

Étude et analyse de la phrase nominale arabe en HPSG

Abdelkarim Abdelkader, Kais Haddar, Abdelmajid Ben Hamadou

Université de Sfax – MIRACL, ISIMS
abdelkader.abdelkarim@laposte.net, kais.haddar@fss.rnu.tn,
abdelmajid.benhamadou@isimsf.rnu.tn

Résumé

Dans cet article, nous proposons une démarche d'analyse syntaxique pour les phrases nominales arabes à l'aide du formalisme des grammaires syntagmatiques guidées par les têtes HPSG. Pour ce faire, nous commençons par étudier la typologie de la phrase nominale arabe en précisant ses différentes formes. Puis, nous élaborons une grammaire HPSG traitant ce type de phrase et qui respecte la spécificité de la langue arabe. Ensuite, nous présentons une démarche d'analyse syntaxique se basant sur une approche ascendante et sur le mécanisme d'unification. Enfin, nous donnons une idée sur l'implémentation et l'expérimentation du prototype réalisé.

Mots-clés : phrase nominale arabe, grammaire HPSG, schéma adapté, analyse syntaxique.

Abstract

In this paper, we propose a parsing approach for Arabic nominal sentences using the Head-Driven Phrase Grammar formalism HPSG. To elaborate this approach, we begin by studying the typology of the Arabic nominal sentence specifying its different forms. Then, we elaborate a HPSG grammar for this type of sentence. The elaborated grammar respects the Arabic language specificity. Next, we present a parsing approach based on a bottom-up unification process. Finally, we illustrate the implementation and experimentation of the realised prototype.

Keywords: arabic nominal sentence, HPSG grammar, parsing.

1. Introduction

L'analyse syntaxique est une étape importante dans plusieurs applications de traitement automatique de la langue arabe. Le but de l'analyse qui peut se baser sur une grammaire générant toutes les structures syntaxiques de la langue selon un formalisme bien précis est de trouver, entre autre, les représentations syntaxiques des phrases. Parmi, les structures de phrases les plus rencontrées dans la langue arabe sont celles des phrases nominales. La phrase nominale possède une structure assez particulière et bien différente de celle de la phrase verbale. Cette structure n'a pas été bien étudiée dans les analyseurs arabes existants. Ceci est dû, d'une part, à la complexité de la structure de ce type de phrase qui peut contenir des phénomènes assez complexes, et d'autre part, au choix du formalisme utilisé et aux systèmes de règles de production élaborés.

Ainsi, l'un des objectifs de ce travail est d'étudier la structure de la phrase nominale arabe. Cette étude est basée sur des livres de grammaires et sur des discussions avec des linguistes. À partir de l'étude effectuée, nous voulons aussi dégager un système de règles permettant de trouver toutes les représentations syntaxiques possibles des phrases nominales arabes. L'étude et l'analyse sont basées sur le formalisme HPSG adapté pour la langue arabe. L'intérêt de

cette grammaire est triple : d'une part, en tant que système formel, elle facilite la définition des algorithmes d'analyse et ce d'autant plus qu'il y a équivalence avec un modèle mathématique de grammaire formelle ; d'autre part, elle permet l'intégration de ressources hétérogènes, à la fois parce qu'elle utilise le même type de représentation (les structures de traits) et parce qu'elle permet l'accumulation harmonieuse de contraintes sans leur imposer d'ordre d'application ; enfin, elle établit une séparation stricte entre les données linguistiques et les programmes de traitement. L'utilisation de HPSG pour la langue arabe a nécessité une modification des schémas standard pour pouvoir intégrer les spécificités de la phrase nominale arabe.

Dans cet article, nous commençons par présenter quelques travaux effectués sur l'analyse syntaxique de la langue arabe. Ensuite, nous donnons une idée sur la typologie de la phrase nominale arabe. Puis, nous proposons une démarche permettant l'analyse de la phrase nominale arabe en utilisant une grammaire HPSG adaptée. Enfin, nous présentons notre prototype réalisé et nous clôturons par une conclusion et des perspectives.

