Un système de recouvrement des ellipses de la langue arabe

Kais Haddar & Abdelmajid Ben Hamadou
FSEG Sfax B.P.1088 Sfax 3018 Tunisie

Résumé. Le phénomène d'ellipse est particulièrement difficile à traiter d'une manière automatique car il n'est pas toujours facile de pouvoir caractériser, les éléments que l'on est en droit de supprimer. Dans le présent article, nous proposons une caractérisation formelle du phénomène d'ellipse et un algorithme de localisation des parties elliptiques d'une phrase. Nous présentons ensuite un certains nombre de critères de classification des phrases elliptiques pour préparer la phase de recouvrement proprement dit. Enfin, nous avons élaboré des algorithmes de recouvrement des phrases elliptiques complexes. L'heuristique proposée et la classification des phrases elliptiques complexes nous ont permis aussi de traiter l'effet de l'interaction entre ellipse et anaphore.

1 Introduction

L'ellipse est un phénomène linguistique qui se manifeste par l'omission d'une partie d'un énoncé qui n'a pas une incidence sur la compréhension globale. L'objectif étant d'alléger la formulation et d'éviter la redondance. Le problème des ellipses est particulièrement difficile à traiter d'une manière automatique, car il peut apparaître en même temps avec d'autres phénomènes linguistiques complexes comme les conjonctions et les comparatifs [14]. Aussi, il est quasiment impossible, de caractériser les éléments que l'on est en droit de supprimer. En première approche, à part les constituants principaux de la phrase, tout type de structure complémentaire peut être élidé, mais on rencontre toujours des contres-exemples et aucune règle générale n'a été mise en évidence.

L'analyse syntaxique peut se faire sur la phrase en temps qu'unité ou en décomposant cette phrase en propositions avant de procéder à l'analyse (i.e. une analyse syntaxique de propositions). Notre approche se base sur l'utilisation d'une grammaire de propositions et par conséquent sur la décomposition de la phrase en propositions. L'originalité de cette approche réside dans le fait qu'elle traite des phrases elliptiques complexes et qu'elle permet leur classification. Cette classification est utilisée pour construire des algorithmes efficaces de recouvrement [6].

Dans cet article, nous nous intéressons au traitement automatique du phénomène d'ellipse dans les corpus écrits de la langue arabe. Nous commençons par présenter des travaux liés en les comparant avec notre approche. Nous introduisons l'analyseur syntaxique de propositions et la typologie des ellipses de la langue. Ensuite, nous proposons une caractérisation formelle des phrases elliptiques et un algorithme de localisation des parties elliptiques basé sur cette caractérisation. Nous donnons une méthode de classement des phrases elliptiques basée sur l'occurrence des types d'ellipses rencontrés. Cette étude servira ultérieurement comme base pour la phase de recouvrement proprement dit des phrases elliptiques. Enfin, nous élaborons des algorithmes de recouvrement des phrases elliptiques complexes. L'heuristique proposée et la classification des phrases elliptiques complexes nous ont permis aussi de traiter l'effet de l'interaction entre ellipse et anaphore tout en validant les idées et la démarche présentées dans cet article

2 Travaux liés

Diverses méthodes ont été utilisées pour traiter les ellipses dans leurs différentes formes et dans des contextes différents. Citons particulièrement:

- l'extension de la grammaire en ajoutant des règles explicites ou des méta-règles qui prévoient les ellipses [5], [9], [15].
- le relâchement des contraintes [3],[16]: Sur un échec, l'analyseur peut être relancé en négligeant les contraintes du langage.
- l'appariement approximatif [10]: Il s'agit d'établir une correspondance approximative entre la phrase elliptique à traiter et un sous ensemble de règles de la grammaire.

La théorie linguistique traditionnelle recense les différentes formes d'ellipse suivant la structure syntaxique des propositions elliptiques [4], [5]. Les travaux syntaxiques concernant ces formes proposent le copiage de la structure syntaxique à partir de la représentation de la proposition bien formée vers celle de la proposition elliptique [5],[8], [11]. Les approches utilisées sont généralement basées sur une grammaire de phrases complexes qui ne peut pas localiser facilement les parties elliptiques d'une phrase. Dans ces approches, l'occurrence de la proposition elliptique ne joue aucun rôle. Dans notre approche, par contre, nous distinguons des classes d'ellipse suivant la structure syntaxique des phrases elliptiques en tenant compte des considérations suivantes: l'occurrence des propositions elliptiques de la phrase, l'alternance de proposition bien formée/proposition elliptique et la forme d'ellipse.

A l'opposé des approches mentionnées, notre approche est basée sur un autre type de grammaire à savoir une grammaire d'ATN de propositions que nous utilisons pour identifier la proposition bien formée et pour localiser les propositions elliptiques.

