Un système multi-agent pour la détection et la correction des erreurs cachées en langue arabe

(1) Chiraz Ben Othmane Zribi, (2) Fériel Ben Fraj, (3) Mohamed Ben Ahmed

Laboratoire RIADI – Université La Manouba
ENSI, La Manouba, Tunisie
(1) Chiraz.benothmane@riadi.rnu.tn, (2)Feriel.BenFraj@riadi.rnu.tn,
(3)Mohamed.BenAhmed@riadi.rnu.tn

Mots-clés : Erreurs orthographiques cachées, Détection, Correction, Système multi-agent, Analyse linguistique, Langue arabe

Keywords: Hidden spelling errors, Detection, Correction, Multi-Agent System, Linguistic analysis, Arabic language

Résumé

Cet article s'intéresse au problème des erreurs orthographiques produisant des mots lexicalement corrects dans des textes en langue arabe. Après la description de l'influence des spécificités de la langue arabe sur l'augmentation du risque de commettre ces fautes cachées, nous proposons une classification hiérarchique de ces erreurs en deux grandes catégories ; à savoir syntaxique et sémantique. Nous présentons, également, l'architecture multi-agent que nous avons adoptée pour la détection et la correction des erreurs cachées en textes arabes. Nous examinons alors, les comportements sociaux des agents au sein de leurs organisations respectives et de leur environnement. Nous exposons vers la fin la mise en place et l'évaluation du système réalisé.

Abstract

In this paper, we address the problem of detecting and correcting hidden spelling errors in Arabic texts. Hidden spelling errors are morphologically valid words and therefore they cannot be detected or corrected by conventional spell checking programs. In the work presented here, we investigate this kind of errors as they relate to the Arabic language. We start by proposing a classification of these errors in two main categories: syntactic and semantic, then we present our multi-agent system for hidden spelling errors detection and correction. The multi-agent architecture is justified by the need for collaboration, parallelism and competition, in addition to the need for information exchange between the different analysis phases. Finally, we describe the testing framework used to evaluate the system implemented.

1 Introduction

Le problème des erreurs cachées présente une incommodité pour les scripteurs lors de la saisie de leurs textes. Ces fautes ne sont autres que des erreurs orthographiques qui produisent pour autant des mots lexicalement corrects. La présence d'un vocable au sein d'un contexte syntaxique ou sémantique qui n'est pas le sien peut rendre insensée toute la phrase. L'exemple suivant illustre ce phénomène :

Exemple : Le jardinier utilise le *gâteau* pour bêcher la terre
Le mot *"gâteau"* est introduit dans un contexte qui ne lui est pas approprié. Cette faute de frappe peut être corrigée en la changeant par le vocable *"râteau"*.

Les statistiques réalisées par Eastman et Oakman (1991) (Verberne, 2002), affirment que les erreurs cachées comptent 25% parmi toutes les erreurs orthographiques commises et contenues dans leur corpus de référence. Mitton (1987) leur attribue une valeur plus grande à savoir : 40% parmi toutes les erreurs orthographiques étudiées. Ces deux valeurs assez importantes ont rendu l'étude de ce genre d'erreurs une nécessité en soi. En effet, plusieurs recherches ont été entreprises dans le but de remédier à ce problème. Nous pouvons, alors, citer comme exemples : les recherches de Golding (Golding, Dan, 1996) qui a étudié ce genre d'erreurs en langue anglaise. Il a ainsi proposé de multiples méthodes comme la méthode de Bayes (Golding, 1995), la méthode des trigrammes des parties du discours (Golding, Schabes, 1996) et la méthode à base de réseaux neuronaux dite Winnow (Golding, Dan, 1999). Le chinois a été aussi traité par les deux chercheurs Jianhua et Xiaolong (Xiaolong, Jianhua, 2001). Le suédois a fait l'objet d'une recherche pareille avec Bigert et Knutsson (Bigert, Knutsson, 2002).

