Proceedings of the International Joint Conference IBERAMIA/SBIA/SBRN 2006 - 4th Workshop in Information and Human Language Technology (TIL'2006), Ribeirão Preto, Brazil, October 23–28, 2006. CD-ROM. ISBN 85-87837-11-7

A geometric approach to meaning computation: automatic disambiguation of French adjectives.

Fabienne Venant.

Laboratoire LALIC, ISHA, Université Paris IV- Sorbonne.
fabienne.venant@ens.fr

Abstract: This paper presents a new kind of model for meaning construction within the framework of continuous mathematics. Language is considered as a morphodynamic system following the basic principles of Gestalttheorie. In our model, linguistics units acquire meaning in a semantic space. The validity of this model is tested by its implementation on a French adjectival lexicon. Our dynamic method for meaning computation allows us to take the different factors of adjectival polysemy into account.

1 Introduction

First of all, we want to understand the processes of meaning comprehension. How do words like *vie (life), vue (vision), volant (steering wheel),* whose meanings are inert when they are alone, become alive in an utterance like *au volant la vue c'est la vie (when you drive, to see is to keep alive)*? Of course each word can have a lot of different meanings, depending on context. This phenomenon -called polysemy- is constitutive of language, and is the basis of its richness. However it is quite difficult to formalize. In most models of language polysemy is considered as a kind of artefact. In these models, polysemy amounts to very little: a choice in a list of pre-existing meanings. The omnipresence of polysemy always leads this kind of computation to combinatorial explosions. To avoid this problem, we want to give a central place in meaning construction to polysemy: that is why we define our model within the framework of continuous mathematics. This model was first proposed by Victorri [4]. It is deeply rooted in Gestalttheorie. Each linguistic unit is associated with a semantic space, where its different meanings are organised according to semantic proximity. The other units of the utterance define a potential function, which allows us to determinate the region of the semantic space corresponding to the meaning of the unit studied within the utterance studied.
Our basic result is that this model is computationally implementable with great success. Two mains goals are therefore aimed at: firstly, to develop general tools for meaning computation. Secondly, to account for theoretical studies of French adjectives.

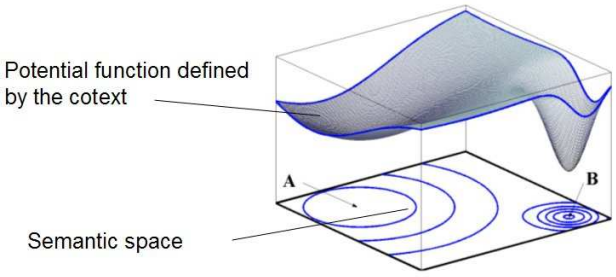

Fig. 1: Model

2 French adjectives

We choose to work on the French adjectival lexicon, because it is the subject of many works in linguistics. This literature enables us to form a very precise idea of the phenomenon we want to study. Very few people work on automatic adjective sense disambiguation, at least in French [2]. Therefore the French adjectival lexicon constitutes an ideal domain for experimental research. It is almost unexplored from a computational viewpoint, although very well described from a linguistic viewpoint [5]. We so drew out from the literature the most specific factors of French adjectival polysemy in order to test how the model can process them.
We worked only on French attributive adjectives and on two main factors:

- An attributive adjective (like *sec* in *un homme sec-a thin man*) is always linked to a noun. It is this noun which mostly constrains the meaning of this adjective, even if other units like the article can play a role. We first studied the influence of the noun, through the analysis of the French adjective *sec (dry, severe, brusque…).*
- The meaning of a French attributive adjective can also change a lot with word order. In French, if an adjective is placed before or after the noun it can sometimes make a subtle difference. But there are many cases in which the place of the adjective plays a significant role in the construction of meaning. We studied this phenomenon, and how our system can bring it out, through the analysis of the French adjective *méchant* (*nasty, naughty, wicked, bad*…)

3 Semantic space

In order to compute the meaning of an adjective in a given sentence, the first step consists in associating each adjective with a semantic space. The method will be illustrated by a visualization of the semantic space associated with the French adjective *sec*. Let's take a look at its semantic. *Sec* is a very polysemic adjective, but according to the French dictionaries we can group its meaning in six main areas:

1) Lacking water*: sable sec (dry sand).*
2) thin, bony: *un homme grand et sec* (*a tall and thin man*).
3) sterile, unproductive: *rester sec aux questions du professeur* (*cannot answer the teacher's questions).*
4) lacking sensitivity, egoist *: avoir le cœur sec* (*to have dry heart*).
5) brief, abrupt, lacking sweetness : *coup sec* (*blunt blow*).
6) alone *: atout sec* (in playing cards, *a singleton trumps*).
The semantic space is built though the algorithm proposed by Ploux and Victorri [7]. This algorithm relies on the analysis of a graph of synonymy. This graph is provided by the Dictionnaire Electronique des Synonymes (DES, www.crisco.unicaen.fr).
For example, for the adjective *sec* the DES gives a list of 63 synonyms, and builds their graph: two adjectives are in relationship if their synonymy is certified by the dictionary. Fig. 2 shows an excerpt of the graph of synonymy of *sec*.

