# Une structure sémantique pour l'interprétation des énoncés en langue arabe

Anis ZOUAGHI (1), Mounir ZRIGUI (1) et Mohamed BEN AHMED (2)

(1) Labo RIADI – Unité Monastir
Faculté des sciences de Monastir, Tunisie
anis.zouaghi@fsegma.rnu.tn
mounir.zrigui@fsm.rnu.tn
(2) Labo RIADI – Ecole nationale des sciences de l'informatique
à Mannouba – Tunisie
mohamed.benahmed@riadi.rnu.tn

## Résumé – Abstract

Dans le contexte de la compréhension automatique de la langue arabe, nous nous sommes fixés comme objectif la détermination d'un couple (structure de représentation du sens (SRS), processus de construction de ce sens (PCS)), noté par CI (SRS, PCS). Dans le présent article, nous proposons une approche componentielle différentielle, basée sur une étude statistique de corpus pour la réalisation de la structure SRS. Nous rejoignons une approche anthropocentrée, en rejetant le point de vue exclusivement calculatoire du sens, et en donnant un rôle primordial au(x) sujet(s) humain(s), au cours du processus d'interprétation PCS.

In the context of automatic comprehension of the arabic language, we stick to the determination of a couple noted CI (SRS, PCS), where SRS stands for representation structure of meaning and PCS for building process of that meaning.
In the hereby article, we propose a differentiel componential approach, based on a statistic study of corpus to realize the structure SRS. We rejoin an anthropocented approach, while redjecting the point of view exclusively calculatorial of the meaning, and giving prime part to human subject(s), during interpreting process PCS.

## Keywords – Mots Clés

Sémantique componentielle, Approche anthropocentrée, Analyse de corpus, Poids sémantique.

Componentiel semantic, Anthropocentred approach, Corpora analysis, Semantic weight.

# Introduction

Etant conscient, de la complexité d'une modélisation exclusivement calculatoire du sens, en se basant par exemple sur des formalismes logiques et mathématiques, comme la SDRT (Kamp et al., 1998), ou l'ASL (Landauer et al., 1997), nous avons opté pour une approche componentielle et anthropocentrée basée sur une analyse statistique de corpus. En effet, selon une approche anthropocentrée: l'interprétation d'un énoncé est le fruit d'une coopération entre la machine et l'utilisateur. Donc l'intervention du sujet humain dans le processus de construction du sens PCS, permet de réduire le degré de complexité de celui-ci. En plus cette approche est validée par les sciences cognitives, puisque le résultat d'interprétation dépend du sujet humain en interaction avec la machine.

Cette approche s'inscrit dans le courant des travaux réalisés en informatique par (Tanguy., 1997) et (Beust., 1998, 2002). Et dans le cadre des travaux traitant le problème de sens en linguistique de (Rastier., 1987) et (Mel'cuk., 1993). L'originalité de notre approche par rapport à celles cités ci-dessus, est que la structure SRS, s'accorde avec les modèles actuels sur la lecture de textes (O'brien et al., 1998), (Albrecht et al. 1993) et (Kintsch., 1998), sur le fait que la compréhension met en jeu plusieurs niveaux de représentation du texte: la structure de surface, la représentation sémantique et le modèle de situation.

En effet, la structure SRS que nous proposons, permet de représenter la forme morphosyntaxique (structure de surface) de la lexie ainsi que de l'énoncé à interpréter. Elle permet aussi une représentation sémantique, basée sur une description componentielle du sens, en utilisant de traits sémantiques. En plus le processus de construction du sens PCS, tient compte du contexte d'énonciation. Le modèle permet d'attribuer des poids sémantiques aux différents traits sémantiques associés aux lexies, selon le contexte d'énonciation. Ces Ps, sont déterminés à partir d'une étude statistique d'un corpus journalistique préalablement lemmatisé, constitué d'articles de journaux.