3 Analyseur syntaxique de propositions

L'analyseur syntaxique peut être réaliser en utilisant une grammaire de phrases complexes. Mais la spécification d'une telle grammaire est compliquée et a besoin d'efforts considérables. Il est difficile, en outre, de localiser les parties elliptiques de la phrase. Ainsi, nous préconisons l'utilisation d'une grammaire de propositions, ce qui nous permet d'abord de minimiser l'effort nécessaire pour écrire une grammaire de phrases, ensuite, nous pouvons facilement identifier la proposition bien formée d'une phrase elliptique et aussi localiser les propositions elliptiques et enfin, nous pouvons détecter les fragments elliptiques d'une proposition elliptique. Comme dans [12], nous ne considérons pas un fragment d'une phrase d'exception comme étant une forme d'ellipse, mais plutôt comme un modificateur NP déplacé.

La grammaire de propositions que nous utilisons pour notre approche est une grammaire contexte sensitive de propositions. Pour l'implémenter, nous construisons un réseau de transitions récursifs (RTN) [17]. Ensuite, nous le transformons en un réseau de transitions augmentés (ATN) afin d'ajouter le principe de sous-catégorisation et les contraintes d'accord qui peuvent faciliter le processus de résolution et permettent d'éviter l'ambiguïté.

Les principales composantes de notre analyseur syntaxique sont:

- Un ensemble de catégories lexicales pour les mots de la langue Arabe et un lexique dans lequel chaque mot est associé à un certain nombre de catégories [1];
- Une grammaire Arabe de propositions utilisant les mêmes catégories que le lexique et tout constituant nécessaire pour spécifier les structures des propositions bien formées;
- Un programme d'analyse syntaxique dont les données sont la proposition à analyser, le lexique et la grammaire Arabe de propositions et dont la sortie est la description structurelle de la proposition donnée en entrée.

4 La typologie des ellipses de la langue arabe

Comme nous l'avons déjà indiqué ci-dessus, l'ellipse se manifeste par l'absence d'un ou plusieurs mots d'une phrase qui reste cependant compréhensible. Ces mots seraient nécessaires uniquement pour la construction régulière de la phrase. Le sens de la phrase est complété en général en se référant à une partie du discours antérieur.

Exemple

(1) استعملت الحاسوب كذلك [استعمل] بقية الطلبة [الحاسوب]

J'ai utilisé l'ordinateur et les autres étudiant aussi.

Les mots en crochets sont les mots omis.

Nous donnons à présent une typologie des ellipses de la langue arabe inspirée de celle proposée dans [11], [13] et [14].

4 -1 Forme figée

Certaines expressions, et notamment les formules de vœux et de politesse, certains ordres ou certaines exclamations sont elliptiques. Pour comprendre ces expressions il n'est pas nécessaire de reconstruire une forme complète. C'est pourquoi, elles sont appelées ‘‘fausses ellipses’’.

Exemples

(2) [أتمنى لك] عيدا سعيدا *Bonne année*

(3) [أحذر] النار, النار *Au feux, au feux*

Ces formes figées d'ellipse ne nous intéressent pas car elles peuvent être résolues au niveau lexical. Mais, nous nous intéressons plutôt aux ellipses syntaxiques et sémantiques qui obligent le lecteur à chercher, dans le contexte, les éléments qui manquent et sans lesquels le message serait incompréhensible. La phrase (1) en est un exemple.

4 -2 Ellipse nominale

L'ellipse nominale se manifeste par l'omission de la partie essentielle d'un syntagme nominal (i.e., la tête) qui sera cherchée dans une partie du discours antérieur. Donc, dans une phrase qui contient une ellipse nominale l'un des syntagmes nominaux est incomplet.

Exemple

(4) الأخ الأكبر له من العمر عشرسنين و [الأخ] الأصغر ستة سنوات

Le frère aîné a dix ans et le plus jeune a six ans.

Dans l'exemple (4), le syntagme nominal contenant l'ellipse est représenté par un article et un adjectif (le plus jeune). L'exemple traité renferme une ellipse du nom. Ainsi, l'ellipse syntaxique dans la langue arabe est principalement une ellipse du nom.

4 -3 Ellipse d'un syntagme entier

L'ellipse d'un syntagme entier se distingue de l'ellipse nominale par la nature du constituant omis qui peut être tout un syntagme (nominal ou verbal). Ce syntagme sera cherché dans un contexte précédent. Ainsi dans cette même catégorie, nous pouvons distinguer les formes suivantes:

- l'ellipse du sujet:

(5) اكتشف الشرطي اللصوص فهربوا [اللصوص]

Le policier découvre les voleurs et [les voleurs] s'enfuient.

- l'ellipse du verbe:

(6) دخل الاستاذ ف[دخل] الطلبة

Le professeur est entré puis les étudiants [sont entrés].

- l'ellipse du verbe et du sujet:

(7) اكلت البنت تفاحا ثم [اكلت البنت] اجاصا

La fille a mangé des pommes puis [La fille a mangé] des poires.

- l'ellipse du verbe et du complément:

(8) اكلت البنت تفاحا و كذلك [اكل] أخوها [تفاحا]

La fille a mangé des pommes et son frère [a mangé des pommes]aussi.