En ce qui concerne la langue arabe, aucun travail n'a été réalisé sur les erreurs cachées malgré l'importance de l'entreprise d'une telle recherche. La langue arabe présente, en effet, des spécificités qui rendent le risque de commettre une erreur cachée plus important que pour les autres langues. Nous nous sommes donc proposés de nous intéresser à ce problème en construisant un système permettant à la fois de détecter et de corriger ce type d'erreurs pouvant survenir dans des textes arabes. A cause de la complexité de ce travail, nous avons été amenés à émettre certaines hypothèses pour restreindre les champs de nos investigations. Nous avons considéré alors l'arabe non voyellé avec une seule erreur cachée par phrase. L'erreur consiste en une seule faute d'édition à savoir ; l'ajout d'un caractère, l'omission d'un caractère, la substitution d'un caractère par un autre ou l'interversion de deux caractères adjacents. Des statistiques ont en effet montré que l'une (seulement) de ces opérations est à l'origine d'une erreur orthographique dans 90% des cas (Ben Hamadou, 1993).

Dans ce qui suit, nous décrivons dans la première section les spécificités de la langue arabe qui ont participé à augmenter le risque de commettre des erreurs cachées lors de la saisie des textes. Dans la seconde section, nous présentons la classification que nous avons adoptée pour ces erreurs. Sur la base de ladite classification et de nos besoins, nous avons opté pour une architecture multi-agent dont la conception est décrite dans la troisième section de l'article. La quatrième section est consacrée, quant à elle, à la description des résultats de l'évaluation du système mis en place.

2 Quelles difficultés pour la langue arabe ?

Pour la langue arabe, le problème des erreurs cachées s'aggrave et se complique encore plus que pour d'autres langues (notamment indo-européennes). En effet, on se trouve confronté dans cette langue, à des contraintes d'écriture et à diverses ambiguïtés : telles que l'agglutination des enclinomènes aux formes simples, l'ambiguïté grammaticale, et la proximité lexicale des formes textuelles.

2.1 Le phénomène d'agglutination

L'agglutination est l'ajout de préfixes appelés *'proclitiques'* et de suffixes dits *'enclitiques'* aux formes simples pour obtenir ce qu'on a appelle des "formes agglutinantes" ou encore des "hyperformes". Les proclitiques et les enclitiques forment l'ensemble des enclinomènes[1] de la langue arabe. Une erreur cachée peut être la conséquence d'une opération d'ajout ou d'omission d'un enclinomène à un radical. Cette opération peut donner naissance à une forme textuelle licite, mais malencontreusement, injectée dans un contexte qui ne lui est pas approprié.

2.2 L'ambiguïté grammaticale

Les mots en arabe présentent une ambiguïté grammaticale. Les statistiques réalisées par (Debili et al., 2002) sur un texte en langue arabe confirment cette ambiguïté. L'auteur a constaté l'importance du taux d'ambiguïté grammaticale pour les textes voyellés qui est égale à 5,63 en moyenne. Ce taux est augmenté par l'absence des voyelles pour atteindre en moyenne 8,71 catégories grammaticales par forme textuelle. L'ambiguïté grammaticale est l'une des causes de l'abondance des erreurs cachées en langue arabe car ce genre de fautes peut être dû à une confusion dans l'interprétation grammaticale des formes textuelles.

2.3 La proximité lexicale

La caractéristique des mots arabes la plus remarquable pour notre problématique est celle de la proximité lexicale. En effet, les mots en langue arabe sont très proches graphiquement les uns des autres, nous pouvons l'affirmer en nous basant sur l'étude réalisée par (Ben Othmane Zribi, 1998) à ce propos. Cette expérience consistait à appliquer les quatre opérations d'édition, précédemment citées, sur tous les mots d'un dictionnaire de la langue. Parmi les formes, automatiquement construites, on a procédé par dénombrer les graphies correctes, comptant ainsi ce qu'on appelle *"le nombre de mots approchants ou lexicalement voisins"*. Ces comptages nous ont donné une idée claire sur la similarité dite aussi le degré de ressemblance entre les vocables d'une langue. Ainsi, avons-nous constaté que les mots en langue arabe sont beaucoup plus proches les uns des autres avec un nombre moyen de formes voisines de 26,5 : valeur importante comparée à celle calculée pour la langue française égale à 3,5 et celle relative à l'anglais égale à 3. De ce fait, la ressemblance typographique des mots

---

[1] Ce sont des particules affixes représentant les pronoms personnels, les conjonctions, les prépositions, etc.

en arabe augmente le risque de tomber sur des erreurs cachées, comme elle augmente la taille de la liste des candidats à la correction lors de la phase de correction.