# 1 L'approche

L'approche que nous avons adoptée pour la détermination du couple CI (SRS, PCS), est basée sur une analyse de corpus, et d'une coopération homme machine. L'analyse de corpus nous a permis de déduire les lexies représentatives de chaque domaine présent dans le corpus, ainsi que d'étudier la distribution de ces lexies. Cette analyse constitue la première étape d'interprétation : l'interprétation macro sémantique, c'est à dire la détermination du sens d'une lexie en fonction du contexte d'énonciation. La coopération homme machine constitue la deuxième étape d'interprétation. L'utilisateur déclare un ensemble de traits sémantiques, permettant à la machine, de répartir les lexies extraites en sous ensembles appelés classes sémantiques. La figure 1, décrit les étapes contribuant à l'interprétation d'un énoncé.

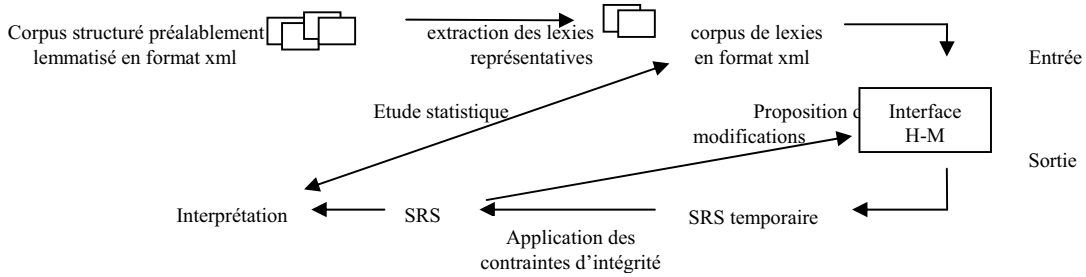


Figure 1 : Description des étapes d'interprétation

## 2 Structure de Représentation du Sens SRS

La structure proposée permet une analyse interprétative en terme d'isotopies. Pour cela, les lexies sont décrites via des traits sémantiques notés Ts. Comme dans le modèle Anadia (Nicolle et al., 2002), nous rejoignons un point de vue componentiel et différentiel. Selon ce point de vue, un Ts permet la description ainsi que la structuration des lexies. Selon notre structure, chaque lexie est décrite par trois traits sémantiques élémentaires de natures différentes.

Les deux premiers traits permettent successivement de préciser à quels domaines et classes sémantiques peut appartenir chaque lexie. Alors que le troisième trait permet d'opposer un signifié avec d'autres signifiés sémantiquement proches, c'est à dire de distinguer ou identifier la signification d'une lexie par rapport aux significations des autres lexies appartenant à la même classe sémantique. Donc le premier trait sémantique permet une description macro sémantique, alors que le troisième trait permet une description micro sémantique. Nous nous sommes convenu de noter successivement ces trois traits sémantiques élémentaires par: $Ts_{mac}$, $Ts_C$ et $Ts_{mic}$. Ainsi, la signification d'une lexie est déterminée par la donnée du triplet Ts ($Ts_{mac}$, $Ts_C$,$Ts_{mic}$).

$Ts_{mac}$ est attribué automatiquement à la lexie à interpréter. Ceci est réalisé au cours du processus d'extraction des lexies représentatives d'un domaine (morceau du corpus). Notre corpus contient par exemple, un ensemble d'articles contenus dans la balise <رياضة 'sport'> </رياضة 'sport'>. Donc le terme رياضة (sport), est un exemple de trait macro sémantique $Ts_{mac}$.