5 Quelques définitions et notations

Compte tenu de ce qui a été exposé, nous pouvons considérer qu'une phrase elliptique est constituée de deux parties distinctes: une partie elliptique précédée par une partie bien formée pouvant être utilisée comme référence pour résoudre l'ellipse. la partie elliptique peut être constituée par une succession de propositions elliptiques généralement séparées par des mots outils de liaison (que nous appelons connecteurs) qui sont des conjonctions de coordination, de subordination et d'opposition. Pour caractériser d'une manière formelle ce phénomène, nous avons besoin d'introduire quelques notations et définitions.

5 -1 Proposition elliptique

Soit V un alphabet fini. Soit V^* l'ensemble de mots généré par V. $V^+ = V^* \backslash \{\varepsilon\}$ où ε est le mot vide.

Définition 1. Nous appelons proposition toute séquence finie P de mots de V^+, $P=w_1w_2...w_m$, $w_i \in V^+$ et $1 \le i \le m$. Nous notons $\mathcal{P}$ l'ensemble de toutes les propositions qu'il est possible de construire à partir de V^+.

Soit G une grammaire sur V^+ décrivant ainsi un sous-ensemble $L(G) \subseteq P$. Pour vérifier si une proposition donnée P est générée par G ou non, nous introduisons la fonction suivante:

$$analyse : \mathcal{P} \rightarrow \mathcal{B} = \{true, false\}$$

$$P \mapsto \begin{cases} true \ si \ P \in L(G) \\ false \ sinon. \end{cases}$$

Définition 2. Une proposition P est dite bien formée si elle appartient à L(G).

P *bien formée* $\Leftrightarrow$ *analyse*(P)

Notons $\mathcal{S}$ l'ensemble des séquences finies de propositions de P.

Définition 3. Soit $S=P_0P_1...P_n \in \mathcal{S}$. Nous définissons le contexte d'une proposition P_i, $1 \le i \le n$ par la sous-séquence $S_i=P_0P_1...P_{i-1}$ formée de propositions qui précèdent P_i. Nous notons par W^{Si} l'ensemble des mots constituant S_i.

Remarque. Le contexte de P_0 est la séquence vide.

Définition 4. Soient $S=P_0P_1...P_n$ une séquence de $\mathcal{S}$ et $P_i = w_{i1}w_{i2}...w_{im}$ une proposition de $\mathcal{P}$, $0 \le i \le n$. Soit la fonction *is_recoverable* définie par:

$$is\_recoverable(P_i,S)=\begin{cases} true \;\; si \neg analyse(P_i) \wedge W^{Si} \neq \varnothing \wedge \exists w'_{i1},\ldots,w'_{i(m+1)} \in W^{Si} \cup \{\varepsilon\} \\ \qquad\qquad \wedge analyse(w'_{i1}w_{i1}w'_{i2}\ldots w'_{im}w_{im}w'_{i(m+1)}) \\ false\ sinon. \end{cases}$$

Une proposition P_i,$1 \leq i \leq n$, est dite elliptique si *is_recoverable*(P_i,S).

Donc, une proposition P_i d'une séquence S est elliptique si elle n'est pas bien formée et si elle peut être complétée par des mots de son contexte pour en devenir une. Notons que cette définition n'exclue pas certaines phrases non elliptiques comme par exemple la phrase en anglais (*he is going out and it raining*). La proposition (*it raining*) est traité ici comme une proposition elliptique. Ce problème n'a pas d'incidence sur la rigidité de notre définition ni sur le processus de détection puisque nous restreindrons notre champ de recherche aux corpus corrects.

5 -2 Phrase elliptique

Soit C un sous-ensemble de V^* que nous appelons ensemble de connecteurs contenant aussi le mot vide ε.

Définition 5. Une phrase Ph est une séquence de propositions séparées par des connecteurs de C.

$Ph = P_0c_1P_1...c_nP_n$ où $c_i \in C$, $P_i \in P$.

Définition 6. Une phrase Ph est dite elliptique si
1) la première proposition de Ph est bien formée
2) l'une au moins des propositions restantes est elliptique.

Exemple 1. Soit la phrase Ph:

يجب ربط السلك الازرق بالنهائي أ والسلك الاصفر بالنهائي ب او السلك الاسود بالنهائي ج.

Le câble bleu doit être connecté au terminal A et le câble jaune au terminal B ou le câble noir au terminal C.
P_0: Le câble bleu doit être connecté au terminal A.
P_1: le câble jaune au terminal B.
P_2: le câble noir au terminal C.
Ph est elliptique car P_0 est une proposition bien formée et P_1 et P_2 sont elliptiques.

5 -3 Classification des phrases elliptiques

Nous allons maintenant nous intéresser aux phrases elliptiques et plus particulièrement à leur construction en essayant de dégager des classes sur lesquelles nous pouvons appliquer des algorithmes de recouvrement.

