A SET-THEORETIC APPROACH TO LEXICAL SEMANTICS

DOMINIQUE DUTOIT

MEMODATA
23 rue des Boutiques
14000 Caen, France.

RESUME

UNE APPROCHE ENSEMBLISTE DU SEMANTISME LEXICAL

Nous présentons les résultats d'un travail mené depuis plusieurs années par la société MEMODATA sur la structuration sémantique du lexique français. Ce travail a fait l'objet d'un premier contrat de recherche avec le Ministère de la Recherche et de la Technologie et aboutit aujourd'hui à un dictionnaire logiciel de plus de 100 000 mots et locutions regroupés analogiquement et synonymiquement (DICOLOGIQUE).

Le titre "dictionnaire analogique" regroupe des réalités très hétérogènes. Pour notre part, nous évitons le travail de Sisyphe portant sur les associations stéréotypées pour nous concentrer sur la structuration des champs sémantiques. Ainsi, nous classons, à l'aide de l'ordinateur, et selon l'expertise humaine, les milliers de faits linguistiques que les dictionnaires de langue s'attachent à recenser et à figer dans l'usage.

Une partie importante de l'article décrit d'une façon formelle le dictionnaire des champs lexicaux : il est une structuration de mots et d'ensembles de mots prenant l'aspect d'un graphe très fortement poly-hiérarchique. A la base de ce graphe, nous trouvons des ensembles "primitifs" : des ensembles qui n'ont pas de contenant autre que le dictionnaire tout entier. A l'autre extrémité se situent les mots définis par les ensembles auxquels ils appartiennent, ainsi que par l'ensemble de leurs successions d'appartenance. Dans DICOLOGIQUE, les successions d'appartenance des ensembles constituent différents niveaux d'une même "quasi-définition" d'un mot. Il existe 9 types d'ensembles :

- Quatre ensembles "Liste" composés de mots ayant une équivalence de sens et de catégorie grammaticale (nom, verbe, adjectif, adverbe).
- Un ensemble "Classe" destiné à recevoir des énumérations en extension (la zoologie par ex.).
- Un ensemble "Termes liés", au contenu assez hétérogène de termes n'ayant pu donner lieu à la création de listes dans un "Thème" donné.
- Un ensemble "Thème" capable d'énumérer tout le champ lexical d'une notion. Entre autre, il peut contenir des thèmes.
- Un ensemble "Description" employé en cas de nécessité définitoire.
- Un ensemble "Caractéristique" qui regroupe des mots dont les signifiés comportent un même trait saillant.

Les énoncés mathématiques correspondent à des fonctionnalités du dictionnaire électronique que nous avons illustré par des exemples empruntés à celui-ci :

- Croisement de concepts (verbes exprimant "faire tomber" et "couper", substantifs désignant une "coiffure" du "Pape", synonymes de "voler" pour des "abeilles"...).
- Edition de listes ou de thèmes (la liste "Penser" éditera 500 verbes structurés, synonymes potentiels.
- Recherche des quasi-définitions d'un mot (le terme polysémique "abattre" est présenté dans l'article).

Avec ces différents exemples, nous comprenons comment le dictionnaire des champs lexicaux peut être consulté utilement par un utilisateur humain. Nous posons alors le problème de son exploitation par la machine elle-même.

En reprenant les définitions mathématiques du dictionnaire, nous travaillons sur l'indexation automatique des thèmes d'un petit texte paru dans la presse. Nous développons une stratégie d'analyse strictement lexicale. Elle est basée sur une détermination des ensembles capables de cerner les sujets abordés par l'énoncé. En définitive, le dictionnaire peut travailler sur de petits textes (nous n'intégrons aucune syntaxe) en-dehors de toute démarche d'ingénierie préparatoire.

Actuellement, DICOLOGIQUE contient plus de 100 000 mots et 15 000 ensembles typés. Les 350 000 observations d'appartenance directes des mots, créées par un expert humain, développent un graphe de 4 000 000 de successions d'héritage que nous améliorons sans cesse. Les outils de base que nous construisons, tel le SEMIOGRAPHE pour la recherche documentaire, nous permettent d'évaluer la progression de la qualité des interprétations que nous obtenons.

Nous concluons notre article par notre souhait de rencontrer, lors du COLING, des partenaires, français ou étrangers, qui voudraient avec nous échanger des travaux sur ces questions.

A SET-THEORETIC APPROACH TO LEXICAL SEMANTICS

D.DUTOIT, MEMODATA.