2. État de l'art

Les recherches dans le domaine de traitement automatique de la langue arabe ont débuté vers les années 1970, avant même que les problèmes d'édition de textes arabes ne soient complètement maîtrisés. Les premiers travaux concernaient, notamment, les lexiques, la morphologie et la correction orthographique. Les recherches concernant la syntaxe de la langue arabe n'ont pas encore atteint un stade évolué. Ceci est dû, d'une part, à la complexité de la langue arabe et, d'autre part, à l'insuffisance des recherches sur cette langue. En effet, les analyseurs existants sont orientés plutôt vers la morphologie que vers la syntaxe proprement dite à l'exception de quelques travaux traitant chacun un phénomène particulier de la syntaxe arabe (Elleuch, 2004 ; Maaloul, 2005).

Dans ce contexte, plusieurs travaux ont été effectués sur l'analyse morphosyntaxique dont nous citons les travaux utilisant le formalisme des grammaires d'arbres adjoints (Tree Adjoining Grammars ou TAG) présentés par Ghenima (1996) et Habbash et Rambow (2004). Dans Ghenima (1996), un système de voyellisation assisté par ordinateur a été proposé. Ce système permet de générer, à partir d'un mot non voyellé, en passant par une analyse morphologique et en utilisant une base de données lexicale, tous les mots voyellés qui peuvent lui correspondre, qu'ils soient des verbes ou des noms. Le système proposé a utilisé une grammaire de l'arabe «non voyellé» afin de réduire les possibilités de voyellisation d'un mot.

D'autres analyseurs morphosyntaxiques basés sur le formalisme AGFLs (Affixs grammars over finite lattice) ont été conçus soit pour faciliter la correction grammaticale des textes en arabe (Ouersighni, 2002), soit pour définir une grammaire de reconnaissance des syntagmes nominaux en arabe en vue de réaliser un système d'indexation automatique des documents écrit (Abbas, 1996).

Outre, un système de correction orthographique de textes arabes a été réalisé par Gader (1995). Il s'agit, dans un premier temps, de la mise en place d'un vérificateur d'orthographe utilisant un analyseur morphologique de textes arabes non-voyellés, et qui fait appel à un lexique de 200 000 entrées. Ce vérificateur orthographique est capable de traiter les textes arabes indépendamment de leurs degrés de voyellisation. Dans un deuxième temps, le travail a proposé un module de correction. Ce module a été intégré dans le vérificateur d'orthographe.

Parmi les travaux menés sur l'étude de la phrase arabe avec le formalisme HPSG, nous citons les deux travaux de recherche présentés dans Elleuch (2004) et Maaloul (2005). Ces deux travaux ont pour objectif la construction d'une grammaire HPSG arabe permettant de traiter aussi bien des phrases verbales simples que des phrases verbales complexes contenant des phénomènes linguistiques assez délicats tels que la coordination, l'ellipse et les relatives. La grammaire conçue est la base de l'analyseur syntaxique utilisé par le prototype AICOO (Maaloul, 2005).

Enfin, nous citons le système MASPAR décrit dans Aloulou (2003). Ce système est basé sur une approche multi agents. Chaque agent de ce système traite soit un niveau bien déterminé de l'analyse (*i.e.*, l'analyse lexicale, l'analyse syntaxique) soit un phénomène linguistique (*i.e.*, l'ellipse, l'anaphore).

3. Typologie de la phrase nominale arabe

Comme nous l'avons déjà indiqué, la langue arabe possède deux types de phrases : la phrase nominale et la phrase verbale. La phrase nominale se compose d'un topique (Mubtada') et d'un attribut (Khabar). Elle exprime, généralement, une constatation ou une définition et le verbe est sous-entendu. L'attribut s'accorde avec le topique en genre et en nombre si ce dernier est au singulier, au duel ou au pluriel (s'il est relatif aux êtres humains). L'exemple (1) illustre ce type d'accord.