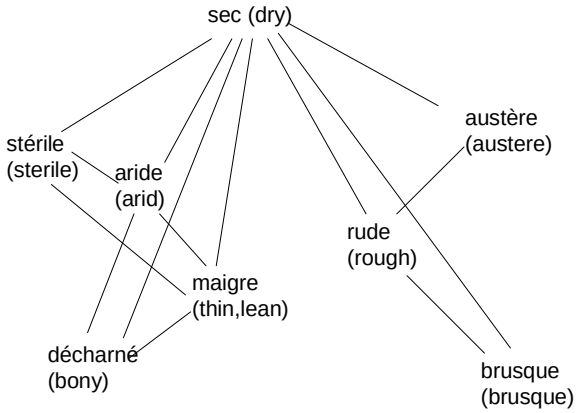

Fig. 2. An excerpt of the graph of synonymy of sec

The underlying idea is that only one synonym is generally not enough to define a meaning of the unit studied. We can see in Fig. 2 that *maigre* is at the same time synonym of *aride* and of *décharné*, which correspond to two different meanings of *sec*. In order to build the semantic space, we use cliques of the graph. A clique in a graph is a maximal set of pairwise adjacent vertices, or -in other words- an induced subgraph which is a maximal complete graph. In the graph on Fig. 2, there are four cliques: <*Aride-décharné-maigre-sec*>; <*Aride-maigre-stérile-sec*>; <*Austère-rude-sec*> and <*Brusque-rude-sec*>. We can consider as a first approximation that a clique corresponds to a precise meaning of the word studied. Each clique corresponds to a point in the semantic space. (cf [7]). When considering the subgraph formed by a word and all its synonyms, the cliques of this subgraph constitute overlapping parts that cover all of the set of the synonyms of the word studied. Each synonym of the word studied belongs at least at one clique, and vice versa. The set of cliques associated with a lexical unit thus characterizes the structure of the set of synonym of

this word. We then define the semantic space as the Euclidian space generated by the synonyms of the lexical unit studied:
Let u_1, u_2… u_n denote the synonyms, and c_1, c_2… c_p the cliques. U_1, u_2… u_n correspond to the coordinated axes of the Euclidian space. In this space, the clique c_k corresponds to a point whose coordinate relatively to u_j is x_{ki} and

$$x_{ki} = 1 \text{ if } u_i \in c_k \text{ and } x_{ki} = 0 \text{ si } u_i \notin c_k \quad \textbf{(1)}$$

Ploux and Victorri showed that the canonical Euclidian distance does not work in this space. This distance doesn't account for real semantic proximity because it gives the same weight to all the cliques and all the synonyms. Ploux and Victorri proposed to use the chi-square distance:

$$d^2(c_k, c_l) = \sum_{i=1}^{n} \frac{x}{x_{\bullet i}} \left(\frac{x_{ki}}{x_{k\bullet}} - \frac{x_{li}}{x_{l\bullet}} \right)^2 \quad \textbf{(2)}$$

This distance works better because each synonym is balanced according to the number of cliques it belongs to, and each clique according to the number of synonyms it contains. The more a synonym belongs to different cliques, the less it is specific, the less the role it plays in meaning discrimination is important. The more a clique contains non specific synonyms, the nearer its corresponding point to the origin of the space. We developed Visusyn a software which can automatically build the semantic space and propose a 2D visualization of the cloud of cliques using a Principal Component Analysis [1]. Fig. 3 shows the semantic space associated with *sec*. It accounts for the six main zones previously presented.

