

# Conception d'un catégoriseur morphologique fondé sur le principe d'Eric Brill dans un contexte Multi-Agents

Ahmed Haddad<sup>1</sup>, Henda Ben Ghezala<sup>2</sup>, Malek Ghenima<sup>3</sup>  
Laboratoire RIADI, Ecole Nationale des Sciences Informatiques

## 1. Introduction

Il est admis de nos jours que les systèmes de qualité de traitement automatique des langues ne peuvent pas faire l'économie d'une analyse syntaxique des phrases et des textes. Cette analyse permet de déterminer les structures syntaxiques des phrases d'un texte, nécessite plusieurs connaissances telles que, les caractéristiques lexicales, les caractéristiques morphologiques des différents mots, etc. De façon schématique, nous pouvons considérer que cette analyse linguistique comprend « classiquement » les opérations suivantes :

- Découper le texte en phrases et segmenter ces phrases en séquences d'unités lexicales (mot, expressions, etc.).
- Déterminer pour chaque mot déjà segmenté ses caractéristiques morphologiques, et comment ses unités lexicales s'articulent les unes avec les autres pour former des groupes syntaxiques de niveau supérieur. Toutefois, la diversité et la complexité des connaissances à mettre en œuvre dans un système d'analyse automatique de la langue naturelle, et plus particulièrement pour l'arabe, constituent une problématique à part entière pour les chercheurs dans ce domaine et limitent les résultats escomptés de tels systèmes (Aloulou 2003).

Dans cet article, nous commençons par présenter les ambiguïtés typiques de la langue arabe qui peuvent perturber le bon fonctionnement de l'analyse automatique. Nous décrivons par la suite les caractéristiques des systèmes multi-agents en choisissant celles les plus adaptées à notre domaine et nous justifions l'apport de cette approche pour l'analyse linguistique.

## 2. Schéma global de l'analyse syntaxique

Pour effectuer une analyse syntaxique, nous avons besoin habituellement de deux composantes : une composante déclarative, qui correspond aux connaissances linguistiques et une composante procédurale, qui incarne la stratégie d'analyse. Cette stratégie d'analyse correspond à la manière avec laquelle l'analyseur utilise les connaissances linguistiques ou les règles qu'il a à sa disposition pour déterminer quelles structures peuvent être assignées à la phrase analysée.

### 2.1. L'ambiguïté syntaxique de l'arabe

La langue arabe est une langue à flexions qui emploie, pour la conjugaison du verbe et pour la déclinaison du nom, des indices d'aspect, de mode, de temps, de personne, de genre, de nombre et de cas qui sont en général des affixes. Certaines des caractéristiques de la langue arabe peuvent être source d'ambiguïté pour l'analyse automatique. Dans ce qui suit nous allons nous limiter aux ambiguïtés qui ont une incidence directe sur l'analyse syntaxique.

### 2.1.1. *Problèmes d'absence des voyelles*

Les signes de 'voyellation', qui sont réalisés, lorsqu'ils sont notés, sous la forme de signes diacritiques placés au dessus ou au dessous des lettres, apparaissent dans certains textes (coran hadith) ou littéraires (poésie classique, notamment) : On dira qu'ils sont édités en graphie voyellée. La non voyellation génère plusieurs cas d'ambiguïtés lexicales et morphologiques. Nous donnons à titre d'exemple le mot non voyellé ("كتب") [ktb] qui possède 16 voyellations potentielles, représentant 9 catégories grammaticales différentes.

### 2.1.2. *Problèmes de l'ordre des mots dans la phrase*

En passant à un niveau supérieur au mot, à savoir la phrase nous remarquons que l'ordre des mots en arabe est variable. D'une manière générale, on met au début de la phrase le mot sur lequel on veut attirer l'attention et l'on termine sur le terme le plus long ou le plus riche en sens ou en sonorité. Cet ordre, relativement libre des mots, provoque des ambiguïtés syntaxiques artificielles dans la mesure où il faut prévoir dans la grammaire toutes les règles de combinaisons possibles d'inversion de l'ordre des mots dans la phrase.