INTRODUCTION

We present the results of the work carried out over several years by the Memodata Company on the structure of the French lexicon. This work has been accomplished thanks to a first research contract with the Ministry of Research and Technology and today has lead to a dictionary of more than 100 000 words and phrases grouped analogically and synonymously.

If we understand quite well how a dictionary like this can be used with ease by humans, we set the problem of the identification of meaning by a computer. We will evaluate how Dicologique adds information complentary to the information contained in semantic nets. Thanks to a somewhat unusual construction method and the systematic classification of words according to their meaning, we are progressing to a continuous system of localisation of the meaning itself. On the map we created, it is possible to compute the meaning due to lexical semantics for any sentence written in natural language...

1) GENERAL PURPOSE OF ANALOGICAL DICTIONARIES

The purpose of dictionaries grouped under the name "analogical" is always the ease of the passage from a word to an idea and the inverse passage from an idea to a word. This aim is reached by the make up of lists of stereotyped associations and of semantic fields. The first approach does not have the same likelihood of ending with a satisfactory result as the second.

The stereotyped associations depend on the idea of time, the background and the experience of each individual. Their record can only be a track of the associative memory from the individual.

On the other hand, the dictionary of semantic fields is perfectly workable at any time ; it is based on the linguistic conventions that the language dictionaries have tried for centuries to record and to normalize.

It is not possible, by definition, to construct a dictionary made up of stereotyped associations, whereas it is possible to work on the complexity of hundreds and thousands of linguistic facts which we have classified.

We will give a mathematical description of the dictionary of the semantic fields. This approach is in parallel with concrete examples derived from the database.

2) DESCRIPTION AND UTILISATION OF THE DICTIONARY

The dictionary of analogies and synonyms that has been set up is a structure of sets and words which the conceptual figure (1) shows. The objects "words" (W_u) are represented in the rectangles and the objects "sets" (C_i) in the parallelograms.

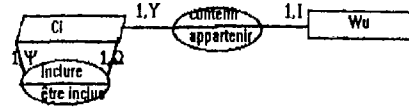

Fig 1: **G,** *the conceptual model of the dictionary*

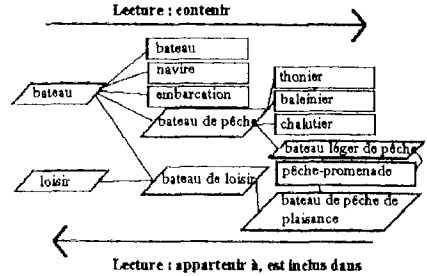

Fig 2: Example of a microstructure

2.1) MOVING FROM THE LEFT TO THE RIGHT

When moving through Dicologique from the left to the right, we move from the general to the specific.

a) Definitions linked to the motion from the left to the right

* Suppose we take C_j, a set contained in C_i.

The function $P(C_i, C_j)$ measures the depth of the inclusion.

Example :

P(bateau [1], bateau de pêche [2]) = 1

P(bateau, bateau) = 0

[1] bateau : boat

[2] bateau de pêche : fishing boat

* Set C_i can contain 1 to Ψ sets C_j (j goes from 1 to Ψ, Ψ is the number of subsets children of C_i).

* Suppose we take $C(C_i)$ the set containing all the direct subsets, or children, of C_i.

$C(C_i) = \{C_j$, such that $0 <= j <= \Psi$ and $P(C_i, C_j) = 1\}$

Example :

C(bateau) = {(bateau de pêche), (bateau de loisir[3])}

* A set C_i contains 1 to Y terminal words W_u (u goes from 1 to Y, Y is the number of terminal words children of C_i).

* Suppose we take $A(C_i)$ the set of words directly children of C_i.

$A(C_i) = \{W_u$ such that $1 <= u <= Y\}$

Example :

A(bateau) = {bateau, navire[4], embarcation[5]}

In our example, function A produces the synonyms for boat. Because of their position on the graph we will often consider them as lexical prototypes.

* Suppose we take a function $M(C_i)$. $M(C_i)$ counts the number of words in a set C_i. $(M(C_i) = \omega)$

* Suppose we take $U(C_i)$, the contents of the lexical field of C_i, i.e. the set of words contained in C_i.