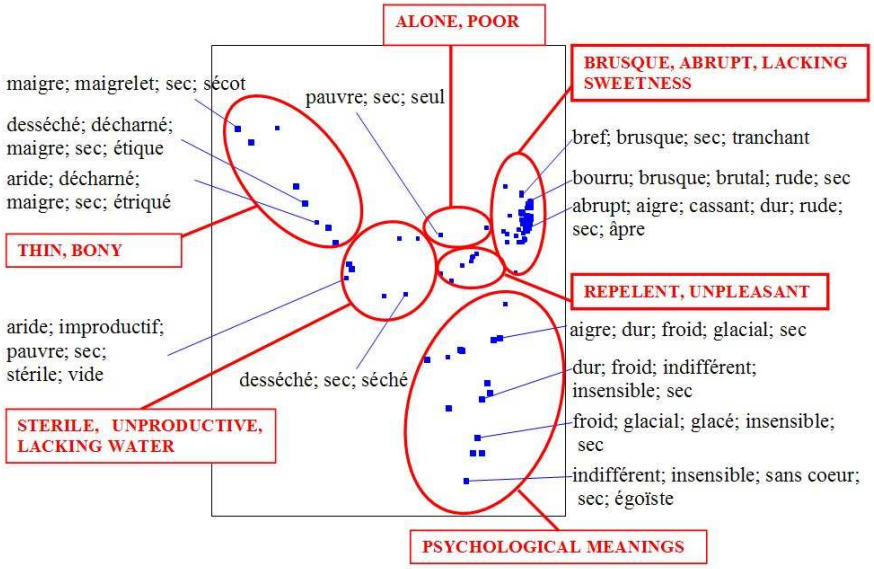

Fig 3. Semantic space associated with *sec*

4 Influence of the noun

In a first work [8], we disambiguated *sec* in context by automatically finding which synonyms matched better the meaning of *sec* when used with a given noun.
In order to do this, we associated a potential function with each synonym. Our tool Visusyn can compute the value of the function in a given point by seeing whether the synonym belongs to the corresponding clique or not. It can also propose 2D or 3D visualizations of these functions. The basins of a function represent the meaning zones of the semantic space in which the synonymy between the word and the given synonym is relevant (Fig. 4).
We then selected 20 nouns among the most frequently used with *sec* in a large corpus (Frantext, http://atilf.atilf.fr/frantext.htm). We associated a potential function to each noun. Visusyn can use data (frequencies of cooccurrences) from the corpus to compute an affinity degree between a noun and a given clique. It thus uses these degrees to compute the potential function associated with the noun. The value of the function in each point depends on the affinity degree between the noun and the corresponding clique. The basins of the function determine the zones of the semantic space corresponding to the meaning of *sec* when used with the associated noun.

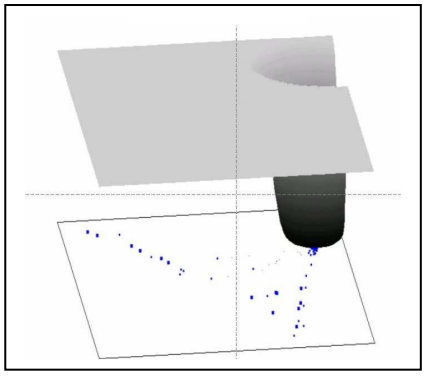

Fig. 4. Potential function associated with the adjective *brusque*

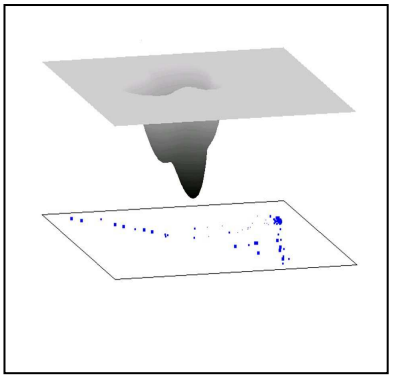

Fig. 5. Potential function associated with the noun *fleur* (*flower*)

Fig. 5 shows the potential function associated with *fleur (flower)*. In this case the noun *fleur* forces *sec* to take a precise meaning in the zone 'lack of water'.
The method of disambiguation consists in comparing the function of a synonym to the function of the noun under focus. The more the functions overlap, the more you can replace *sec* with its synonym without changing the meaning of the syntagm.
We compute the overlap rate for the 20 nouns and the five adjectives selected. We then compare the results automatically processed with the human answers on the same task, which is to select among 5 synonyms of *sec* those which better describe the meaning of *sec*, when used with a given noun. The rate of success is 79% (for more details on the tools used, see [8]). This result encourages going further. By analysing errors, we found out several ways of improving our method.