## 3. **Approche Agent pour les systèmes de TALN**

Nous admettons que la langue arabe étant intrinsèquement ambiguë, il est normal que la grammaire sensée la décrire soit ambiguë, et donc difficile qu'un analyseur déterministe puisse être envisagé. Enrichir la composante lexicale de chaque mot, permet d'une part de réduire au minimum le nombre de règles syntaxiques, et d'autre part de réduire au maximum le nombre d'ambiguïtés qui se génèrent au cours de l'analyse.

Pour cela, et afin de prendre en considération ces problèmes d'ambiguïtés au cours de l'analyse de la langue nous proposons l'intégration de l'analyse syntaxique dans une architecture Multi-agent en vue d'améliorer les performances du système en termes de qualité de solutions et de temps d'exécution. Nous commençons dans cette section par justifier le choix de l'approche agent et exposer brièvement les contributions faites sur le système de TALN avec des approches outre les approches séquentielles.

### 3.1. Justification de l'approche Agent pour les systèmes de TALN

L'application de l'approche agent pour un système d'analyse de la langue arabe consiste à définir une société d'agents et de définir les interactions possibles entre eux. L'avantage de la définition d'une interaction entre les différents agents avec une exécution parallèle ou pseudo-parallèle, est à l'origine de l'approche adoptée.

Dans la littérature, plusieurs systèmes de traitement automatique de la langue, basés sur une approche agent ont vu le jour. Nous citons à titre d'exemple, le système HEARSAY II qui est un système de compréhension de la parole, le système HELENE et qui est un système pour la compréhension de comptes-rendus d'hospitalisation. Le système CAMEL qui est un système de Compréhension Automatique de Récits, Apprentissage et Modélisation des Echanges Langagiers. Le système TALISMAN qui est une architecture multi-agents gouvernée par des lois, pour l'analyse morpho-syntaxique du français écrit. Ces systèmes se distinguent les uns des autres par la méthode de découpage du système en agents (découpage fonctionnel ou découpage objet), la méthode de communication (communication directe ou par tableau noir), la gestion des conflits (utilisation d'un agent médiateur ou superviseur, négociateur), etc. Le tableau suivant illustre une comparaison faite entre les différents systèmes d'un point de vue TAL :

	Spécificité	Couverture de la langue	Morphologie	Syntaxe	Sémantique
<b>HEARSAY II</b>	Compréhension de la parole	Vocabulaire de 1 000 mots		Analyseur simultané syntaxique et sémantique (« grammaire sémantique » hors-contexte)	
<b>CARMEL</b>	Compréhension automatique de récits	15 000 entrées lexicales	Analyse morphologique flexionnelle	Analyseur simultané syntaxique et sémantique (grammaire de cas et grammaire systémique)	
<b>HELENE</b>	Compréhension de comptes rendus d'hospitalisation	320 entrées lexicales	Analyse morpho-lexicale	Analyseur syntaxique (grammaire LFG)	Analyse sémantique et événementielle
<b>TALISMAN</b>	Analyse morphosyntaxique du français écrit	Langue en général (50 000 entrées lexicales)	Analyse morphologique flexionnelle	Analyse syntaxique (grammaire horscontexte de constituants)	

*Tableau 1. Comparaison des systèmes d'un point de vue TAL*

Les critères retenus pour comparer l'architecture des systèmes étudiés sont présentés dans le tableau 2.

	Type de contrôle	Mode de communication	Type d'agents	Utilisation du parallélisme
<b>HEARSAY II</b>	Centralisé (gestionnaire de tâches)	Indirecte (tableau divisé par niveaux)	Sources de connaissances	
<b>CARMEL</b>	Mixte (superviseur et sous-superviseurs)	Indirecte (plusieurs tableaux)	Experts (« processus » ou sources de connaissances)	Oui (plusieurs machines)
<b>HELENE</b>	Centralisé (contrôleur du tableau)	Indirecte (tableau)	"Modules"	
<b>TALISMAN</b>	Mixte (lois globales et locales)	Directe (envois messages et protocole d'apprentissage)	Agents cognitifs	Pseudoparallélisme (simulé)