$Ts_C$ et $Ts_{mic}$, sont construits via la coopération homme machine. L'utilisateur propose des traits sémantiques aux différentes lexies, et c'est à la machine de construire la structure SRS: celle-ci est obtenue, à la suite de l'application de contraintes d'intégrités CI et de règles de déduction RD par le processus PCS. Une des contraintes d'intégrité à vérifier par le PCS par exemple, est la contrainte CI1 suivante: Un trait micro sémantique $Ts_{mic}$ ne doit être affecté qu'à une seule lexie au maximum. Formellement, cette contrainte est décrite par la relation suivante:

CI1: $\forall (l_1, l_2) \in CS_i \times CS_j$ avec $i \neq j$, si $Ts_{mic} \rightarrow l_1$ alors $Ts_{mic} \not\rightarrow l_2$ ; où le symbole $\rightarrow$ dénote: implique l'interprétation de la lexie, et CS désigne classe sémantique.Et deux lexies sont considérées synonymes, si la règle RD1 suivante est vérifiée:

RD1: Si $(l_1, l_2) \in CS \times CS$ et on a : $Ts_{mic} \rightarrow l_1$ et $Ts_{mic} \rightarrow l_2$, alors $(l_1, l_2) \in$ Syn.

Chaque domaine $D_i$ est représenté par un arbre sémantique $AS(D_i)$, dont la racine est le trait $Ts_{mac}$. Les premiers antécédents représentent les classes sémantiques $Ts_c$ pouvant appartenir à ce domaine. Les lexies sont réparties dans une classe sémantique selon leurs natures. Les nœuds d'ordre deux décrivent alors la nature des lexies. Et enfin chaque nœud queue de cette arbre représente un trait micro sémantique $Ts_{mic}$ spécifique à une lexie donné l. La figure 2 suivante permet d'illustrer la structure SRS.

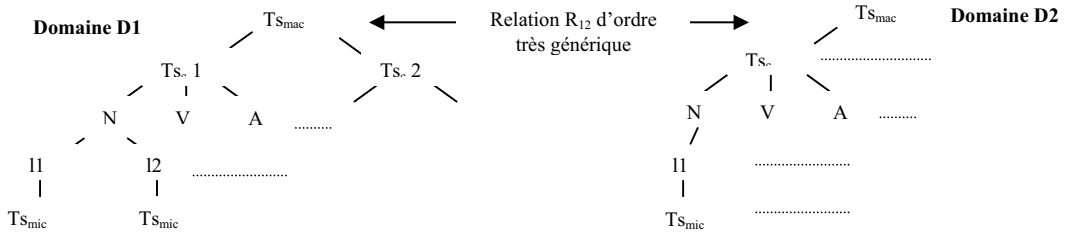


Figure 2 : structure de représentation du sens SRS du couple interprétatif CI.

# 3 Le Processus de Construction du Sens PCS

Nous considérons, comme nous l'avons signalé précédemment, une interprétation en terme d'isotopies. Dans ce paragraphe, nous allons nous concentrer plutôt sur la présentation de la méthode permettant l'attribution des traits sémantiques Ts ($Ts_{mac}$, $Ts_C$, $Ts_{mic}$) en cas d'ambiguïtés. La figure 3 illustre ce genre d'ambiguïté, où une même lexie (la lexie 1 dans la figure) peut appartenir à plusieurs arbres (domaines) de la structure SRS définie:

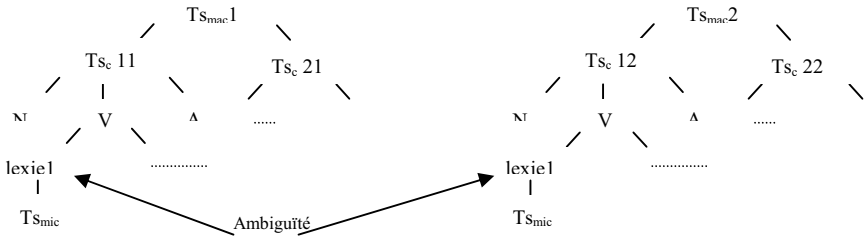


Figure 3 : Un exemple d'ambiguïté provenant des mots polysémiques.

Dans ce cas là, et pour ne pas bloquer le processus d'interprétation PCS, nous avons défini un modèle mathématique permettant d'attribuer des masses (poids sémantiques notés Ps) aux traits sémantiques candidats à la description du sens d'une lexie, selon le contexte d'énonciation. Ce modèle est inspiré de la théorie des fonctions de possibilités (Zadeh., 1978), et il ressemble à ce qui a été proposé par (Le Ny., 1979).