3 Typologie des erreurs cachées

Pour détecter les erreurs cachées, une simple analyse morphologique s'avère insuffisante puisque ces erreurs engendrent des formes morphologiquement correctes mais erronées sur le plan syntaxique ou sémantique (voir même pragmatique). Ainsi, une phrase contenant une erreur syntaxique est lexicalement correcte mais la structuration de ses mots est incorrecte. On parle alors d'un dérèglement syntaxique. Par contre, une phrase contenant une erreur sémantique est dénuée de sens puisque l'erreur est un mot venant s'intercaler dans un contexte sémantique qui n'est pas le sien.

3.1 Les anomalies syntaxiques

Les dérèglements grammaticaux peuvent être de différents types. Ainsi, la classe des anomalies grammaticale peut être subdivisée en des sous-classes d'erreurs syntaxiques qui se présentent comme suit :

- Les erreurs d'accord : Ce sont des dérèglements syntaxiques qui sont dus au non-respect des contraintes d'accord qui gèrent la langue. Ils causent des incompatibilités au niveau des informations morpho-syntaxiques relatives aux mots d'une même phrase. Exemple : (الكبيرة) "رفع المسافر الحقيبة الكبير", ("Le voyageur a soulevé le grand valise », la forme correcte est : la grande).

- Les erreurs liées à la transitivité : La transitivité est un lien syntaxique (et sémantique) entre un verbe et un complément d'objet lui succédant en général. L'absence de ce complément ou son apparition là où il ne faut pas rend incorrecte la phrase. Exemple : (بالحمّى) "مرض الرضيع الحمى" ("le nourrisson est tombé malade fièvre", la forme correcte est : à cause de la fièvre).

- Les erreurs d'agrammaticalité : Ce genre de fautes concerne l'agencement des catégories grammaticales au sein de la phrase. Exemple: (مكتبه) "جلس المدير في نكتبه" ("le directeur s'est assis dans nous l'écrivons", la forme correcte est : son bureau).

3.2 Les anomalies syntaxiques

Tout comme pour le niveau syntaxique, les anomalies sémantiques peuvent être scindées en des sous classes d'erreurs dont nous pouvons citer :

- Les incompatibilités sémantiques : Une incompatibilité sémantique consiste en l'injection d'un mot dans un contexte sémantique qui n'est pas le sien. Exemple : (سمكة) "وجد الصياد سكة كبيرة" ("Le pécheur a trouvé une grande voie", la forme correcte est : poisson ; en langue arabe le mot poisson est au féminin)

- Les incomplétudes sémantiques : L'omission ou l'insertion là où il ne faut pas d'une conjonction de coordination ou toute autre particule rendent parfois la phrase dénuée de sens (Aloulou, 1996). Exemple : "ضربت الولد بكى (فبكى) " ("j'ai frappé le garçon il a pleuré", la forme correcte est : ensuite il a pleuré).

4 La solution proposée

La complexité de notre problème ainsi constatée, ainsi que la hiérarchie présentée dans la classification des erreurs cachées dénotent la nécessité d'une interférence entre les différentes phases d'analyse pour le traitement de ce genre d'erreurs. En effet, la détection et la correction des erreurs syntaxiques nécessitent la contribution des connaissances sémantiques. De même que, le traitement des erreurs cachées de type sémantique nécessite des retours en arrière syntaxiques pour une meilleure détection et correction.

En conséquence, nous avons opté pour une architecture multi-agent où les différents agents travaillent en : collaboration, compétition, coordination et parallélisme afin d'atteindre l'objectif global du système. Chacun d'eux apporte sa contribution à la solution finale. Tous s'organisent dans une société commune où ils peuvent discuter et coopérer.

5 L'architecture du système

Pour qu'un système de vérification des textes en langue naturelle soit efficace, il doit avoir à sa disposition un ensemble d'informations linguistiques concernant ces textes. C'est pour cette raison que nous nous sommes proposés d'effectuer une analyse morpho-syntaxique des textes à traiter. Ces textes analysés serviront comme entrées pour notre système. La figure1 illustre l'architecture générale de notre système.

5.1 Le groupe syntaxique d'agents

Ce groupe est formé de quatre agents à savoir ; l'agent Accord, l'agent Transitivité, l'agent Compatibilité syntaxique et l'agent superviseur des trois premiers. Le superviseur syntaxique reçoit alors le texte à vérifier et l'envoie phrase par phrase à ses collègues de la même société d'agents.