Les gens [sont] nombreux الناس كثيرون (1)

Notons que la traduction en français de l'exemple (1) ne donne pas une phrase nominale.

En effet, le topique et l'attribut se présentent sous plusieurs formes qui peuvent changer la structure de la phrase et par la suite la grammaire que nous allons proposer. Dans ce qui suit, nous détaillons ces formes.

3.1. Formes de topique

Selon l'étude que nous avons effectuée, le topique se manifeste sous plusieurs formes. En effet, il peut être un nom propre, un pronom personnel, un nom défini non humain, un nom et un adjectif, une préposition et un nom, un nom suivi d'une conjonction de coordination suivie d'un autre nom ou un nom suivi d'un déterminant suivi d'un pronom attaché. Les phrases (2) à (6) en sont des exemples.

Mohamed [est] intelligent محمد ذكي (2)

La maison [est] grande البيت كبير (3)

Il [est] dans la classe هو في الفصل (4)

Les gens tous ont quitté الناس جميعهم قد غادروا (5)

L'argent et les enfants [sont] le décor de la vie réelle المال و البنون زينة الحياة الدنيا (6)

Notons que la liste des formes présentées ci-dessus n'est pas complète. En fait, il existe d'autres formes de topique telles que les syntagmes relatifs et les syntagmes annexionnistes. Ces formes sont traitées ultérieurement lors de l'élaboration de la grammaire.

3.2. Formes d'attribut

Comme nous l'avant déjà indiqué, l'attribut peut prendre plusieurs formes. Il peut être un nom, un adjectif, un pronom personnel, un pronom d'interrogation, un nom et un adjectif, un

nom suivi d'un autre nom défini, une préposition et un nom ou tout un syntagme verbal. Les phrases (7) à (10) en sont des exemples.

La vie [étant] une lutte الحياة كفاح (7)

L'homme [est] un animal parlant الإنسان حيوان ناطق (8)

Le miel [est] dans les tasses العسل في الأكواب (9)

L'État aide les pauvres الدولة تساعد المحتاجين (10)

L'exemple (10) indique que lorsque le sujet vient avant le verbe la phrase est considérée comme phrase nominale. Dans la phrase de l'exemple 2, l'attribut est un adjectif « ذكي » *intelligent* qui ne joue pas le rôle d'un modificateur. Par contre, dans la phrase de l'exemple 8, le participe « ناطق » *parlant* qui est considéré comme une épithète qualifie l'attribut « حيوان » *animal*. Ainsi, l'épithète peut être un adjectif ou un nom dérivé. Elle prend le genre, le nombre et le cas du nom qu'elle qualifie. L'adjectif attribut est généralement indéfini.

4. Analyse de la phrase nominale arabe avec le HPSG

Rappelons que le HPSG repose sur la notion DI/PL (Dominance immédiate/précédence linéaire) qui consiste à représenter séparément les informations d'ordre linéaire (ou de précédence linéaire) indiquées par des PL-règles et les informations hiérarchiques (ou de domination immédiate) (Pollard et Sag, 1994 ; Sag et Wasow, 1999). HPSG qui est basé sur la structure attribut valeur (SAV) utilise une forme de représentation extrêmement générale : les schémas de domination immédiate. Ces schémas correspondent aux descriptions de structures. Ils permettent, d'une part, de définir la bonne formation des structures, et d'autre part, de gérer le partage des structures en empiétant en quelque sorte sur le rôle des principes (Blache, 1995). Notons que les schémas standard du HPSG sont au nombre de six. Ils indiquent un certain type de relation entre une racine et ses descendants et se distinguent donc essentiellement par le type de fils entrant en jeu.