Définition 7. Une phrase elliptique $Ph_e = P_0c_1P_1...c_nP_n$, $n \geq 1$, est dite homogène si toutes les propositions P_i, $1 \leq i \leq n$, sont elliptiques et ont la même forme d'ellipse.

Exemple 2. Reprenons la phrase elliptique de l'exemple 1. Puisque P_1 et P_2 ont la même forme d'ellipse Ph est donc homogène.

Définition 8. Une phrase elliptique $Ph_e=P_0c_1P_1...c_nP_n$, $n>1$, est dite hétérogène si chaque deux propositions elliptiques successives P_i et P_{i+1}, $1 \leq i < n$, n'ont pas la même forrme d'ellipse.

Remarque importante. Une phrase elliptique ne contenant pas deux propositions elliptiques successives est systématiquement *hétérogène* quelque soit le type d'ellipse des propositions qu'elle renferme.

Exemple 3. Soit la phrase Ph:

يجب ربط السلك الازرق بالنهائي أ والسلك الاصفر بالنهائي ب اويجب عليك ترك النهائيات مفتوحة.

Le câble bleu doit être connecté au terminal A et le câble jaune au terminal B ou tu dois laisser les terminaux libres.
Ph est hétérogène car P_0 et P_2 sont bien formées et P_1 est elliptique.

Définition 9. Soit $Ph_e= P_0c_1P_1...c_nP_n$, $n>1$ une phrase elliptique. Ph_e est dite mixte si elle n'est ni *homogène*, ni *hétérogène*.

Exemple illustratif. Nous notons ici P_i : une proposition non elliptique; P_i^{tj} : une proposition elliptique de type t_j. Soit:

$Ph_1 = P_0c_1P_1^{t1}c_2P_2^{t1}$,
$Ph_2 = P_0c_1P_1^{t1}c_2P_2^{t2}$,

$Ph_3 = P_0 c_1 P_1^{t1} c_2 P_2$,
$Ph_4 = P_0 c_1 P_1^{t1} c_2 P_2^{t1} P_3 c_4 P_4^{t2}$.

Selon les définitions 8, 9 et 10, Ph_1 est homogène, Ph_2 et Ph_3 sont hétérogènes et Ph_4 est mixte.

Toutes les définitions que nous venons de donner (i.e., de 1 à 10) , vont nous permettre d'établir des critères à appliquer pour localiser les propositions elliptiques et pour déterminer la classe d'appartenance des phrases elliptiques en vue d'effectuer ultérieurement le recouvrement proprement dit.

6 La méthode de détection

Conformément à la caractérisation formelle des ellipses de la langue arabe, la méthode de détection que nous proposons, s'effectue en trois principales étapes: la recherche des connecteurs, l'identification de la proposition bien formée de référence et l'étiquetage des propositions restantes. Ces trois étapes permettent donc de localiser les propositions elliptiques dans la phrase analysée. Elles préparent aussi l'étape de classification.

- **Etape 1:** Rechercher tous les connecteurs existants dans la phrase analysée en se basant sur un lexique des connecteurs. Le résultat obtenu est constitué d'une liste des propositions et d'une autre contenant les connecteurs.
- **Etape 2:** Effectuer un filtrage à partir des listes de propositions et de connecteurs déjà dégagées basé sur l'identification de la plus longue proposition bien formée de la phrase. Une fois la proposition bien formée est identifiée, les deux listes sont mises à jour d'une manière définitive en supprimant des deux listes les composantes de la proposition bien formée.
- **Etape 3:** Attribuer à chaque proposition de la liste des propositions restantes une étiquette signifiant que la proposition est elliptique ou non. L'étiquetage permet de localiser les propositions elliptiques de la liste des propositions restantes.

Ainsi, pour qu'une phrase soit elliptique, il faut qu'elle commence par une proposition bien formée et qu'elle contient au moins une proposition marquée par *etiquette* avec la marque 0 [7].

7 Détermination de la classe de la phrase elliptique

D'après les définitions 7, 8 et 9, nous avons besoin pour déterminer la classe d'une phrase elliptique, de connaître les formes des ellipses des propositions elliptiques qui se succèdent. Rappelons que dans le cas où la phrase ne contient pas des propositions elliptiques qui se succèdent, elle est systématiquement hétérogène et donc nous n'avons pas besoin de connaître les formes d'ellipses des propositions (elliptiques). La détermination de la forme d'ellipse d'une proposition (elliptique) n'est pas toujours une opération facile à mettre en œuvre. C'est pourquoi nous proposons une heuristique nettement plus facile à mettre œuvre, applicable sur les propositions elliptiques successives. Cette heuristique se base sur le comportement des connecteurs: quand certains connecteurs de la langue arabe se succèdent dans une même phrase, les propositions qui les suivent ont toujours le même type d'ellipse (si elles sont elliptiques).

Pour représenter cette heuristique, nous proposons l'utilisation d'une matrice M[k,k] sur l'ensemble {0,1}, où k = card(C), appelée ''Matrice de cohérence des connecteurs'' et définie de la manière suivante.