- **L'agent Accord** vérifie la validité des contraintes d'accord en appliquant un ensemble de règles d'accord (on compte environ 800 règles).
- **L'agent Transitivité** essaie de détecter les anomalies pouvant exister entre les verbes et leurs compléments d'objet.
- **L'agent Compatibilité syntaxique** vérifie la structuration des Catégories Grammaticales des Hyperformes (HyperCGs) dans la phrase en considérant des séquences ternaires d'HyperCGs. Il utilise à cet effet, une matrice à trois dimensions qui indique la validité de ces séquences ternaires.

Les travaux de ces trois agents sont contrôlés par le superviseur. En effet, une fois que l'un d'eux détecte une anomalie, il informe ses collègues pour qu'ils arrêtent leurs traitements respectifs et annonce la nouvelle au superviseur pour que ce dernier déclenche le processus de correction.

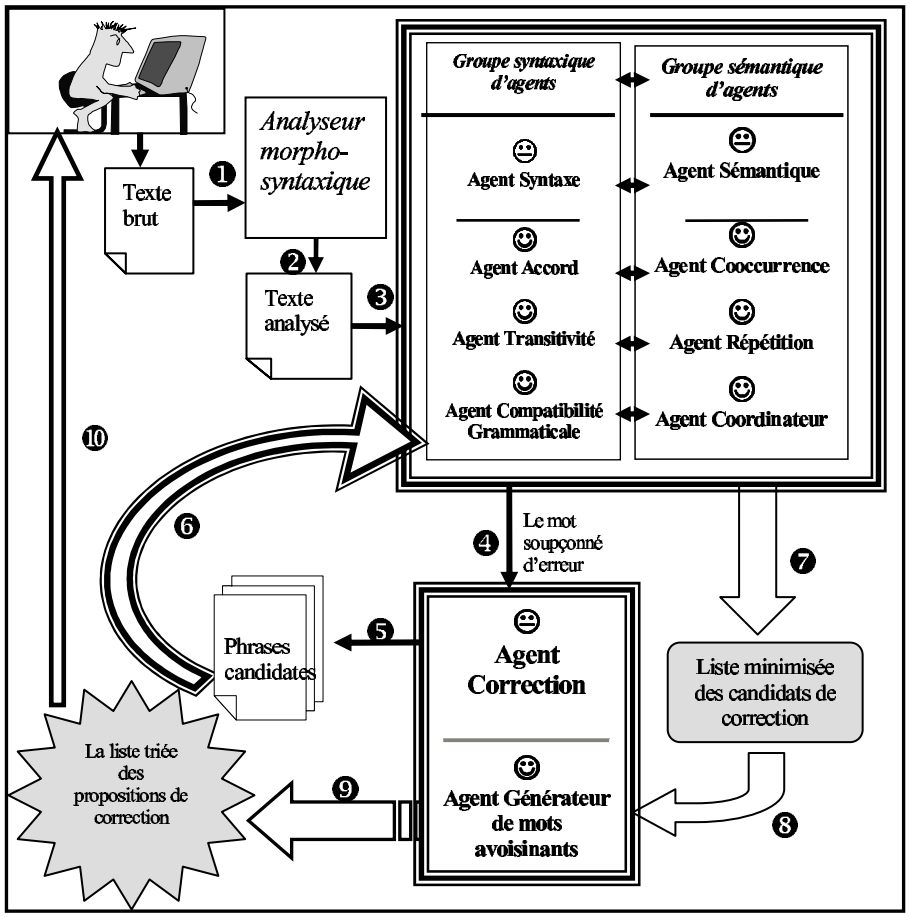

Figure 1 : Architecture du système Multi-Agent de détection/correction des erreurs cachées

5.2 Le groupe sémantique d'agents

Ce groupe est aussi constitué de quatre agents. Le premier est le superviseur qui envoie le texte découpé en phrases aux autres agents du même groupe. Les trois autres sont : l'agent Cooccurrence, l'agent Répétition et l'agent Coordinateur.

Notons :

$Ph = \{m_1, ..., m_i, ..., m_n\}$: la phrase soumise à la vérification sémantique.