*Tableau 2. Comparaison des systèmes d'un point de vue multi-agents*

#### 4. Description du système

Dans ce qui suit nous présentons une modélisation de notre système d'analyse de textes arabes en termes d'agents. Notre système « MOUHALIL » est une association d'agents intelligents sous un contrôle décentralisé pour une parfaite autonomie et une communication directe qui assure l'interconnection entre tous les agents du système. Notre système hiérarchique d'agents est basé sur le principe d'Eric Brill suivant des niveaux d'analyses bien définies pour l'analyse morphosyntaxique qui sera détaillé dans l'article.

	Type de contrôle	Mode de Communication	Type d'agents	Utili du parallé
<b>MOUHALIL</b>	<b>Décentralisé</b> agent autonome qui a sa propre capacité de réflexion et d'action	<b>Directe</b> Chaque agent s'interconnecte avec son environnement par l'envoi des messages	Chaque agent est un système multi-agents	Oui

*Tableau 3. Tableau descriptif du système « MOUHALIL »*

Chaque agent du système est chargé d'analyser une phrase du texte selon un prétraitement déjà réalisé précédemment par des mini-agents qui se chargent de la phase de génération de la base des règles qui est une phase primordiale dans le catégoriseur d'Eric Brill.

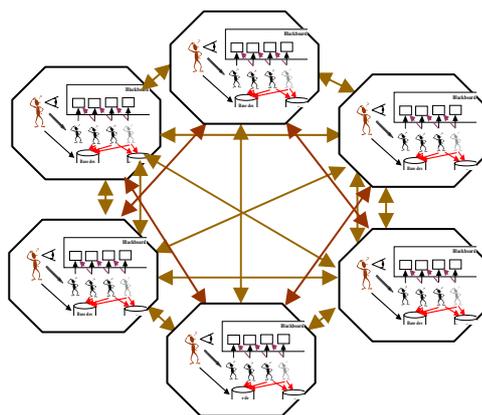


Figure 1. Schéma descriptif du système « MOUHALIL »

Chaque agent va décider parmi les règles possibles dans la base des règles celle qui corrige le plus grand nombre d'erreurs en générant le moins possible de nouvelles erreurs. Pour cela, chaque agent de notre système "MOUHALIL" est lui-même un système multi-agent.

	Type de contrôle	Mode de communication	Type d'agents	Utilisation du parallélisme
<b>Mini-Agent du MOUHALIL</b>	Centralisé (gestionnaire de tâches et moniteur tableau)	Indirecte (tableau divisé par niveaux) <i>morpho syntaxique</i>	Agents réactifs et cognitifs ( )	Oui (pour certains cas)

Tableau 4. Tableau descriptif des agents du système « MOUHALIL »

#### 4.1. Les types d'agents

En général, on distingue deux principaux types d'agents : réactifs ou cognitifs. Une décomposition de ces deux types d'agents suivant leurs compétences peut être trouvée dans Ferber (1995). Le mode de fonctionnement des agents cognitifs est comparé à une organisation de type social et celui des agents réactifs s'inspire, par métaphore, de l'organisation biologique. Le manque de mémoire locale et l'absence de mécanismes de raisonnement des agents réactifs sont des inconvénients pour le traitement des phénomènes linguistiques qui nécessitent la mise en place d'heuristiques complexes ; nous nous intéressons donc aux agents cognitifs. Nous dirons qu'un agent cognitif est capable de raisonner à partir de représentations de son environnement, des autres agents, de ses propres connaissances et de celles des autres.

Deux types d'agents (ou mini-agent) sont présents dans le système : Un agent hybride qui est suivant le cas, soit réactif (agit suivant les informations reçus par l'agent coordinateur, qui contrôle le tableau noir), soit cognitif (en consultant son expertise et les informations de probabilité nécessaires pour décider la nouvelle catégorie). Et un agent coordinateur (superviseur, correcteur, coordinateur, validateur,...) qui se charge de la validité des liens lexicaux, la validation s'appuie sur le principe du catégoriseur d'Eric Brill.