Soit $c_i , c_j \in C$.

$M[c_i,c_j] = 1$ signifie que la proposition qui suit c_i, si elle est elliptique, contient le même type d'ellipse que la proposition qui suit c_j dans une phrase donnée où ces deux propositions se succèdent.

$M[c_i,c_j] = 0$ autrement.

Ceci nous permet de définir une fonction qui partage l'ensemble des connecteurs d'une phrase elliptique donnée en partitions. Le nombre de ces partitions détermine la classe d'appartenance de la phrase elliptique [7].

8 Processus de recouvrement

La démarche que nous proposons, permet le recouvrement des phrases elliptiques appartenant aux différentes classes déjà présentées dans le paragraphe précédent.

Partant de la définition 6 qui considère qu'une phrase elliptique Ph_e a la forme suivante $Ph_e=P_0 c_1 P_1 c_2 P_2 ... c_n P_n$ avec $n \geq 1$ et les deux conditions correspondantes, nous pouvons introduire deux types de phrases elliptiques: *la phrase elliptique simple* si n = 1 et *la phrase elliptique complexe* si n > 1.

Nous pouvons distinguer deux types de recouvrement:

- le recouvrement des phrases elliptiques simples et
- le recouvrement des phrases elliptiques complexes construit à partir du premier recouvrement.

8 -1 Recouvrement des phrases elliptiques simples

La forme complète d'une phrase elliptique simple est restituée par le parallélisme de construction entre la proposition bien formée P_{bf} et la proposition elliptique P_e. Pour ce faire, nous appliquons des règles de réduction

dépendant de la grammaire de propositions utilisée pour obtenir une structure de la proposition plus au moins élémentaire et des règles explicites intégrant des contraintes locales. Notons que, l'utilisation d'une grammaire de propositions et l'intégration des contraintes rendent l'utilisation des règles explicites plus efficace.

L'algorithme de recouvrement des phrases elliptiques simples que nous proposons, est constitué de deux étapes principales:

- **Etape 1:** Transformer une phrase $m_1\ m_2\ ...\ m_n$ en une chaîne de couples $(m_1,c_1)(m_2,c_2)\ ...(m_n,c_n)$ où m_i est un mot et c_i est la catégorie lexicale correspondante. A chaque catégorie sont reliés les caractéristiques morpho-syntaxiques (genre, nombre et aspect humain/non-humain) s'il s'agit d'un nom et (temps, personne, genre, nombre et aspect humain/non-humain) s'il s'agit d'un verbe. Ensuite, transformer la chaîne $(m_1,c_1)(m_2,c_2)\ ...(m_n,c_n)$ en une autre chaîne de couples $(M_1,C_1)(M_2,C_2)...(M_k,C_k)$ avec $k \leq n$ où M_i est un syntagme et C_i est la catégorie syntaxique qui lui correspond.
- **Etape2:** Appliquer des règles explicites sur la chaîne $(M_1,C_1)\ (M_2,C_2)\ ...\ (M_k,C_k)$ qui correspond à P_{bf} et sur la chaîne $(M'_1,C'_1)\ (M'_2,C'_2)\ ...\ (M'_k,C'_k)$ qui correspond à P_e en tenant compte du type du connecteur ainsi que des contraintes locales pour compléter la phrase elliptique [6].

Exemples de règles explicites relatives à l'ellipse du sujet

R1: P_{bf}: SV + SN_{Pbf}	P_{bf}: SV + SN_{Pbf}
⇒	
P_e: SV / SV + SN	P_e: SV + SN_{Pbf} / SV + SN_{Pbf} + SN

R2: P_{bf}: SV + $SN_{Pbf(1)}$ + $SN_{Pbf(2)}$	P_{bf}: SV + $SN_{Pbf(1)}$ + $SN_{Pbf(2)}$
⇒	
P_e: SV / SV + SN	P_e: SV+$SN_{Pbf(2)}$ / SV+$SN_{Pbf(2)}$ + SN
c: " ف "	
Contrainte: $SN_{Pbf(2)}$, SV_{Pe} s'accordent en genre et en nombre	

Exemple 1. Soit la phrase suivante: " استيقظ علي فشرب حليبه "

Ali s'est levé et a bu son lait

L'algorithme de détection décompose la phrase en:

P_{bf}: " استيقظ علي ", P_e: " شرب حليبه ", c: " ف ".

C'est un cas d'ellipse du sujet. A l'aide de l'algorithme de recouvrement, la phrase sera complétée comme suit en utilisant la règle R1 définie ci-dessus:

" استيقظ علي فشرب (علي) حليبه ".

Dans certains cas d'ellipse du sujet, il est difficile de décider lequel des syntagmes nominaux de P_{bf} doit être choisi pour compléter P_e. Pour lever cette ambiguïté, nous appliquons des contraintes relatives aux accords en genre, en nombre, en personne et en trait humain/non-humain. L'exemple suivant illustre ce problème.