$U(C_i) = \{W_u$, with $1 <= u <= \omega$ and such that it exists a set C_j, containing W_u and such that $P(C_i, C_j) >= 0)\}$

Example :

U(bateau) = {bateau, navire, embarcation, thonier[6], baleinier[7], chalutier[8], pêche-promenade[9] (day boat), U(bateau léger de pêche[10]), U(bateau de pêche de plaisance)[11]}.

This function allows to edit, with their structure or without, 1444 verbs currently contained in the set "changer" (to change/to alter).

Comment : according to our example, the function U(loisir[12]) would provide a result quite different from the actual dictionary Dicologique. In fact "leisure" is a structure with several thousands of words and several levels of intermediate sets we have not shown in figure (2).

b) Property of the graph derived from these definitions

The existance of the function $M(C_i)$ for all sets C_i infers that the structures of inclusion are without loops, i.e. there are sets which are not contained in any other set but the set of the graph G (root node) itself.

Sets only contained in the set of the graph G are named "primitives".

c) Semantic and grammatical characteristics of the sets and words

In Dicologique there are 9 types of sets :

- four types of sets named "lists".

They give the quasi synonyms, i.e. lists composed of words which are equivalent in meaning and identical in grammar. We have the following types of grammatical sets : noun, verb, adjective, adverb.

- the set named "class"

A set of this type contains nouns which can be subsumed under the same concept.

In our example in figure (2), "bateau" is a set containing on the one hand words which represent its prototype values ("bateau", "navire", "embarcation") and on the other "classes" of specific boats.

- the set named "related words"

Generally, the contents and utilizations of this type of set are rather various. We need it, for example, to represent the link between "baleinier" and "baleine"[13], which is not shown in figure 2 so as not to weigh down the graph.

- the set named "theme"

This set contains all the concepts and words associated in a particular semantic field. It may also contain other sets such as "related words" or smaller "themes".

- the set named "description"

It contains the constituents organically connected to a concept. It is only used when absolutely necessary for a definition.

- the set named "characteristics"

It subsumes words having the same outstanding feature. For example, our set of class "bateau léger de pêche" could be found under a set characterized by the feature "small" which differenciates this class from other classes of boats.

As for the words, we have provided them with the usual characteristics, i. e. their morphological classes (grammar) and their usage labels (colloquial, archaic, literary ...) which contain the

[3] bateau de loisir : pleasure boat
[4] navire : vessel
[5] embarcation : other synonym of boat
[6] thonier : tunny-fishing vessel
[7] baleinier : whaleboat
[8] chalutier : trawler
[9] pêche promenade : day boat
[10] bateau léger de pêche : small fishing boat
[11] bateau de pêche de plaisance : pleasure boat for fishing
[12] loisir : leisure
[13] baleine : whale

usual information associated to each word in every dictionary.

d) Use of the previous definitions in Dicologique

Moving through the dictionary from the general to the particular is a process widely put into practice by users who may either search a precise term to be discovered by intersection of associated concepts or intend to edit a lexical field or a classified list.

Intersection of concepts

The logic of the sets takes into account the logical "and", "or" and negation. Here are some examples of searches which are always based on the intersection of sets edited by the function U(Cj) :

- Search of a precise term

Search of the name of the "coiffure du Pape[14]". The intersection of "Pape[15]" (theme of 162 words) and "chapeau[16]" (list of 180 words) or "couronne[17]" (theme of 33 words) produces the words "tiare[18]", "calotte[19]" etc. in 10 seconds on the micro computer.

- Search of words to express an idea

Search of the verb to express the idea of "faire tomber[20]" and "couper[21]". The intersection of the two lists of corresponding verbs converges on about 20 words (abattre, décapiter, étêter, ébrancher ...).

- Search of synonyms according to a context

The synonym of "voler[22]" such that the meaning is more suitable for a bee. The words "butiner[23]" and "voltiger[24]" are immediately produced.

Edition of lexical fields

It principally has two aims :

- to search among very wide lists for the terms which help to get a precise idea.

For example, the list of verbs "penser[25]" contains about 500 structured verbs that allow to move continously through the whole field concerned.

- to have a precise idea about a structure.

This is especially interesting for the sets containing predetermined taxonomies.
The edition of the set of class "animals" presents the scientific taxonomy of the animal world. About 4100 indexed animals can be visualized in a structure of 500 classes.