Dans la sous-section suivante, nous proposons une adaptation des schémas standard du HPSG pour la langue arabe afin de traiter essentiellement les phrases nominales. Puis, nous présentons une démarche d'analyse syntaxique permettant de reconnaître les différentes formes de la phrase nominale arabe discutées précédemment.

4.1. Adaptation des schémas de Domination Immédiate

L'application de HPSG sur la phrase nominale arabe a nécessité une modification des schémas standards. Cette modification est basée sur l'étude linguistique que nous avons déjà effectuée et permettant d'intégrer les particularités de la phrase nominale arabe. Les schémas ainsi adaptés et ci-après introduits se lisent de droite à gauche puisque la phrase arabe s'écrit de droite à gauche.

Schéma 1 (Règle de spécification 1 (R.S 1))

La racine est un signe saturé, le fils tête est un syntagme (sans dépendance non bornée à encaisser), le fils-comps est un singleton. Le schéma adapté de la figure 1 indique qu'un syntagme saturé (ARG-S est un singleton vide) qui ne catégorise aucun constituant dont la tête sous catégorise un spécificateur. Ce schéma permet de représenter, généralement, un nom précédé par un pronom démonstratif (هنا، ذلك) ou un nom précédé par un déterminant (جل، بل).

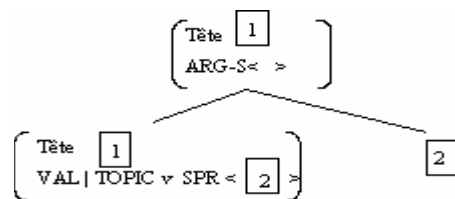


Figure 1. Schéma 1 modifié

Le schéma 1 modifié décrit aussi une phrase nominale entière. Ainsi, l'attribut est la tête et le topique, qui précède l'attribut, est le spécificateur. De plus, les phrases nominales interrogatives (pronom d'interrogation suivi d'un nom) respectent ce schéma.

Schéma 2 (Règle de spécification 2 (R.S 2))

La racine est un signe sous-catégorisant un sujet, le fils tête est un signe lexical.

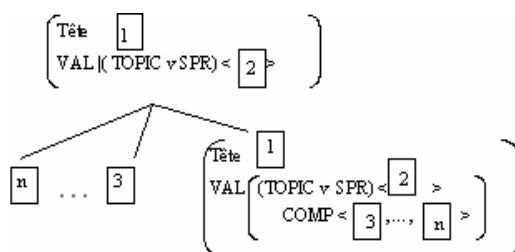


Figure 2. Schéma 2 modifié

Le schéma de la figure 2 décrit les syntagmes verbaux pour lesquels il reste à sous catégoriser un topique. Ainsi, ce schéma décrit la phrase de l'exemple 8 « الدولة تساعد المحتاجين » *l'État aide les pauvres*. Le syntagme « المحتاجين » *les pauvres* est le seul complément du verbe « تساعد » *aide*, le topique « الدولة » *l'État* spécifie le groupe verbale « تساعد المحتاجين » *aide les pauvres*.

Schéma 3 (Règle de complémentation (R.C))

La racine est un signe saturé, le fils tête est un signe lexical.

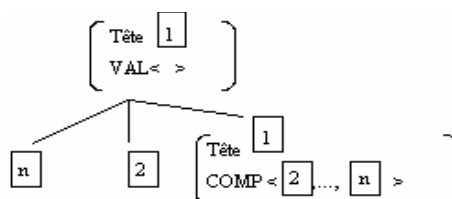


Figure 3. Schéma 3 modifié

Typiquement les syntagmes prépositionnels (prép + SN) (جار و مجرور) ou les syntagmes nominaux (N + SN) (مضاف و مضاف إليه) respectent le schéma 3 modifié de la figure 3. Ainsi, les SAVs des syntagmes « في البيت » *dans la maison* et « راجل دلو » *le fils de voisin* sont formées à partir de ce schéma.

Schéma 4 (Règle de modification (R.M))

Le trait MOD du fils modificateur sélectionne le fils d'un signe.