Notre modèle de désambiguïsation est formalisé comme suit: soit une lexie $l \in L$, (où L est l'ensemble des lexies arabes) dans un énoncé, induisant un problème d'interprétation au cours du processus PCS. On postule qu'il existe une quantité finie de possibilités, dont les parties (appelées poids sémantiques et notés Ps), seront distribuées sur les traits sémantiques candidats à être représentatifs du sens de cette lexie l, notés $TsC(TsC_{mac}, TsC_C, TsC_{mic})$. En faite cette quantité de possibilités, est en relation directe avec le nombre de contextes, où peut être utilisé cette lexie l, et elle est déduite à partir de la structure SRS construite. Notre modèle s'inscrit donc dans la voie ontique.

Considérons par exemple la lexie طار (verbe voler). Cette lexie peut avoir deux significations différentes, selon le contexte de son utilisation (énoncés (1) et (2)). Le trait sémantique TsC (حيوانات ″animaux″, عصافير ″oiseaux″, تنقل هوائي ″déplacement aérien″) est un trait candidat pouvant être attribué à la lexie طار (voler). Pour simplifier la représentation, nous allons raisonner sur les traits $TsC_{mic}$. Ainsi, la lexie طار (voler) admet deux significations différentes, en lui attribuant par exemple le trait micro sémantique $TsC_{mic}$ = /تنقل هوائي ″déplacement aérien″/ dans l'énoncé (1), et le trait $TsC_{mic}$ = /إحساس قوى ″sentiment fort″/ dans l'énoncé (2).

(1) طار العصفور (l'oiseau a volé). Et (2) طار من شدة الفرحة (il a sauté de joie).

Ainsi, dans cet exemple les traits micro sémantiques candidats sont : /تنقل هوائي/ et /إحساس قوى /. Et par conséquent les deux possibilités d'affectation d'un $TsC_{lmic}$ à la lexie طار (voler), sont: {/تنقل هوائي/ → طار; /إحساس قوى/ → طار}.

Considérons que nous voulons interpréter l'énoncé (1), donc il faut déterminer le contexte d'énonciation pour pouvoir attribuer le $TsC_{mic}$ approprié à la lexie طار. Ceci est réalisé, en déterminant le poids sémantique Ps de chacune des $TsC_{mic}$, dans l'énoncé (1), et en

sélectionnant ensuite le $TsC_{mic}$ possédant le poids sémantique le plus élevé. La somme des poids sémantiques attribués aux TsC contribuant à la description des différentes significations d'une lexie polysémique est toujours inférieure ou égale à un. On peut écrire alors que:

$$\sum_{TsC->L} m(TsC) \leq 1$$

Avec Ps = m (TsC) est le poids d'un trait sémantique candidat TsC attribué à une lexie l. Et le symbole TsC → l (TsC implique l'interprétation de la lexie l) pour dire que la somme est prise sur toutes les TsC possibles, qui permettent la description du sens de la lexie l.

Revenons maintenant à l'énoncé (1), où nous voulons déterminer le poids sémantique de chacun des deux traits /تنقل هوائي/ et /إحساس قوى/. Ces poids sont déterminés à partir de la formule générale FG suivante:

$$FG: \quad m(TsC) = n_C(l + lco)/N$$

Avec $n_C(l + lco)$ est le nombre de cooccurrences de la lexie dont on veut déterminer la signification (noté l) avec la ou les lexie(s) qui la(les) suit(vent) ou la(les) précède(ent) dans la même phrase (notée(s) lco). N est le nombre total des lexies dans le corpus entier.