Firstly, on one hand the partition of the semantic space with only five synonyms is not sufficient. But on the other hand we can not compute the potential function for each synonym of *sec* because of the computation time. So we divide (by hand) the semantic space in 6 zones corresponding more or less to the 6 main meanings previously determined. We now want our system to decide automatically which of these six zones correspond to the meaning of *sec* used with a given noun. The method is similar to the previous one: we now associate each zone and not only each synonym with a potential function. This function is computed according to the cliques belonging to the zone.
We also use a bigger corpus, that is to say the French newspaper *Le Monde* for ten years. We work on the fifty most frequent nouns used with *sec*. For each noun we compare its potential function with that of each meaning zone. We thus compute an affinity rate between each noun and each zone. This method gives better results than the previous one but also generates a high rate of no answer. Those silences concern nouns whose frequencies of use with *sec* and its synonyms are very low.

5 Using distributional classes

In order to assign a meaning to *sec* even when it is used with a scarcely used noun, we want to associate a potential function not to a single noun, but to a class of nouns. The idea is to follow the principle of distributional analysis which says that semantic can emerge from syntax. We thus put together the nouns which share the same lexico-syntatic contexts within the same class. A lexico-syntatic context is for example 'to have *sec* as an attributive adjective'. We want to extract from the corpus classes of nouns having the same influence on the meaning of *sec*. We want for example to automatically associate the class [*bruit (sound), coup (blow)*], with the zone 'lack of sweetness' or the class [*fruit (fruit), haricot (bean), legume (vegetables)*] with the zone 'lack of water'. To build these classes, we use the method described by Jacquet and Venant [6]. We use the output of the parser SYNTEX [3] to build a distributional space associated with the corpus. This space is the Euclidian space generated by all the lexico-syntactic contexts. To find the class of a given name in this distributional space, we look among the neighbours of this noun for those also used with *sec* as an attributive adjective. We then select the nearest (using the chi-square distance). Fig. 6 shows that the class computed for a given noun changes with the adjective under study. C*oup* used with the attribute adjective *sec* is associated with *bruit* and *geste* (*gesture*). That class enables the system to compute the correct meaning for the adjective *sec* in the syntagm *coup sec*. The class associated with *coup* when used with the attribute adjective *audacieux (daring, bold)* is not the same that the previous one. Each class enables to compute the correct meaning of the adjective in the syntagm under study.

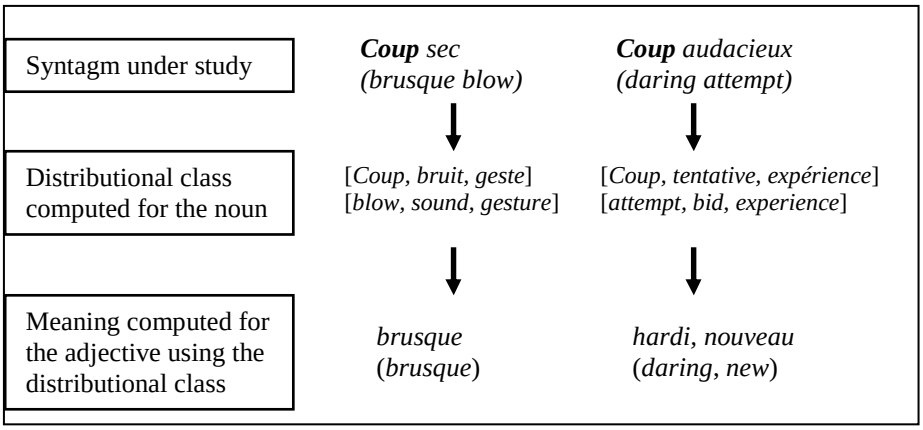

Fig. 6 : distributional classes associated with the noun *coup*.

We compute these distributional classes for each noun and use them in our disambiguation instead of the name itself. The results are far better than using the noun alone. Table 1. shows some results.

Table 1.

Noun	Meaning zones	Affinity
Corps (body)	Maigre, décharné (thin, bony)	93%
humour	Sens psychologiques (psycholgical meanings) Manque de douceur (Lack of sweetness) manque d'eau, improductif (Lacking water, unproductive)	25% 22% 18%
cheveu (hair)	manque d'eau, improductif (Lacking water, unproductive)	77%
Son (sound)	Manque de douceur (Lack of sweetness) Sens psychologiques (psycholgical meanings)	42% 33%
Été (summer)	Sens psychologiques(	73%
Hiver (winter)	Sens psychologiques (psycholgical meanings)	100%

The first comment we can make is that if the meaning computation was adequate with the noun alone, it is still adequate when we replace the name with its distributional class. Some subtle can even appear. For example the computation of the noun *humour* alone only selected the zone 'psychological meanings'. Using the distributional class we can see that *un humour sec* is also a humour lacking sweetness.