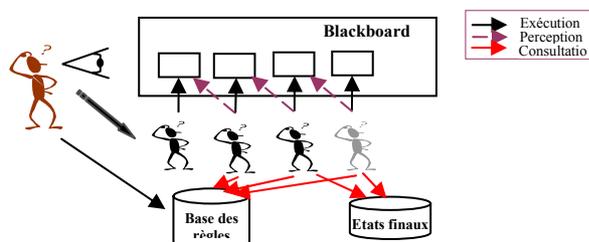


Figure 2. Schéma du système multi-agents

L'agent superviseur a un comportement basé sur le principe d'étiquetage d'Eric Brill. Le résultat décidé par le premier agent hybride sera directement contrôlé par l'agent superviseur, le deuxième agent avant d'écrire son résultat (réflexion) dans le tableau noir, il vérifie la compatibilité de son

résultat avec le résultat du premier agent, si c'est ok alors il l'écrit dans le tableau, sinon il révisé son choix et propose une autre décision, et il refait le même travail jusqu'à une solution homogène avec le premier agent, si aucune décision n'est valide, alors il remet en cause la décision du premier agent en envoyant un message à l'agent superviseur, celui la impose au premier agent de changer sa décision, le troisième agent a le même comportement que le deuxième agent, mais en plus il vérifie si sa décision est un état final ou non, si oui c'est ok, sinon il envoie un message au superviseur pour qu'il active le quatrième agent, le quatrième agent suit le même comportement que le deuxième agent, il ne dispose que d'états finaux.

#### 4.2. Module d'expertise associé à l'agent hybride

Ce module fournit à l'agent les outils nécessaires pour vérifier la morphologie du mot afin d'extraire la catégorie syntaxique. Ce module est constitué de deux types de connaissances : les CSM et les matrices phonologiques pour l'analyse morphologique, et les statistiques tirées à partir d'un corpus pour l'analyse syntaxique.

##### 4.2.1. Analyse morphologique

En admettant que l'arabe classique possède les trente-trois phonèmes et comme l'arabe classique n'a retenu comme phonème que trente-trois combinaisons parmi les centaines que permet le stock universel de traits à l'intérieur d'un même segment, ils sont liées à des restrictions combinatoires et séquentielles qui sont énoncées sous la forme des CSM.

##### 4.2.2. Les conditions de structures morphématiques (CSM)

Les Conditions de Structures Morphématiques (CSM) sont des règles qui régissent la génération des mots dans la langue arabe : un mot qui enfreint une condition ne peut pas appartenir à l'arabe.

**Cadre théorique:** Soit  $x$  l'ensemble des traits possibles définis par la théorie linguistique. Soit  $C$  l'ensemble des 28 consonnes de la langue arabe. Soit  $C_1...C_{|mot|}$  un mot, avec  $C_1, C_{|mot|} \in C$ . Soit  $MP[j][k]$  la matrice phonologique (avec  $1 \leq j \leq 14$  et  $1 \leq k \leq 28$ ) cette matrice représente l'ensemble des traits des consonnes de l'arabe. Soit  $V$  l'ensemble des 6 voyelles de la langue arabe. Soit  $C_1V_1C_{|mot|}V_{|mot|}$  un mot voyellé, avec  $V_1, V_{|mot|} \in V$ . Soit  $MPv[j][k]$  la matrice phonologique des voyelles (avec  $1 \leq j \leq 14$  et  $1 \leq k \leq 6$ )

Les linguistes dénombrent cinq CSM qui régissent la formation des mots arabes. Ces conditions sont classées en deux types: les restrictions combinatoires et séquentielles.

##### 4.2.3. Restrictions combinatoires :

Ces restrictions régissent les spécifications des traits correspondant aux phonèmes de la langue arabes. Dans ce cas trois règles sont à énoncer :

1) *CSM1 : tous les phonèmes sont [-aspirés]*

Tout phonème de l'arabe est une colonne de  $x$  spécifications correspondant à ces  $x$  traits, les ( $x$ -quatorze) spécifications qui ne sont pas représentées découlent automatiquement des quatorze présentes en vertu de conditions propres à l'arabe classique. La condition CSM1 distingue l'arabe classique de nombreuses langues naturelles qui opposent phonèmes aspirés et non aspirés. C'est l'existence de telles restrictions valables pour tous les phonèmes de l'arabe classique, qui a permis de ne faire figurer que quatorze traits, parmi  $x$  traits possibles définis par la théorie linguistique.