2.2 MOVING FROM THE RIGHT TO THE LEFT

This is the opposite of the previous work. It corresponds to moving from the particular to the general.

a) Definitions related to the motion from the right to the left

A set named C_j may be included in 1 to Ω sets C_i (i going from 1 to Ω, Ω being the number of "parents" (main sets) of C_j).
This function permits, therefore, to move upward (towards the root node) in the structure of the sets.
The Ω classes C_i constitute the "quasi definition" of C_j.
Example :
In figure (2), the set "bateau de pêche de plaisance" is defined by the set of sets h = {{bateau de pêche}, {bateau de plaisance}}.
A terminal word W_u can belong to 1 to I sets C_i (i going from 1 to I, I being the number of "parents" (main sets) of W_u).
In Dicologique the direct questioning of a word gives, as in all dictionaries, the "(quasi)-definition" of the word.

The I classes C_i give the "quasi definition" of the term W_u.

Example :
In figure (2), the search of "baleinier" gives the list "bateau de pêche" and the set of linked words "baleine".

b) Properties of the G graph we defined thus

It exists for every non primitive object of G, 1 to Ξ series of connections which link it to one of these primitives.

All series of connections for an object taken together constitute the inheritance H of this object.

c) Use of the previous definitions in Dicologique

The table (3) represents the result of a part of the search of the polysemous word "abattre". We have limited the reproduction of the result to the polysemous zone only. The left column shows the sets containing "abattre" directly. The column in the middle, the type of set concerned. The right column shows the number of elements in the set concerned.
This result is produced by the computer immediately.

---

[14] coiffure du Pape : Pope's headgear
[15] Pape : Pope
[16] chapeau : hat
[17] couronne : crown
[18] tiare : tiara
[19] calotte : scullcap
[20] faire tomber : to knock over
[21] couper : to cut
[22] voler : to fly
[23] butiner : to gather pollen
[24] voltiger : to fly about
[25] penser : to think

Name of the set	Type of C	M(Cj)
Sens 1		
Couper[26]	List (L)	81
Couper un arbre[27]	L	13
Détruire[28]	L	262
Faire tomber[20]	L	52
Sens 2		
Coucher[29]	L	6
Faire tomber[20]	L	52
Fort[30]	Caract	23
Sens 3		
Bétail[31]	Rel. Words	24
Couper[26]	L	81
Faire tomber[20]	L	52
Tuer[32]	L	56
Sens 4		
Avion[33]	Rel. Words	18
Faire tomber[20]	L	52
Détruire[28]	L	262

Figure3 : Recherche du mot "abattre"

In the first place the elements of the above table lead to the following searches which correspond to moving from the left to the right :

* Search of synonyms of "abattre" with the meaning of "détruire".

The edition of the structured set "détruire" (L) produces the 262 verbs which constitute the set "détruire" in alphabetical order.

* Search of synonyms of "abattre" with the meaning of "couper" and "faire tomber".

We apply the logical function "AND" to these sets of verbs and about 20 verbs are produced. 7 to 8 verbs will be left if we add the list "tuer" as a supplementary constraint. The processing takes 4 to 5 seconds.

Moving from the particular to the general it is possible to enlarge the idea of "détruire" to the sets which include it.

The structured edition of 1400 words contained in the "changer" list takes less than a minute on the micro computer.

The motion from the particular to the general offers much fewer functions than the inverse motion. At the very most it is used to locate. Often the consultation of Dicologique is motivated by the search of synonyms. The edition of the contents of the terminal nodes of figure 3 appears to be largely enough for human users.

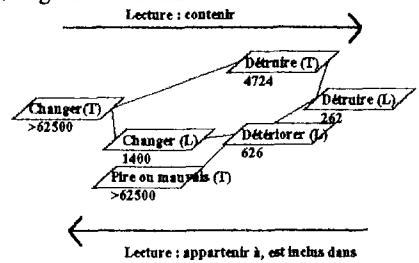

Figure 4 : a part of the conceptual environment of "détruire" (L)

But the position of the computer is very different : if humans possess the structures necessary for interpreting the terms, i. e. the linguistic heritage and the knowledge of the world, the computer for its part possesses neither of them. This is why we try to supply it with the lexical knowledge absolutely necessary in a coherent system.

Obviously this knowledge is situated in the inheritance from the "primitives" to the words.

Let's see an example in which the motion from the right to the left is applied to a problem of automatic indexing in information retrieval.

2.3 An application to information retrieval

We want the system to retrieve the lexical key elements of the following small piece of text :

"The accident, on Friday, took place in foggy weather. The two cars that crashed into each other caused a pile up of about 50 vehicles on the congested national dual carriageway."