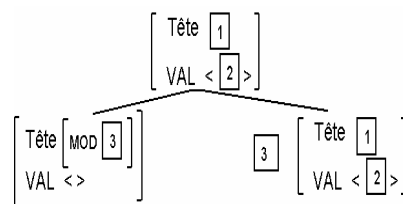


Figure 4. Schéma 4

Le schéma de la figure 4 permet de contrôler la sélection d'une tête par son modificateur via le trait MOD. Comme le français et l'anglais, le modificateur en arabe peut être soit une épithète, pour le nom soit un adverbe, pour le verbe.

Le reste des schémas DI du HPSG ne sont pas modifiés. Les principes qui contrôlent les schémas restent valables aussi pour les schémas adaptés à la langue arabe présentés ci-dessus. Quant aux traits, qui constituent l'élément de base d'une structure HPSG, leur adaptation a nécessité d'une part, l'ajout de quelques traits et d'autre part, la modification de certains autres (Maaloul, 2005).

4.2. Démarche d'analyse syntaxique utilisée

Dans le présent paragraphe, nous présentons la démarche d'analyse syntaxique permettant de reconnaître les différentes formes de phrases nominales discutées dans la troisième section selon une grammaire HPSG. La méthode, que nous proposons, nécessite le passage par trois phases : le découpage de la phrase en mots, le chargement des SAV's des mots et l'analyse syntaxique proprement dite. La figure 5 illustre les différentes étapes de la démarche d'analyse syntaxique poursuivie.

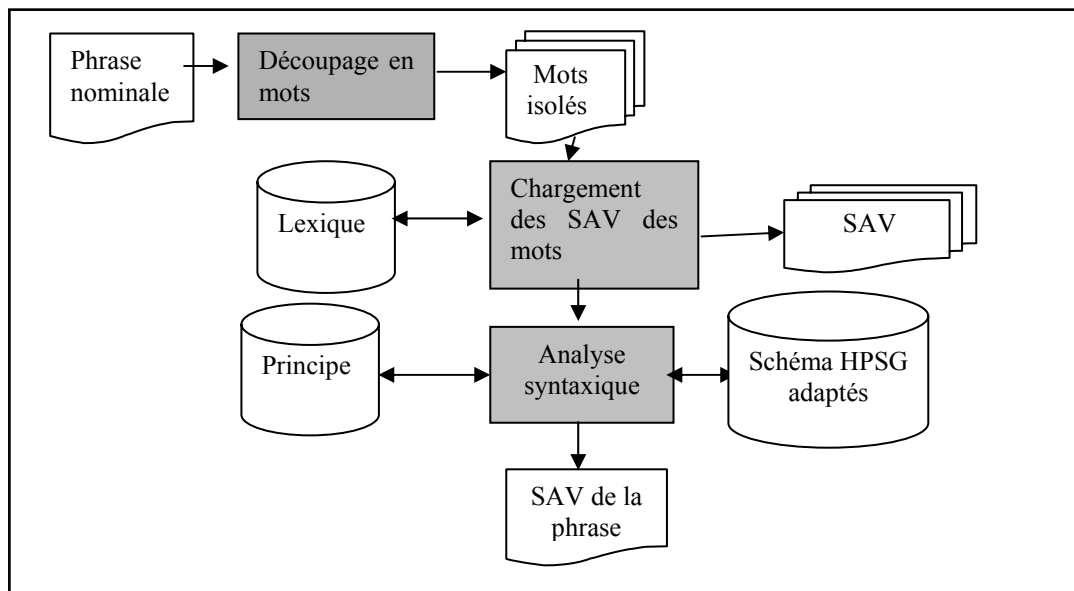


Figure 5. Démarche d'analyse syntaxique

La phase de découpage qui consiste à découper la phrase en mots utilise comme indicateurs de séparation, les espaces, certains pronoms et certaines prépositions. Ainsi, le découpage de la phrase de l'exemple 11

البيت كبير

(11)

La maison est grande

donne les deux mots « البيت » la maison et « كبير » grande.