$C = \{c_{-k}, ..., c_{-1}, c_1, ..., c_k\}$: l'ensemble des mots entourant le mot analysé (en considérant une fenêtre de taille k).

$L = \{l_1, \ldots, l_i, \ldots, l_n\}$: l'ensemble des lemmes des mots de la phrase.

- **L'agent Cooccurrence** vérifie pour chaque mot de la phrase s'il possède des affinités sémantiques avec son contexte. Il procède alors de deux façons, ne pouvant être qualifiées de différentes mais plutôt de complémentaires. Cet agent va, tout d'abord, chercher s'il existe des collocations[2] entre le mot cible d'analyse et les mots l'avoisinant dans le même contexte. Les collocations, si elles sont trouvées, vont

[2] Une collocation est une association habituelle de deux ou plusieurs termes (collocats) au sein d'un discours.

permettre de conforter chaque mot dans le contexte où il a été mis. Ainsi, pour chaque vocable est attribuée une valeur dite information mutuelle calculée à l'aide de la formule suivante :

$$I(m_i) = \max_{k \leq j \leq k} \text{Log} \frac{p(m_i, c_j)}{p(m_i) \times p(c_j)}$$

Avec $p(m_i)$ la probabilité d'observer m_i, $p(c_j)$ la probabilité d'observer c_j et $p(m_i,c_j)$ la probabilité de les observer ensemble.

Affirmer que le mot m_i est en collocation avec son contexte revient à évaluer la valeur $I(m_i)$ de la façon suivante :

- si $I(m_i)$ est positive, alors m_i est en collocation avec son contexte.
- si $I(m_i)$ s'approche de la valeur nulle alors m_i n'a pas de rapport avec son contexte.
- si $I(m_i)$ est négative alors m_i a des distributions complémentaires avec son contexte.

Outre les collocations, l'agent Cooccurrence va chercher s'il existe des cooccurrences ordinaires entre chaque mot cible d'analyse et son entourage. Pour ce faire, nous avons choisi d'utiliser la formule des probabilités conditionnelles de Bayes suivante :

$$p(m_i|C) = \frac{p(C|m_i) \times p(m_i)}{p(C)}$$

Plus la valeur $p(m_i/C)$ est élevée, plus le mot analysé m_i a d'affinité sémantique au sein de son contexte C.

- **L'agent Répétition**, pour sa part, cherche si le lemme de la forme textuelle à analyser se répète au sein du texte lui même. Il se base sur le principe qui dit que *'les mots ou plus précisément les lemmes des mots d'un texte ont tendance à se répéter dans le texte lui-même'*. En effet, d'après des comptages réalisés (Ben Othmane Zribi, Ben Ahmed, 2003) sur un corpus textuel en langue arabe appartenant à un domaine bien particulier, nous savons qu'une forme textuelle peut apparaître en moyenne 5,6 fois alors qu'un lemme peut apparaître en moyenne 6,3 fois et ce dans le même texte. Nous calculons alors pour chaque lemme sa fréquence d'occurrence au sein de tout le texte, avec la formule suivante :

$$p(l_i) = \frac{\text{nombre d'occurrences de } l_i}{\text{nombre total de lemmes}}$$

Nous considérons alors que plus la valeur de la probabilité $p(l_i)$ est élevée plus il y a affinité sémantique entre le mot m_i à vérifier dont le lemme est l_i et le texte où il a été mis.

- **L'agent Coordinateur** englobe les résultats trouvés par les deux agents Cooccurrence et Répétition dans la formule :

$$F(m_i) = \alpha * I(m_i) + \beta * p(m_i|C) + \lambda * p(l_i)$$

Avec $F(m_i)$ la fréquence totale d'apparition du mot m_i au sein du texte, α, β et λ sont trois coefficients liés aux trois probabilités contextuelles calculées, les valeurs associées à ces coefficients ne peuvent être prédites, mais il faut qu'elles soient obtenues à travers des tests et des comparaisons de pertinence. Toutefois, nous estimons que les valeur de α et β sont plus importantes que celle de λ vu que l'importance du contexte voisin du mot analysé est, sans aucun doute, plus grande que celle du contexte lointain de ce même mot.