Exemple 2. Soit la phrase suivante: "اكتشف الشرطي اللصوص فهربوا"

Le policier découvre les voleurs qui s'enfuient

L'algorithme de détection décompose la phrase en:

P_{bf}: " اكتشف الشرطي اللصوص ", P_e: " هربوا ", c: " ف "

C'est un cas d'ellipse du sujet. A l'aide de l'algorithme de recouvrement, la phrase sera complétée par le choix du syntagme "اللصوص" (*les voleurs*) à la place du syntagme "الشرطي" (*le policier*) en appliquant les contraintes d'accords en genre et en nombre intégrées dans la règle R2: " اكتشف الشرطي اللصوص فهربوا (اللصوص) ".

8 -2 Recouvrement des phrases elliptiques complexes

L'algorithme de recouvrement des phrases elliptiques complexes est une extension de l'algorithme de recouvrement des phrases elliptiques simples. Nous distinguons, en effet trois types de recouvrement: le recouvrement par propagation, le recouvrement en cascade et le recouvrement par alternance.

8 -2 -1 Recouvrement par propagation

Il est appliqué aux phrases elliptiques homogènes. Il consiste à appliquer le recouvrement simple sur la proposition bien formée, le premier connecteur et la première proposition elliptique, ensuite à compléter les autres propositions elliptiques restantes avec le même fragment trouvé.

8 -2 -2 Recouvrement en cascade

Il est appliqué aux phrases elliptiques hétérogènes. Il consiste tout d'abord à appliquer le recouvrement simple sur la première proposition elliptique et le connecteur correspondant en prenant comme référence la proposition bien formée qui la précède ensuite à considérer la proposition recouvrée comme réfèrence pour le recouvrement de la proposition suivante si elle est elliptique. Dans le cas où cette dernière n'est pas elliptique, elle sera utilisée comme référence pour la proposition suivante et ainsi de suite jusqu'à atteindre la fin de la phrase.

8 -2 -3 Recouvrement par alternance

L'algorithme de recouvrement par alternance est appliqué aux phrases elliptiques mixtes. Rappelons d'abord qu'une phrase elliptique mixte contient au moins une séquence de propositions elliptiques ayant partout la même forme d'ellipse, et qu'à la suite de cette séquence nous rencontrons nécessairement soit une proposition bien formée, soit une proposition elliptique ayant une forme d'ellipse différente. L'idée de l'algorithme consiste à effectuer le recouvrement par propagation sur chaque séquence de propositions elliptiques ayant la même forme d'ellipse. Si après le recouvrement de toutes les propositions d'une séquence nous rencontrons une proposition bien formée, alors nous faisons appel à l'algorithme de recouvrement en cascade et nous prenons la proposition bien formée comme référence pour le recouvrement de la proposition suivante. Si au contraire la proposition rencontrée est elliptique, d'une forme d'ellipse différente, la dernière proposition recouvrée de la séquence sera considérée comme référence pour le recouvrement en cascade de cette proposition elliptique.

8 -2 -4 Recouvrement des ellipses en présence d'anaphores

Rappelons que l'interaction entre ellipse et anaphore se présente lorsque la proposition bien formée P_{bf} d'une phrase elliptique Ph_e possède un syntagme nominal complément contenant un pronom personnel affixe et que ce syntagme est omis dans une ou plusieurs propositions elliptiques qui suivent P_{bf}. Donc, les formes d'ellipse dont le recouvrement peut être influencé par l'anaphore sont l'ellipse du verbe et du complément ou l'ellipse du verbe et du deuxième complément. Si nous mettons en considération ce phénomène d'interaction, le recouvrement des ellipses en présence d'anaphores devient complexe et ambigu car l'interprétation de l'anaphore d'une proposition recouvrée d'une phrase elliptique n'est pas toujours évidente. En effet, il existe deux interprétation (ou lectures) possibles de l'anaphore: la lecture stricte et la lecture souple. La première signifie que le référent pronominal (de la proposition bien formée), après recouvrement, est remplacé dans la proposition elliptique par le syntagme nominal sujet de la proposition bien formée. La deuxième signifie que le référent pronominal, après recouvrement, est remplacé dans la proposition elliptique par le syntagme nominal sujet de cette proposition. Quand il y a un seul pronom dans la phrase elliptique ceci résulte en une simple stricte/souple ambiguïté dans l'interprétation pronominale. Cependant, en ajoutant d'autres pronoms les possibilités croissent et une autre interprétation sera possible: c'est l'interprétation mixte.

Plusieurs méthodes ont été envisagées pour traiter l'ellipse en présence d'anaphore. Citons en particulier les travaux de Dalrymple, shieber et Pereira [4] qui utilisent les formes logiques (le formalisme du λ-calcul) pour caractériser ce phénomène et autres travaux préconisent la combinaison de la théorie des dépendances et la théorie d'indexation comme moyen pour résoudre les possibilités d'interprétation des pronoms dans des phrases elliptiques sans se référer à une représentation sémantique. Tous ces travaux traitent ce problème du point de vue sémantique contrairement à notre travail qui essaye de traiter le problème de point de vue plutôt syntaxique.