Strategy for resolution :

The strictly lexical analysis of this short passage will be based on the calculation of the surface corresponding to each concept (set) covering one or more words of this document : the more abstract a concept, the bigger its surface.

For every set Cj we have defined its cardinality M(Cj) and its depth in comparison with the primitive P(Ci, Cj), with Ci as a primitive and considering a specific passage from the general to the particular.

Suppose we take Max P(Ci, Cj) the maximum depth of the graph whatever j might be. We know, from experience, that this maximum depth is attained in detailed enumerations of words which have very precise meanings and designate concrete things.

[26] couper : to cut
[27] couper un arbre : to cut down a tree
[28] détruire : to destroy
[29] coucher : to lay down
[30] fort : strong
[31] bétail : beast
[32] tuer : to kill
[33] avion : aeroplane

We define the function (Cj) to measure the distance between the concept Cj and the concept that maximizes P(Ci,Cj).
The surface of a concept is given by the formula as follows :

$$S(C_{jh}) = \Delta(C_{jh}) * M(C_j)$$

where h specifies the series of connections h for the calculation of the surface of C_j.

Notes :

1) Each concept possesses h evaluations of its surface accounting for its h series of connections.
2) In reality $\Delta(Cj)$ takes into account a complementary piece of information : the semantic characteristics of sets. A set of the type "list" introduces a more stretched arc than a set of the types "theme" or "related words".
3) If Dicologique is a general dictionary capable of resolving problems of information retrieval referring to general language, it is very easy to adjust it to a precise problem (for example, the thesaurus of a specific undertaking). One only has to stretch each arc situated on the passage of the series of connections of each term of the thesaurus.
The surfaces S, which are situated in a norm reference, describe a map of meaning on which all the continuous calculations of Euclidian geometry are made possible.
To resolve our problem, it is possible to keep the mathematical expressions very simple : a simple mean.
Each word of the document is recognized in the dictionary as it activates all the sets containing it according to their specific weights $S(C_j)$ which depend on their series of connections.
Finally, each set will have been activated k times.
The analysis will take into account as the most relevant set the one which presents the smallest ratio of $S(C_j)/k$.
$S(C_j)/k$ measures the weight of the concept in the text.
Our dictionary produces the beginning of a hierarchy in accordance with the conceptual sets which summarize the text given in the example :
1° : car ; 2° : accident ; 3° : road.
The other sets have negligible weights.
The complete analysis (but useless) takes 5 minutes on a compatible computer.
We have finished with the description of the results of our work on the semantic structure of the lexicon. We think it might be interesting if we add a description of the method we use for constructing the map of meaning we have presented above.

3) METHOD OF STRUCTURING AND EXISTING CONTENTS OF THE DICTIONARY

It is difficult to present completely the method of structuring we use as we lack of the supplementary page this would involve. To put it simply, all of the 100 000 words and current phrases, the 15 000 typified sets have been created manually and are continually worked over again, thanks to a structuring tool specially developed for this purpose.
The 350 000 observations of direct connections of words, the 4 000 000 series of inheritance which we run currently are, proportional to our efforts, increasing daily and of better quality.
The point at issue in this iterative procedure is, of course, to make appear in Dicologique tendencies towards sets of "primitives" in the linguistic sense. The number and articulation of these primitives should be both sufficient and necessary to define the objects depending on them.

FURTHER PROJECTS FOR THE FIRM

Our will is to continue working on the structure of the lexicon, as to the quantity of vocabulary represented of course, but principally with regard to the quality of the structure of the lexicon.
Moreover, we have been developing basic tools such as the Semiographe for information retrieval which allow us to evaluate the progression of the quality of interpretations we obtain. If we hope to find partners during COLING 92 who want to join us in our current research projects, we are also very interested in meeting people from foreign universities and firms who would like to launch a version of Dicologique in another language.

REFERENCES

Rouaix, P (1897) "Dictionnaire des idées suggérées par les mots", Armand Colin.

Martin, E (1990) "L'exploration textuelle assistée par ordinateur ; l'interrogation thématique", Coloquio de Lexicologia e lexicografia, Universidade Nova de Lisboa.

Rastier, F (1987) "Sémantique Interprétative", Paris, P.U.F.

Dutoit, D (1991) "Un nouveau dictionnaire de la langue française", La banque des mots, CILF.

Herr, P (1991) "Les dictionnaires électroniques : quelles caractéristiques pour quelles objectifs", La tribune internationale des langues vivantes.