Pour chaque mot, nous calculons la valeur de $F(m_i)$ qui, comparée aux valeurs des mots voisins et à une valeur seuil, va confirmer ou infirmer la validité de l'existence de ce mot entouré de ses voisins.

Le résultat final de la vérification sémantique est aussitôt envoyé au superviseur pour qu'il déclenche le processus de correction.

5.3 Collaboration des deux groupes d'agents : syntaxique et sémantique

Avant le déclanchement du processus de correction, les deux groupes d'agents syntaxique et sémantique entrent en communication par le biais de leurs superviseurs respectifs. Etant donné que ces deux groupes d'agents sont lancés en parallèle à la recherche d'une erreur cachée dans une phrase, le premier qui trouve doit informer l'autre et demander une confirmation sur l'erreur. Si consensus, il y a, le processus de détection est arrêté et l'erreur est envoyée vers l'agent Correction. Il s'agit dans ce cas là, d'une erreur cachée provoquant un dérèglement à la fois syntaxique et sémantique. Dans le cas contraire, le deuxième groupe d'agents continue de balayer toute la phrase à la recherche d'une autre erreur. Deux cas se présentent : si ce dernier n'a pas trouvé alors il informe le premier groupe, qui, lui a déjà trouvé une erreur, pour qu'il déclenche la correction. Si par contre, il a trouvé une autre erreur, il l'envoie à son tour au premier groupe demandant une confirmation. Si les deux groupes se mettent d'accord sur cette nouvelle erreur, alors c'est cette dernière erreur qui est envoyée pour être corrigée à la place de la première erreur. Enfin, s'il les deux groupes ne se sont pas mis d'accord sur aucune erreur chacun envoie de son côté sa propre présumée erreur à l'agent Correction.

5.4 L'agent Correction

Finalement, l'agent Correction vient corriger les fautes détectées par les deux vérificateurs : syntaxique et sémantique. Il procède alors par la génération de toutes les formes proches de la forme erronée, à une différence d'édition près pour former ainsi une liste contenant les candidats à la correction. Ladite liste est assez longue. Elle contient en moyenne plus de 27 formes candidates et peut atteindre 185 formes : valeur pouvant être augmentée par l'agglutination des enclinomènes (Ben Othmane Zribi, 1998). Pour réduire ce grand nombre de propositions, l'agent Correction substitue la forme erronée par chacune des formes proposées et forme ainsi un ensemble de phrases candidates. Ces dernières seront réinjectées, au fur et à mesure de leur production, dans les deux vérificateurs (autrement dit la partie détection du système). Celles qui contiennent toujours des anomalies seront éliminées et c'est le même sort que subissent les propositions qui leur sont respectives. La liste des propositions restantes est par la suite triée par ordre de pertinence et présentée à l'utilisateur.

6 Expérimentation et résultats

Notre objectif étant la réalisation d'un système capable de détecter et de corriger les erreurs cachées, nous avons alors implémenté à ce stade de notre travail une partie du système, déjà conçu, à savoir le groupe syntaxique d'agents et intégré l'agent Correction de (Ben Othmane Zribi, 1998).

De plus, pour évaluer le système ainsi réalisé, nous avons besoin d'un corpus textuel contenant suffisamment d'erreurs cachées. Chacune de ces dernières doit être identifiée avec sa forme corrective. Toutefois, faute de corpus contenant ce genre d'erreurs sous sa forme naturelle, nous avons choisi de créer notre propre corpus manuellement. Nous avons, ainsi, généré parmi les formes qui existent dans le corpus de test une liste d'erreurs cachées tout en respectant les hypothèses restrictives de nos champs d'investigation. Ce corpus, qui constituera l'entrée de notre système, contient environ 750 formes textuelles non voyellées, dans lesquelles nous avons introduit 100 erreurs cachées du type syntaxique.

6.1 Résultats de l'évaluation de la détection

L'expérimentation de notre système de détection des erreurs cachées a donné des résultats que nous jugeons satisfaisants avec un pourcentage de précision égal à **80%** et un pourcentage de rappel égal à 77%. La présence de bruit 20% (1- Précision) et de silence 23% (1 – Rappel) s'expliquer principalement par les causes citées ci-dessous :

- La largeur de la portée de vérification dans les phrases manipulées. En effet, malgré la phase de segmentation, le nombre de mots constituant certaines phrases reste important ce qui contrarie les principes de vérification de certains agents qui travaillent à base de phrases courtes.