# 4 Résultats

Sur un corpus de petite taille (environ 100 pages), notre système appliqué à des énoncés ne contenant pas des anaphores et des métaphores, fournit des résultats satisfaisants. La figure 4, présente le taux d'erreur τ du système au cours du processus d'interprétation de lexies, dans des énoncés de types différents. Pour cela nous avons considéré des échantillons d'énoncés contenant chacun: 10 énoncés simples (ES), 10 énoncés contenants des mots polysémiques (EP), 10 énoncés métaphoriques (EM) et 10 énoncés anaphoriques (EA). Le taux d'erreur τ est déterminé à partir de la formule suivante:

$\tau = (\sum_i Ln_i / N) \times 100$; Avec $\sum_i Ln_i$ est la somme des lexies non interprétées correctement, et N: le nombre de lexies total se trouvant dans chaque type d'énoncés.

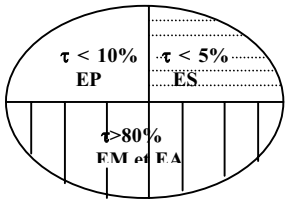


Figure 4: Taux d'erreur du système selon la nature des énoncés: simples, anaphoriques, métaphoriques et polysémiques.

# Conclusion

Dans cet article, nous avons proposé une structure de représentation de sens SRS, et une méthode permettant la désambiguïsation des problèmes dues aux mots polysémiques, lors de l'interprétation d'un énoncé par le processus de construction de sens PCS. Cette approche que nous avons adoptée, constitue une nouveauté pour la langue arabe. En effet la majorité des

systèmes utilisés pour le traitement de l'arabe, se basent sur les techniques de Frames de (Minsky., 75), et sur les modèles des logiques formelles comme dans (Haddad et al., 01). Les résultats actuels sont encourageants et nous comptons les confirmer en améliorant le corpus. Actuellement, nous sommes en cours de tester davantage notre modèle sur un corpus de taille plus grande (un million de mots), et couvrant des thèmes assez variés. Nous prospectons d'améliorer les résultats trouvés, en nous concentrant sur les énoncés anaphoriques et métaphoriques, selon toujours une approche componentielle et anthropocentrée de construction du sens.

# Références

Beust B. (1998), *Contribution à un modèle interactionniste du sens*, Thèse de doctorat d'informatique de l'université de Caen.

Beust B. (2002), Un outil de coloriage de corpus pour la représentation de thèmes, *JADT*.

Haddad B., Yaseen M. (2001), Towards understanding arabic: A logical approach for semantic representation, *Arabic NLP workshop, ACL'01*.

Kamp H., Reyle U. (1998), From discourse to logic, *Dordrecht: kluwer academic publishers*.

Kintsch W. (1998), *Comprehension: A paradigm for cognition*, Cambridge, Cambridge university press.

Landauer T.K., Dumais S.T. (1997), A solution to Plato's problem : the latent semantic analysis theory of the aquisition, induction, and representation of knowledge, *Psychological review*, Vol. 104, pp.211-240.

Nicolle A., Beust P., Perlerin V. (2002), Un analogue de la mémoire pour un agent logiciel interactif, *Revue In Cognito,* Vol. 21, pp. 37-66.

Le Ny J.F. (1979), *La sémantique psychologique,* Paris, PUF.

Mel'cuk I. (1993), Paraphrase et lexique: la théorie sens texte, Actes de *la 4ème école d'été sur le traitement des langues naturelles*, Lannion.

Minsky M. (1975), *A framework for representing knowledge*, New - York, In P. Winston ed.

O'Brien E.J., Rizzella M.L., Albrecht J. E., Hallaeran J.G. (1998), Updating a situational model: A memory-based text processing view, *JEP: Learning, Memory, and Cognitio.*

Rastier F. (1987), *Sémantique interprétative*, Paris, Presses universitaires de France.

Tanguy L. (1997), *Traitement automatique de la langue naturelle: contribution à l'élaboration d'un modèle informatique de la sémantique interprétative*, Thèse de doctorat d'informatique de l'université de Caen.

Zadeh L. (1978), Fuzzy sets as a basis for a theory of possibility, *Fuzzy sets and systems,* Vol.1, pp. 3-28.