Le renard et la cigogne voulaient manger.

Figure 4 : the verbs and their operations

How to categorize verbs of the lexicon? We have shown [Abraham, 1995] that, given the verbal polysemy, it were not the words of the lexicon which were to be categorized, but the concepts held in semantic fields. Thus, micro-semantic fields are proposed in the windows giving access to the lexicon of verbs: for example, in Figure 4, the red pictogram gives access to verbs which concerns animals specifically, being understood as a means to find them easily; but a person may, for example, rush in stretchers, even if the verb rush was found in this specific page concerning animals.

Conclusion

Ontologies represent a referential organization of the world, which allows us to have independent knowledge of languages, more or less shared. Languages cut worldwide, but the partition is not similar from one language to another. The role of language is to communicate meaning: from a shared knowledge of the world they have built, they allow to say anything other than the referential obviousness.
In the problem of disability that we seek to address, change writing and breaking up of the act of writing by at least three stages (Thinking the situation, finding the words, applying the grammar) shows that it is at the level of the organization of semantic domains of the words of lexicon dictionary words that ontologies find their place. In our problem, they give access to words, which are used as a new writing. This writing must be read in order to be interpreted and to access the situation constructed by language.

References

ABRAHAM Maryvonne**Communication pictogramme bidirectionnelle : du pictogramme au texte et**

inversement . Handicap 2008 : 5ème conférence, 10-12 juin, Paris, 2008

[2008a] Abraham M.Y., "Alteration in dialogical communication : the status of the language in the palliation of speech trouble", *ICCTA*, Damas, 2008

[2007a] ABRAHAM Maryvonne **Verbal polysemy in automatic annotation**. FLAIRS 2007 : 20th international conference, may 7-9, Key West, Florida, USA, 2007

[2006a] ABRAHAM Maryvonne **Altération de la communication dialogique : le statut de la langue dans la palliation des troubles de la parole** . Handicap 2006 : 4ème conférence "Nouvelles technologies au service de l'homme", 8-9 juin, Paris, France, 2006

[2005a] Abraham, MY, « représentation et structuration de la polysémie verbale – un exemple - » , 137 :154, La polysémie, sous la direction d'Olivier Soutet , PUPS, travaux de stylistique et de linguistique française : études linguistiques, Paris Sorbonne, Presses de l'Université, 2005, 1vol., 15p.

[2000c] Abraham, MY, *Journal Européen des Systèmes Automatisés* (JESA) vol 34, n°6-7, Handicap 2000 – *Assistance technique aux personnes handicapées* – « Reconstruction de phrases oralisées à partir d'une écriture pictographique », 883 :901, Hermès Sciences

Desclés, J.-P., 1987, "Réseaux sémantiques : la nature logique et linguistique des relateurs", *Langages* , n° 87, 1987, pp. 57-78.

Desclés, J.-P., 1990, Langages applicatifs, langues naturelles et cognition, Paris : Hermès.

Eco, U. *La recherche de la langue parfaite dans la culture européenne*, Seuil, 1994

Jackendoff, R. *Semantics and Cognition* . MIT Press, Cambridge, 1983.

Langacker, R. *Fondation of cognitive grammar*, vol 1. Standford University Press , 1987.

Pauchard J. Le Discours Psychanalytique - « la question de la représentation graphique au XVIIème siècle en Angleterre : du mot à la chose » *revue de l'Association Freudienne*, XVII : 249-270 fev 1997

Pustejovsky J. 1991 Pustejovsky (James). – *The generative lexicon.* Computational Linguistics, vol. 17, n° 4, 1991.

Pustejovsky, J. (1995) *The generative lexicon*. MIT Press, Cambridge, Ma.