Cette première phase est nécessaire pour pouvoir charger, par la suite, les structures attributs/valeurs (SAV) des mots. Donc, la phase de chargement permet de prendre en compte toutes les informations, concernant chaque mot composant la phrase sous forme de SAVs. Ces informations sont stockées dans des fichiers XML formant notre lexique. Chaque fichier correspond à une catégorie lexicale. La taille du lexique actuel est d'environ 1000 mots mais nous sommes en train de construire un autre lexique plus développé avec une représentation normalisée. Notons qu'un mot donné peut avoir plusieurs entrées lexicales et le chargement de la SAV convenable nécessite des retours en arrière.

La phase de l'analyse syntaxique proprement dite consiste à trouver la représentation syntaxique de la phrase nominale en utilisant le mécanisme d'unification et les schémas modifiés pour la langue arabe. L'utilisation de ces schémas consiste à sélectionner le schéma approprié pour construire les structures attributs/valeurs des syntagmes parents de chaque mot de la phrase. En effet, la règle de modification exige qu'un modificateur partage toujours les mêmes traits de la tête, la règle de spécification exige que les traits de tête (*i.e.*, attribut) soient transmis aux spécificateurs (*i.e.*, topique). La construction de la structure attribut/valeur de la phrase est basée aussi sur l'utilisation du principe HFP qui garantit que les traits de tête, quelle que soit la catégorie, seront copiés sur le syntagme parent. L'analyse syntaxique d'une phrase sera alors le résultat d'une unification réussite des informations lexico-syntaxiques associées aux mots de la phrase. Cette unification permet d'aboutir à la construction d'une SAV. En outre nous avons défini une certaine priorité d'application des règles lors de l'unification.

Ainsi, la règle de modification (R.M) doit être appliquée en premier lieu, si la phrase à analyser contient une épithète. Cependant, elle peut être précédée par une règle de coordination car nous pouvons trouver deux épithètes conjointes dans un syntagme nominal. En second lieu, et dans le cas de la présence de deux noms consécutifs tel que le premier soit indéfini, ou d'une préposition la règle de complémentation (R.C) doit être appliquée. La règle de spécification 1 (R.S1) possède une priorité minimale. Elle est plus générale que les autres. La règle de spécification 2 (R.S2) est applicable lorsque l'attribut est un syntagme verbal.

Algorithme Analyser

Entrées : SAVs des mots, L : liste contenant les mots de la phrase à analyser

Sortie : SAV de la phrase

DEBUT

Initialiser L par les mots de la phrase

Pour chaque mot de L

Charger la SAV du mot

Fin pour

Tant que (Nombre de syntagme >2) faire

Parcourir L pour choisir le schéma propriétaire (x)

Syntagme = Application du schéma propriétaire (x)

Ajouter *Syntagme* dans L

Supprimer les constituants du *Syntagme* de L

Fin tant que

Vérifier les formes de deux syntagmes obtenus

Fin

Figure 6. Algorithme d'analyse syntaxique

Pour les phrases nominales contenant un verbe d'existence (كان و أخواتها) ou un verbe inchoatif (فعل من أفعال الشروع), nous devons appliquer la règle de complémentation (R.C) en dernier lieu. Chaque règle appliquée avec succès donne naissance à un nouveau syntagme. L'algorithme de la figure 6 illustre le fonctionnement du processus de l'analyse syntaxique.

Cet algorithme montre que l'application des schémas modifiés se poursuit jusqu'à ce que le nombre de syntagmes de la liste L soit égal à 2. La dernière étape de l'algorithme vérifie si les deux syntagmes trouvés peuvent former une phrase nominale.