- La compétition entre agents, qui a été l'une des principales hypothèses de notre choix d'une architecture multi-agent et qui s'est manifestée par le fait que le premier agent détecteur de faute arrête ses collègues et ce, sans savoir si la faute détectée est ou n'est pas une fausse alarme.

- La non exhaustivité de nos règles linguistiques et l'absence de certaines informations linguistiques. Les notions d'"animé" ou d'"inanimé", pourraient par exemple aider à effectuer une meilleure vérification syntaxique ou sémantique.

6.2 Résultats de l'évaluation de la correction

Cette phase a été testée à deux niveaux ; d'abord après l'obtention de toutes les propositions de correction, ensuite après la minimisation de la liste de ces propositions. Les résultats obtenus sont illustrés dans le tableau ci-après.

	Couverture	**Précision**	**Ambiguïté**	**Proposition**	**Position**
Initialement	100%	100%	100%	82,5	8,7
Après minimisation	93,3%	86,6%	86,6%	18,4	2,8

Figure 2 Evaluation du correcteur des erreurs cachées

Nous remarquons que notre méthode de minimisation de la liste des propositions a permis de diminuer, considérablement, le nombre moyen des propositions de 77% (de 82,5 à **18,4** propositions en moyenne). Cette diminution, bien qu'elle ait réduit l'ambiguïté de notre correcteur de 13,4%, ne s'est pas passée sans dégâts. Elle s'est faite aux dépens de la couverture (diminution de 6,4%) et de la précision (diminution de 13,4%).

7 Conclusions et perspectives

La partie du système ainsi implémentée a donné des résultats assez satisfaisants. Les choix que nous avons adoptés nous ont permis d'atteindre nos objectifs initialement dressés. Cependant, nous estimons que les résultats obtenus peuvent être encore améliorés d'abord par l'amendement des règles linguistiques utilisées et ensuite par la prise en considération des informations sémantiques. Aussi, l'implémentation du groupe sémantique d'agents figure parmi les perspectives proches de nos travaux de recherche.

Références

ALOULOU C. (1996), Utilisation de l'approche multi-critère pour orienter un processus de correction des erreurs d'accord dans des phrases de la langue arabe non voyellée, Mémoire de DEA, Institut Supérieur de Gestion, Université de Tunis III.

BEN HAMADOU A. (1993), Vérification et correction automatique par analyse affixale des textes écrits en langue naturelle : le cas de l'arabe non voyellé, Thèse d'état en informatique, Faculté des Sciences de Tunis.

BEN OTHMANE ZRIBI C. (1998), De la synthèse lexicographique à la détection et à la correction des graphies fautives arabes, Thèse de doctorat, Université de Paris XI, Orsay.

BEN OTHMANE ZRIBI C. et BEN AHMED M. (2003), Le contexte au service de la correction des graphies fautives arabes, *TALN'03*, Nantes.

BIGERT J., KNUTSSON O. (2002), Robust Error Detection : A Hybrid Approach Combining Unsupervised Error Detection and Linguistic Knowledge, In *Proceedings of Robust Methods in Analysis of Natural Language Data (ROMAND'02)*, Frascati, Italie.

DEBILI F., ACHOUR H., SOUISSI E. (2002), La langue arabe et l'ordinateur: De l'étiquetage grammatical à la voyellation automatique, *Correspondances N°71*, Lyon.

GOLDING A. R. (1995), A bayesian hybrid method for context- sensitive spelling correction, In *Proceedings of the Third Workshop on Very Large Corpora*, Cambridge, Massachusetts, USA, pages 39-53.

GOLDING A. R., SCHABES Y. (1996), Combining trigram-based and feature-based methods for context-Sensitive Spelling Correction, *ACL'96*, San Fransisco.

GOLDING A. R., DAN R., (1999), A winnow-based approach to context-sensitive spelling correction, *Machine Learning*, 34(1-3), 107-130.

VERBERNE S. (2002), Context sensitive spell checking based on word trigram probabilities, Mémoire de Mastère, Université de Nijmegen.

XIAOLONG W., JIANHUA L. (2001), Combine trigram and automatic weight distribution in Chinese spelling error correction, *Journal of computer Science and Technology, Volume 17 Issue 6*, Province, China.