2) *CSM2 : tous les phonèmes vocaliques sont [-nasal]*

La condition CSM2 exclut les voyelles nasales de l'inventaire des phonèmes de l'arabe.

Si  $c_i \in C$  et  $c_i \subset C_1...C_k$  (avec  $1 \leq i \leq 28, k = |mot|$ ) alors  $MP[aspiré][i] = [-]$

Si  $v_i \in V$  et  $v_i \subset C_1V_1...C_{|mot|}V_{|mot|}$  (avec  $1 \leq i \leq 6$ ) alors  $MPv[nasale][i] = [-]$

#### 4.2.4. Restrictions séquentielles

Ce sont des restrictions qui lient les spécifications de traits appartenant à des segments successifs de la matrice de l'arabe classique, ces restrictions reflètent le fait que n'importe quelle séquence de phonèmes de l'arabe n'est pas un morphème-racine ou un allomorphe possible (variante combinatoire d'un phonème). Par exemple *مَد* et *كَجَب* sont des séquences permises par la structure de la langue, mais pas *خَخَد*.

Le fait qu'il n'existe aucun morphème-racine dont la représentation phonologique soit "كَجَب" n'est la séquence d'aucune contrainte structurelle, il s'agit seulement d'une lacune accidentelle : Il s'agit d'une combinaison admissible par la structure de la langue, mais qui est absente du lexique. En revanche, des séquences telles que "خَخَد" ou "نَبَد" ne sont pas des morphèmes-racines possibles en arabe classique. La première enfreint la restriction qui est exprimée par la condition CSM4 et la seconde celle qui est exprimée par CSM5 :

*CSM4 : La condition CSM4 exclut de l'ensemble des morphèmes-racines possibles en arabe classique toute séquence de phonèmes formée de deux segments identiques, en première et en deuxième consonne radicale.*

Si  $c, d \in C$  et  $c, d \subset \{C_1 C_2 C_3\}$  (tel que  $c = C_1$  et  $d = C_2$ ) alors ( $c \neq d$ ).

*CSM5 : La condition CSM5 interdit des consonnes identiques qui sont [+continu, +voisé] en première et troisième consonne radicale.*

Si  $[(MP[continu][i,j] = [+], MP[voisé][i,j] = [+])$  et  $(MP[continu][i,j] = [+], MP[voisé][i,j] = [+])$  alors  $c_{i,j} X c_{i,j} \notin C_1 C_2 C_3$ .

#### 4.2.5. Analyse syntaxique

Dans cette phase l'agent consulte le dictionnaire de la langue arabe pour extraire la catégorie du mot traité, bien que ce dictionnaire ne comporte pas véritablement de probabilités, il indique pour chaque mot, en plus des catégories possibles, celle qui est la plus fréquente. Il est construit à l'aide d'un corpus étiqueté.

### 4.3. Module d'expertise associé à l'agent superviseur

#### 4.3.1. Principe d'étiquetage d'Eric Brill

L'analyse d'un texte par cette méthode est basée sur une suite de corrections. Lors de la période d'entraînement, le programme reconnaît lui-même ses faiblesses et les corrige. Une fois entraîné, le programme commence par affecter à chaque mot d'un texte sa catégorie la plus probable et applique, ensuite, des règles de corrections.

#### 4.3.2. La base de règles

Puisque le premier passage sur un texte affecte la catégorie la plus fréquente à chaque mot, de nombreuses erreurs apparaissent. Le principe de construction de la base de règles consiste à chercher la règle qui corrige le plus grand nombre d'erreurs en générant le moins possible de nouvelles erreurs, à l'appliquer et à chercher la règle suivante sur le même principe.