Dans notre algorithme de recouvrement, nous avons tenu compte des deux lectures en présence d'anaphores lorsque la phrase est elliptique simple et des trois lectures lorsque la phrase est elliptique complexe. En effet, l'algorithme fait le choix entre une interprétation stricte, souple ou une interprétation mixte selon le cas. Ce choix d'interprétations se base sur deux facteurs qui sont le type de connecteur et la classe de la phrase elliptique (homogène, hétérogène ou mixte). Les phrases elliptiques homogènes, par exemple, ne peuvent pas avoir des interprétations mixtes contrairement aux autres classes. Pour ce faire, nous avons opté pour l'extension des règles explicites de recouvrement de telle façon qu'elles mettent en considération une des deux lectures (stricte ou souple). Cette démarche nous permet, d'une part, d'exploiter les règles déjà conçus sans recours à d'autres formalismes orientés au traitement sémantique et, d'autre part, de mettre en oeuvre facilement ces modifications.

Les règles explicites de recouvrement concernant l'ellipse du verbe et du complément, l'ellipse du verbe, du sujet et du complément et l'ellipse du verbe et du deuxième complément sont les règles qui peuvent être étendue de telle façon qu'elles mettent en considérations une des deux lectures (stricte ou souple).

Prenons l'exemple d'interaction suivant sur lequel nous pouvons appliquer une règle explicite:

(9) رأى الولد أمّه وكذلك الصديق.

Le garçon a vu sa mère et l'ami aussi.

Avec la lecture stricte, nous obtenons (9'):

(9') رأى الولد أمّه وكذلك رأى الصديق أمّ الولد.

Le garçon a vu sa mère et l'ami a vu la mère du garçon *aussi.*

Avec la lecture souple, nous obtenons (9''):

(9'') رأى الولد أمّه وكذلك رأى الصديق أمّه.

Le garçon a vu sa mère et l'ami a vu sa mère *aussi.*

Notons ici que les deux lectures sont possibles du point de vue sémantique. C'est un cas de préférence sémantique. Mais, nous pouvons avoir des exemples où une seule lecture est possible; l'autre étant absurde du point de vue sémantique ou logique tel que:

(10) باع أحمد منزله وكذلك منير.

Ahmed a vendu sa maison et Mounir aussi.

Seule la lecture souple (10') est valable. La lecture stricte est absurde car le fait de vendre une maison à une date donnée ne peut s'effectuer qu'une seule fois par une personne qualifiée, l'action est unique.

(10') باع أحمد منزله وكذلك باع منير منزله.

Ahmed a vendu sa maison et Mounir a vendu sa maison aussi.

Soit la phrase suivante:

(11) أعطى عليّ صورته لمريم و صالح لسارة.

Ali a donné sa photo à Myriam et Salah à Sarah.

(11') أعطى عليّ صورة عليّ لمريم و (أعطى) صالح (صورة صالح) لسارة.

Ali a donné la photo d'Ali à Myriam et Salah a donné la photo de Salah à Sarah.

La seconde proposition de (11) est ambiguë entre Salah a donné sa photo (d'Ali) à Sarah et Salah a donné sa photo (de Salah) à Sarah. L'intuition est que Ali est parallèle avec Salah et à Myriam avec à Sarah. Avec l'interprétation souple, la phrase est recouvrée de la manière de (11'). Cet exemple représente un cas de parallèlisme. La notion de parallélisme est sémantique puisqu'elle dépend du contenu du discours plus que sa forme mais la syntaxe joue un rôle implicite. Ce genre d'interprétations peut être incorporé dans les règles explicites.

9 Evaluation

Notre système que nous appelons ERASE (**E**llipsis **R**esolution of **A**rabic **SE**ntences) est un prototype dont le but est de valider les idées et la démarche présentées dans cet article. Il permet à l'utilisateur de saisir une phrase arabe constituée de plusieurs propositions, de vérifier si la phrase est elliptique, de déterminer la classe de celle-ci et de retrouver les constituants manquants. Le développement du système ERASE a nécessité un travail de conception, de réalisation, de test et d'évaluation. Pour évaluer le système ERASE, nous avons extrait un corpus de 200 phrases elliptiques de livres de grammaire arabe à travers des textes littéraires pour les élèves de première année et deuxième année secondaire et des journaux quotidiens. Ces phrases appartiennent aux différentes classes des phrases elliptiques *simples*, *homogènes*, *hétérogènes* et *mixtes*. Les exemples de phrases elliptiques traités dans les différents section de cet article sont un échantillon de ce corpus. Le lexique est formé principalement des mots de ces phrases. Chaque phrase est constituée de deux jusqu'à cinq propositions. Chaque proposition étant composée de un jusqu'à quinze mots.