5. Expérimentation et évaluation

Le prototype correspondant à l'analyseur syntaxique contient cinq modules : un module de segmentation de textes en mots, un module de mise à jour de lexique, un module de découpage des phrases en mots, un module de chargement des SAVs et un module d'analyse syntaxique proprement dite. Ce prototype a pour but de valider les idées et la démarche présentées dans cet article. Dans sa version actuelle, il est formé, principalement, d'une interface JAVA où l'utilisateur peut saisir une phrase nominale arabe et l'analyser en activant les boutons de la figure 7. Enfin, si la phrase est correcte, il peut afficher sa SAV. Ainsi, pour tester le prototype sur une phrase donnée, nous pouvons la découper en un ensemble de mots. Ceci est réalisé par un clic sur le bouton **Découper** de l'interface. Aussi, nous pouvons charger la SAV de chaque mot de la phrase étudiée en activant le bouton **Charger SAV**. Autrement, nous analysons la phrase en activant le bouton **Analyser** qui englobe les précédentes opérations.

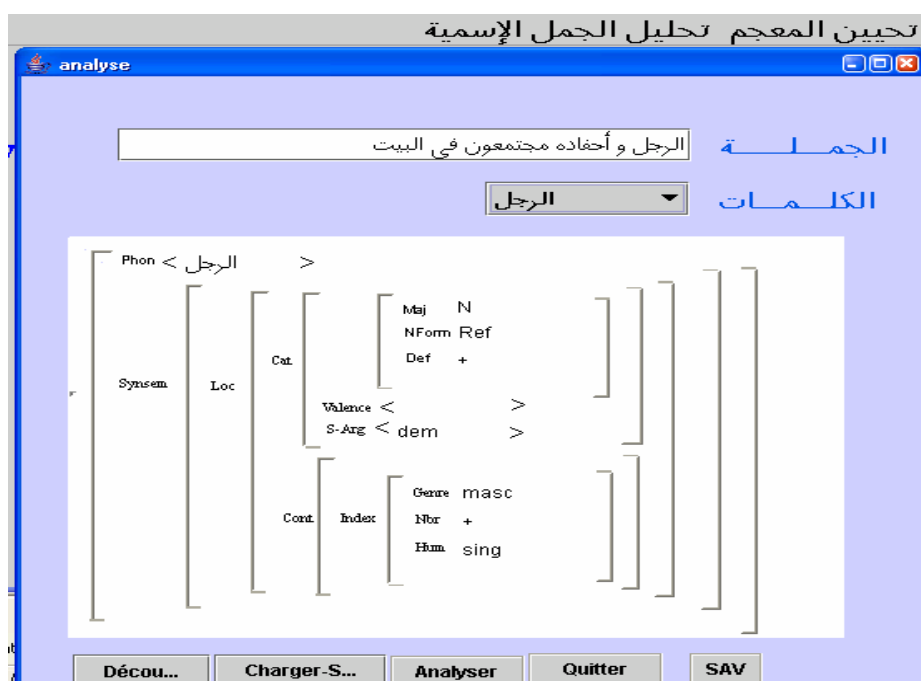


Figure 7. Analyse de la phrase « الرجل و أحفاده مجتمعون في البيت »

Notons que l'interface de notre prototype permet aussi de mettre à jour le lexique. Pour ajouter un nom, par exemple, nous utilisons le menu « تحيين المعجم », et nous choisissons l'option « الأسماء ».

Pour évaluer le prototype réalisé, nous avons extrait un corpus de 100 phrases nominales de livres de grammaires arabes à travers des textes littéraires pour les élèves de huitième année