#### 4.3.3. Schéma d'assignation grammaticale

La méthode d'assignation grammaticale mise au point par Éric Brill au département d'informatique de l'Université de Pennsylvanie est basée sur la recherche et la correction d'erreurs. Pour les mots inconnus, la catégorie qui apparaît le plus fréquemment pendant la période d'entraînement du lexique. Nous avons enrichis cette méthode par l'ajout d'un module de filtrage correspondant au traitement des mots inconnus.

**Module de Filtrage :** Le module de filtrage s'intègre dans la phase de recherche du mot dans le dictionnaire, où chaque mot doit vérifier les conditions de structures morphématiques, toutes les conditions doivent être satisfaites et vérifiées pour que le mot soit un lexème de la langue arabe, sinon

(si une des conditions n'est pas vérifiée, alors c'est un mot inconnu). Dans le cas où toutes les conditions sont satisfaites, alors le mot étudié est un nouveau lexème de la langue arabe (néologisme) et le catégoriseur fournit la catégorie la plus probable.

#### 4.4. Le type de comportement des agents

Le type de comportement interne d'un agent peut varier (égoïste vs altruiste), de même, le type de comportement externe peut être différent (coopératif vs conflictuel). En pratique, il est parfois nécessaire de définir un système où la coopération volontaire et la concurrence entre agents coexistent (ce qui paraît le plus représentatif de la réalité humaine), afin d'améliorer la résolution d'un problème. Ce type de système correspond davantage à un système d'analyse automatique de la langue. En effet, les agents d'un système d'analyse ont un but commun et pour l'atteindre ils doivent coopérer puisque ils possèdent des compétences complémentaires. Il ne faut cependant pas exclure la mise en concurrence de différentes méthodes (réparties dans différents agents) pour atteindre un même sous-but, ce qui peut être profitable aux performances du système. Le rôle de l'agent hybride est de :

Vérifier l'intégrité lexicale du mot traité (si c'est un mot valide pour la langue arabe, c'est-à-dire un mot accepté par les règles de la langue arabe : CSM et ML).

Assigner au mot traité sa catégorie lexicale appropriée suivant des calculs statistiques à partir d'un corpus lemmatisé.

Le rôle de l'agent superviseur :

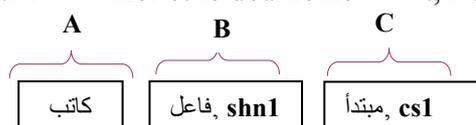
Contrôler les catégories assignées par les agents hybrides, il vérifie les agents par triplet. Cet agent à un fonctionnement basé sur le catégoriseur ou étiqueteur d'Eric Brill. L'analyse d'un texte par cette méthode est basée sur une suite de corrections. Lors de la période d'entraînement, le programme reconnaît lui-même ses faiblesses et les corrige. Une fois entraîné, le système commence par affecter à chaque mot d'un texte sa catégorie la plus probable et applique, ensuite, des règles de correction.

#### 4.5. Les modes de communication

Notre SMA dispose de deux modes de communication. Pour le contrôle des décisions des agents hybrides, nous avons choisis une communication indirecte : les interactions par partages d'informations (blackboard) qui nécessitent l'existence d'une structure de données commune. Concernant la correction des décisions émises par les agents hybrides en se basant sur les règles d'Eric Brill et l'interaction entre l'agent superviseur et les agents hybrides, nous avons choisis une communication directe : intentionnelle par envoi de messages.

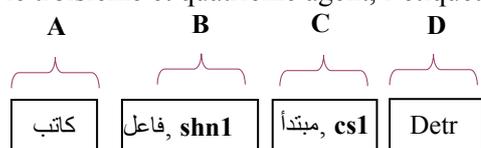
• **La communication par mémoire partagée** nécessite une structure de données commune à l'ensemble des agents, comprenant tous les éléments nécessaires à la résolution d'un problème. Les agents déposent et lisent les informations sur cette zone de données communes appelée « tableau noir » (ou « blackboard »). Les agents hybrides déposent leurs étiquettes (catégorie lexicale et syntaxique), contrôlées par l'agent correcteur, et soit il les valide, soit il dépose de nouvelles étiquettes.