Les résultats que nous avons obtenus sont satisfaisants pour les plupart d'exemples. Ils sont présentés dans le tableau suivant.

corpus de phrases	**correctement recouvrée**	**recouvréé en partie**
200	190	10
	95 %	5 %

Figure 1. Pourcentage des phrases correctement recouvrées.

La figure 1 montre que 5 % des phrases elliptiques ne sont pas recouvrées correctement. Cela est due que les noms composés, comme (الملك حسين) *le roi Hussein*, sont considérés comme deux syntagmes nominaux par le système (voir exemple).

سافر الملك حسين إلى أمريكا وكذلك (سافرت) الملكة (حسين إلى أمريكا).

Le roi Hussein a voyagé au USA et la reine aussi.

Conclusion

Le module d'identification de la proposition bien formée (i.e., implémentation de la fonction *analyse*), le module d'étiquetage et le module de recouvrement des phrases elliptiques simples sont réalisés à l'aide de l'analyseur syntaxique de propositions construit dans le cadre du système CORTEXA [2] qui se base sur le formalisme des ATNs. Dans ce travail, nous avons proposé une caractérisation formelle du phénomène d'ellipse de la langue arabe et un processus de localisation des parties elliptiques d'une phrase. Nous avons aussi établi des critères de classification que nous avons utilisé pour l'étape de recouvrement. Enfin, nous avons élaborer des algorithmes de recouvrement des phrases elliptiques complexes en généralisant celui que nous avons proposé pour les phrases elliptiques simples. Pour tester les résultats du prototype ERASE, nous avons constitué deux échantillons de phrases elliptiques composés, d'une part, de phrases déjà analysées manuellement et ayant servi à l'élaboration des règles explicites et de la matrice de cohérence des connecteurs, d'autre part, de nouvelles phrases elliptiques destinés à tester la robustesse des règles explicites et leur capacité à s'adapter au plus grand nombre de cas possible.

Références

1. L. Belguith and A. Ben Hamadou, Marquage morpho-syntaxique robuste de mots ambigus écrit en Arabe non-voyellé *Forum de recherche en Informatique* (FRI'96) Tunis. (1996)
2. A. Ben Hamadou, Vérification et correction automatique par analyse affixale des textes écrits en langage naturel: le cas de l'arabe non voyellé, *Thése d'Etat en informatique*, Faculté des Sciences de Tunis. (1993)
3. B. K. Boguraev, Recognizing conjunctions without the ATN framework, *Automatic Natural Language Parsing*, Ellis Horwood. (1983)
4. M. Dalrymple, M. S. Shieber, and F. Pereira, Ellipsis and higher-order unification, *Linguistic and Philosophy 14*, pp 399-452. (1991)
5. C. Gardent, A Multi-Level Approach to Gapping, *proc. of the Stuttgart Ellipsis Workshop* Bericht Nr. 29. (1992)
6. K. Haddar, A. Ben Hamadou, Formal Description of Ellipses in Arabic Language and Resolution Process, *In the Proceedings of IEEE ICIPS'97*, Pekin 29-31 October. (1997)
7. K. Haddar, A. Ben Hamadou, An ellipsis detection method based on a clause parser for the Arabic language, *In the Proceedings of the International FLorida Artificial Intelligence Research Society FLAIRS'98, Sanibel USA* 17-20 Mai.(1998)
8. D. Hardt, Verb phrase ellipsis: form, meaning, and processing, *A dissertation in computer and information science*, University of Pennsylvania, (1993).
9. X. Huang, " Dealing with conjunctions in Machine Translation Environment", proc. of 10th Int. Conf. on Comp. Ling., Standford University, Palo Alto, California, July. (1984)
10. K. Jensen, G. E. Heidorn and S. D. Richardson, *Naturel Language Processing: The PLNLP Approach*, Kulwer academic publishers. (1993)
11. S. Lappin, The Syntactic Basis of VP Ellipsis Resolution, *In proc. of the Stuttgart Ellipsis Workshop Bericht Nr. 29.* (1992)
12. S. Lappin, Computational Approaches to Ellipsis Resolution, *In proc. of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, Ireland, 27-31 March. (1995)
13. M. Mast, F. Kummert et al., A speech understanding and dialog system with a homogeneous linguistic knowledge base, *IEEE Transactions on pattern analysis and machine intelligence*, vol. 16 no. 2, pp 179-193. (1994)
14. F. L. Rau, The understanding and generation of ellipses in a natural language system, University of California Berkeley, *Cs csd-85-227(dir)*, March 1985, 30 pages. (1985)
15. G. Sabah, *L'intelligence artificielle et le langage*, Editions Hermès. (1989)
16. R. M. Weischedel and N. K. Sondheimer, An improved heuristic for ellipsis processing, *In* Proceedings of the 20th Annual Meeting of the ACL, pp 85-88. (1982)
17. W. Woods, Transition networks grammars for natural language analysis, CACM 13, 10, pp 591-606. (1970)