et neuvième année de l'enseignement de base. Ces phrases appartiennent aux différents types de phrases nominales. Le lexique (qui est en cours d'amélioration) est formé principalement des mots de ces phrases. Chaque phrase étant composée de deux jusqu'à quinze mots. Notons que notre système analyse les phrases sans avoir recours à une décomposition préalable de phrase en propositions simples. De ce fait, l'analyse nous permet de réduire le temps d'exécution puisque nous appelons une seule fois le module de l'analyse syntaxique proprement dite. En outre, notre analyse conserve la sémantique de la phrase qui risque d'être perdue au moment de la décomposition en propositions. Le fait de travailler sur un corpus bien déterminé de phrases nous a permis d'écarter en quelques sortes le problème d'ambiguïté. Les résultats que nous avons obtenus sont satisfaisants pour la plupart d'exemples. Notons que l'échec de l'analyse syntaxique (s'il existe) peut être expliqué par les raisons suivantes : un mot composé (roi Hussein) ou une structure grammaticale non prévue dans la grammaire proposée.

6. Conclusion et perspectives

Dans cet article, nous avons étudié d'abord la typologie de la phrase nominale arabe. Cette étude nous a permis de dégager une grammaire HPSG traitant essentiellement les phrases nominales. La grammaire dégagée montre l'efficacité de l'utilisation du formalisme HPSG pour la langue arabe. Ensuite, nous avons construit un analyseur syntaxique pour les phrases nominales utilisant les algorithmes classiques de la compilation avec quelques modifications et la grammaire HPSG. Enfin, nous avons conçu un prototype montrant la faisabilité des idées et la démarche proposée.

Comme perspectives, nous comptons, d'abord, remplacer le lexique de notre système par un lexique amélioré et représenté avec un langage pivot normalisé dans le cadre d'un autre projet afin de traiter la majorité des ambiguïtés syntaxiques de la langue arabe. Ensuite, nous voulons tester notre prototype avec un analyseur syntaxique généré par le LKB (Linguistic Knowledge Building) réalisé dans l'université de Stanford.

Références

- ALLOULOU C. (2003). « Analyse syntaxique de l'Arabe : Le système MASPAR ». In *Actes de RECITAL 2003*. Batz-sur-Mer : 419-428.
- ABBAS W. (1996). *Vers une grammaire de reconnaissance du syntagme nominal en arabe*. Rapport scientifique, Université lumière Lyon 2.
- BLACHE P. (1995). « Introduction à HPSG ». In *2LC-CNRS* : 1-24.
- EL-KASSAS D., KAHANE S. (2004). « Modélisation de l'ordre des mots en arabe standard ». In *Actes de JEP-TALN 2004, Traitement Automatique de l'Arabe*. Fès : 259-264.
- ELLEUCH S. (2004). *Analyse syntaxique de la langue arabe basée sur le formalisme d'unification HPSG*. DEA en informatique, Faculté des sciences économiques et de gestion de Sfax.
- GADER N. (1995). *Conception et réalisation d'un correcteur orthographique de texte arabes*. Rapport scientifique, centre de recherche CERSI à l'ENSIB.
- GHENIMA M. (1996). « Analyse morpho-syntaxique de la langue arabe basée sur le formalisme des grammaires d'arbres adjoints (TAG) ». In *Actes de la Conférence NLP +IA/TAL+AI 96 Moncton, N.B.* : 106-110.
- HABBASH N., RAMBOW O. (2004). « Extracting a Tree Adjoining Grammar From the Penn ArabicTreebank ». In *Actes de JEP-TALN 2004, Traitement automatique de l'Arabe*. Fès : 446-454.

- MAALOUL H. (2005). *La coordination dans la langue arabe : étude et analyse en HPSG*. Master en informatique, Faculté des sciences économiques et de gestion de Sfax.
- OUERSIGHNI R. (2002). « L'analyse morpho-syntaxique de l'arabe voyellé ou non voyellé : le système AraParse ». In *Actes du Congrès international pour le traitement automatique de la langue arabe*. Faculté des lettres, Mannouba : 166-179.
- POLLARD C., SAG I. (1994). *Head-Driven Phrase Structure Grammars*. Chicago University Press, Chicago.
- SAG I., WASOW T. (1999). *A formal Introduction Center for the Study of Language and Information*. CSLI publication, Stanford University.