The uses of distributional classes also gives results for 14 of the 20 nouns like *son* or *cheveu* for which no automatic answer was given in the previous calculus but there are still some errors like for *été* or *hiver*. The problem here is that these words cooccurre a lot with *froid* (*cold*) or *glacial*, *glacé (icy)*. Nevertheless, *été froid (cold summer)* is not synonym of *été sec (dry summer)*. In this case, the system fails because *froid*, *glacial*, and *glacé* share a lot of cliques. These cliques correspond to psychological meanings of *sec*. They should not be taken into account when computing the meaning of *été sec* but unfortunatly they do because of the high frequencies of coocurrences. Our method finds some strong chalenges here.

6 Influence of word order

Through the analysis of the behaviour of the French adjective *méchant (wicked)*, we can study the changes in the meaning due to the place of the adjective. *Méchant* is very interesting for this kind of studies. *Méchant* is very frequent and its meaning can change depending on its place in the syntagm. For example *un méchant écrivain* is a mediocre writer whereas *un écrivain méchant* may write very well but is a malicious person. However *méchant* can also have the same meaning before or after the noun. It is the case in *méchant garcon* and *garcon méchant*. Both mean *bad boy*.
We used the method described previously, i.e. a partition of the semantic space and a computation of distributional classes.
Fig. 8 shows the semantic space linked with *méchant*. We part it into three zones:

- **Zone 1**: upper left corner. This zone corresponds to the more general meanings of *méchant*. In these meanings *méchant* is always used before the nouns. It can be used with any nouns: *méchant soleil (nasty sun), méchant avocat (mediocre lawyer), méchante mémoire (deficient memory), méchante voiture (fantastic car).*
- **Zone 2**: upper right corner. This zone corresponds to meanings of *méchant* used to qualify human beings or their acts. *Méchant* means *causing or likely to cause harm, distress, or trouble*: *enfant méchant (naughty child), méchante sorcière (wicked witch), remarque méchante (nasty remark).*
- **Zone 3**: bottom right corner. This zone corresponds to psychological meanings of *méchant* used to qualify a personality, a character, an intention, an emotion. *Méchant* means *having or showing hatred and a desire to harm somebody or hurt their feelings*: *rictus méchant (evil grin), plaisir méchant (malicious pleasure), regard méchant (malicious look).*

The method of disambiguation is somewhat different. We still associate a potential function with each zone, but we associate each noun with two distributional classes. The first [ANTE] characterizes the noun when the adjective precedes it in the syntagm. The second [POST] characterizes the noun when the adjective follows it. Each class is associated with a potential function. We work with 40 nouns. The results show that we can account for change in meaning according to word order. The best scores are obtained for zone 1, i.e. generals meanings, when the adjective is before the noun. It is a well known hypothesis in linguistics that French adjectives with a large extension are placed preferably before the noun. Among the meanings of *méchant*, it is clearly the general value which has the biggest extension. In other words, in the

anteposition structure, the general meaning is the most frequent except in few cases like *cheval* (*horse*), *couleur* (*color*), *eau* (*water*), *farce* (*joke*), *matin* (*morning*), *mot* (*word*), *taureau* (*bull*). In these cases, there is an ambiguity between the general and the behavourial values. Our model can account for these ambiguities. For example the potential function associated with the [ANTE] class of *farce* presents a large basin covering the zone 1 and zone 2. This noun mingles both values. This kind of ambiguity disappears when *méchant* is used after the noun. Used before *bête* (*beast*), *bois* (*wood*), *bruit* (*sound*), *corps* (*body*), *rire* (*laugh*) or *voix* (*voice*), *méchant* takes a general value as well as a behavioural value. After the same nouns, *méchant* takes only a behavioural (or psychological) value. A noun like *coup* (*blow*), *part* (*part*), *regard* (*look*) or *vérité* (*truth*) influence the meaning of *méchant* in the same way, no matter if the adjective is before or after the noun.