Pour le premier et le deuxième agent, l'étiquette est composée de trois champs :



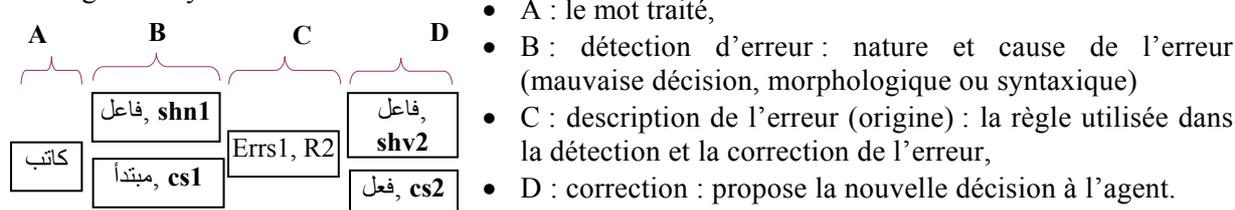
- A : le mot traité,
- B : la catégorie morphologique,
- C : la catégorie syntaxique.

Pour le troisième et quatrième agent, l'étiquette comprend quatre champs :



- A : le mot traité,
- B : la catégorie morphologique,
- C : la catégorie syntaxique,
- D : si le mot traité est déterministe ou non.

• **La communication par messages** est représentée par l'envoi de messages asynchrones suivant les protocoles de communication définis. Ces messages sont envoyés par l'agent correcteur lorsqu'il aperçoit une mauvaise combinaison d'étiquettes qui ne coïncident pas avec les règles qu'il a, donc il envoie des messages aux agents concernés pour qu'ils puissent rectifier leurs décisions, les messages envoyés sont constitués de :



كاتب (katibon) : le mot traité par l'agent hybride1,

فعل (fialon), **shn1** : 1<sup>er</sup> schème nominale dans la catégorie morphologique,

مبتدأ (moptadaa), **cs1** : première catégorie syntaxique,

Errs1, R2 : erreur 1 (lorsque l'agent décide que le mot traité est un non or c'est un verbe), cet erreur est déduite après l'application de la règle R2,

فاعل (faailon), **shv2** : deuxième schème verbale dans la cat morphologique,

فعل (fialon), **cs2** : deuxième catégorie syntaxique.

## 5. Conclusion

Nous avons présenté dans cet article, dans un premier temps, les problèmes d'ambiguïtés inhérents à la langue naturelle arabe et qui perturbent les différentes phases d'analyse et principalement la phase syntaxique. Dans un deuxième temps, nous avons proposé une modélisation du traitement automatique de l'arabe selon une approche multi-agent, faisant coopérer un ensemble hiérarchique d'agents experts selon le principe d'Eric Brill afin d'améliorer la qualité de l'analyse et proposer des analyses morphosyntaxique même en présence d'ambiguïté.

## Références

- ALOULOU C. (2003), « Analyse syntaxique de l'arabe : le système MASPAS », in *Actes de RECITAL 2003*, Batz-sur-Mer, France.
- ABBES R. (2004), *La conception et la réalisation d'un concordancier électronique pour l'arabe*, Thèse non publiée, Institut national des sciences appliquées, Lyon.
- BRILL E. (1993), *A Corpus-Based Approach to Language Learning*, Thèse non publiée, Université de Pennsylvania, Department of Computer and Information Science.
- ESFANDIARI B. (1995), « Systèmes multi-agents et gestion de réseaux », in *Actes des 5<sup>ème</sup> journées nationales PRC-GDR Intelligence Artificielle*. Nancy, février 1995.
- FERBER J. (1995), *Les systèmes multi-agents vers une intelligence collective*. InterEdition, Paris.
- HABAILI H. (1976), *Contraintes de structure morphématique en arabe*, DEA en linguistique, Université de Montréal.
- WARREN K. (1998), *Gestion de conflits dans une architecture multi-agents d'analyse automatique de textes*, Thèse non publiée, Université de Stendhal - Grenoble III.