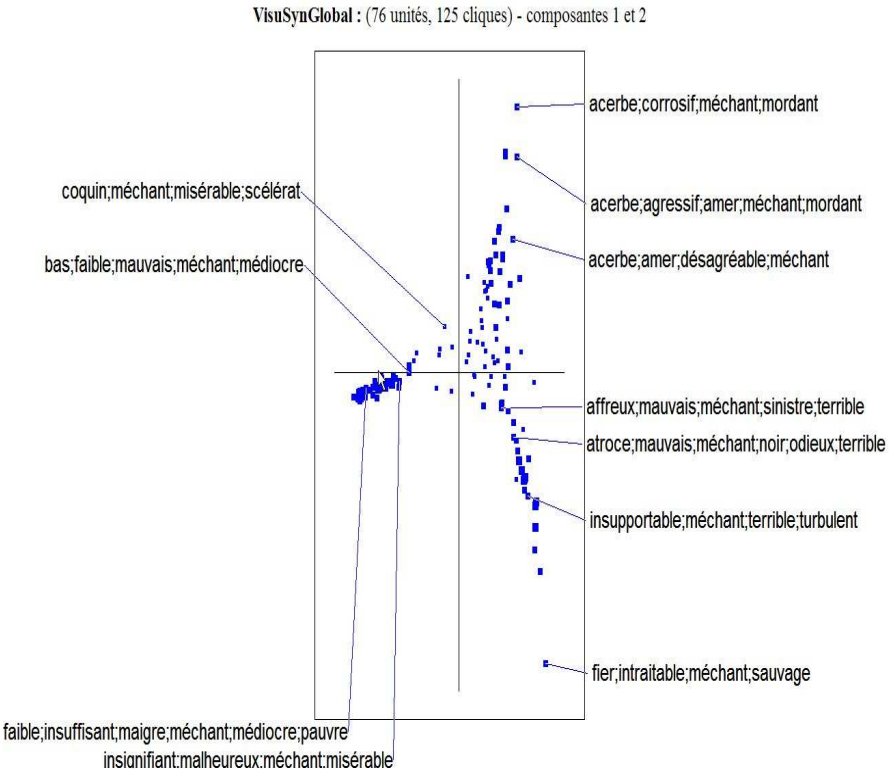

Fig. 8. Semantic space associated with *méchant*

7 Conclusion

This detailed analysis of the French adjectives *sec* and *méchant* shows that the tools we developed for word sense disambiguation are very promising. Just as they are now, they can be utilized to disambiguate any adjective. The different steps in their development brought out various problems linked to synonymy as a descriptive tool. We solved many of them using distributional classes. The cliques show how useful they are in meaning representation as well as in meaning computation. We knew already that using cliques was very efficient in building semantic spaces. Because of their overlapping property all over the semantic space, cliques also constitute a very effective tool for meaning computation. This overlapping combined with the use of potential functions counterbalance the fact that synonymy is a non transitive relation. Using cliques preserves meaning more efficiently than using only synonyms in a one to one relationship. (For example in *coup sévère* (*serious blow*) and *coup sec* (*brusque blow*)). Our model can also account for phenomena as subtle as the change in meaning of an adjective according to its place. Of course this model has some limits and the word is still in progress. The work described in this paper is the first step. We have to go beyond and to develop the system. We want it to be able to give for any couple noun-adjective a correct class for the noun, depending on the adjective, and a correct meaning of the adjective, when used with this noun. We need to make a proper evaluation on a large set of ambiguous adjectives. However, this work shows how continuous mathematic can be relevant for semantic modelization and encourages us in the challenge of using continuity for corpus linguistics (in topics like categorization or variation across register).

References

1. Applied Multivariate Statistical Analysis. (http://www.quantlet.com/mdstat/scripts/mva/htmlbook/).
2. Bouillon P., Viegas E. : The description of adjectives for natural language processing: theoretical and applied perspectives. In: Atelier thématique sur la description des adjectifs pour les traitements informatiques, Institut d'études scientifiques de Cargëse, Corsica,(1999)
3. Fabre C., Bourigault D.:Linguistic clues for corpus-based acquisition of lexical dependencies, Proceedings of the Corpus Linguistics 2001 Conference, UCREL Technical Papers, vol 13, Lancaster University, (2001), pp 176-184
4. Fuchs C., Victorri B.: La polysemie. Construction dynamique du sens. Hermes, Paris (1998)
5. Goes J. : L'adjectif. Entre le nom et le verbe, Paris/Bruxelles, Duculot (1999)
6. Jacquet G. et Venant F. : Construction automatique de classes de sélection distributionnelles, Actes du colloque TALN (2005)
7. Ploux S., Victorri B. : construction d'espaces sémantiques à l'aide de dictionnaires de synonymes. TAL 39 (1) (1998)
8. Venant F., Polysémie et calcul du sens, in Le poids des mots, Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles (JADT), (2004).