
PHARAS : une plate-forme d'analyse basée sur le formalisme HPSG pour l'Arabe standard.

Développements récents et perspectives

Mourad LOUKAM* — **Mohamed Tayeb LASKRI****

** Département d'informatique, Faculté des Sciences et Sciences de l'Ingénieur, Université Hassiba Benbouli Hay Essalem, Chlef, Algérie
loukam@hotmail.com*

***Laboratoire de Recherche en Informatique (LRI), Université Badji Mokhtar Sidi Ammar, Annaba, Algérie
laskri@univ-annaba.org*

RÉSUMÉ. Le formalisme HPSG connaît depuis plusieurs années un essor remarquable dans le domaine du TALN en général, et l'analyse de la syntaxe en particulier. Nous travaillons sur le formalisme HPSG sur le double aspect modélisation/implémentation en vue de son application sur l'arabe standard.

Dans cet article, nous présentons notre projet PHARAS (Plateforme d'analyse basée sur le formalisme Hpsg pour l'analyse de l'ARabe Standard). Il s'agit d'un outil intégré qui offre toute la chaîne d'analyse d'un texte en arabe, voyellé ou non, dans le but de produire son analyse en format HPSG.

MOTS-CLÉS : HPSG , Langue Arabe, TALN.

1. Introduction

Le formalisme HPSG, Head-driven Phrase Structure Grammar (Grammaires syntagmatiques dirigées par la tête), a été conçu par Carl Pollard et Ivan Sag (Pollard et al., 1994). Il s'agit, incontestablement, de l'une des théories les plus en vue actuellement dans le domaine du TALN en général, et notamment dans les travaux concernant la modélisation et le traitement de la syntaxe¹(Abeillé, 2007).

¹ Un symposium international lui est consacré chaque année.

Nous travaillons sur le formalisme HPSG sur le double aspect modélisation/implémentation en vue de son application sur l'arabe standard (Loukam et al., 2007).

Dans cet article, nous présentons notre projet PHARAS² (Plateforme d'analyse basée sur le formalisme Hpsg pour l'analyse de l'ARABe Standard). Il s'agit d'un outil intégré qui offre toute la chaîne d'analyse d'un texte en arabe dans le but de produire son analyse en format HPSG (Loukam, 2007). A l'état actuel, la plateforme dispose des modules nécessaires à l'analyse, notamment : un sous-système d'analyse morpho-lexicale donnant ses résultats conformes au formalisme HPSG (Loukam et al., 2007) et un analyseur syntaxique utilisant l'unification.

2. Travaux connexes

Parmi les outils logiciels supportant le formalisme HPSG, on peut citer (Tseng, 2006) :

- LKB (Linguistic Knowledge Building): est un système de développement grammatical créé par Ann Copestake et son équipe à l'université de Cambridge (Copestake, 2002).

- TRALE : est une plateforme d'implémentation de grammaires HPSG, issue du projet MiLCA et développée à l'Université de Breme (Allemagne).

- Matrix : une plate-forme expérimentale, soutenue par près d'une quinzaine laboratoires de recherches. Il s'agit d'un noyau grammatical universel proposant une signature de base (types généraux, types lexicaux simples, règles de combinaison) et un ensemble de modules paramétrés (questions, négation, coordination, etc.) qui permettent alors de « générer » une analyse sous forme de grammaire typée.

- Enju : un analyseur syntaxique HPSG pour l'anglais, développé au Tsujii laboratory de l'Université de Tokyo (Miyao et al., 2005) et (Ninomiya et al., 2006).

- Babel : un analyseur syntaxique HPSG pour l'allemand, développé à l'Université de Berlin (Müller 2001).

En ce qui concerne le traitement de l'arabe standard, parmi les rares les travaux qui prennent comme cadre de travail le formalisme HPSG, nous pouvons citer le système MASPAR (Bahou et al., 2006) développé à l'université de Sfax.

3. Le projet PHARAS

Le projet PHARAS (Plate-forme d'analyse basée sur le formalisme Hpsg pour l'analyse de l'ARABe Standard) a pour objectif de développer un outil intégré

² Pharas ou Faras signifie en arabe « cheval de haute race »

offrant toute la chaîne de traitement d'un texte arabe (voyellé ou non) en vue d'obtenir son analyse selon le formalisme HPSG (voir figure 1).



Figure 1. *Objectif de PHARAS*

« L'ouverture » du système a été une préoccupation constante lors de la conception. En effet, dans l'espoir de voir la plateforme utilisée dans des applications de TALN diverses, nous avons jugé utile de proposer des formats de sortie « normalisés ». Il s'agit des AVM (Attribute Value Matrix) et XML.

3.1 Fonctionnement général

Un texte en arabe introduit sur PHARAS passe par une série de phases de traitement. Nous pouvons les résumer ainsi :

1. Phase de segmentation du texte : le texte est décomposé en « mots ».
2. Phase d'analyse morphologique : Après avoir segmenté le texte, on soumet chacun des items rencontrés à une analyse morpho-lexicale.
3. Phase de génération des matrices attribut / valeur HPSG : Cette phase est réalisée par l'analyseur morpho-lexical. Elle consiste à générer, pour chaque item, sa structure de traits sous la forme d'une matrice attributs-valeurs (AVM).
4. Phase d'analyse syntaxique: L'analyse syntaxique en HPSG se base principalement sur l'application du processus d'unification. Il opère sur des structures de traits (AVM) des entrées lexicales des différents mots, déjà générées lors de la phase précédente, ainsi que sur les règles syntaxiques (schémas).
5. Phase de production des résultats : il s'agit de présenter sous forme concrète (AVM ou XML) la représentation syntaxique et sémantique du texte analysé.

3.2 Architecture générale

L'architecture de PHARAS repose sur l'interconnexion de plusieurs sous-systèmes faisant intervenir des ressources et des outils divers (voir figure 2).

Nous faisons ci-après une description de cette architecture.

Le sous-système d'analyse morpho-lexicale : après un prétraitement appliqué au texte d'entrée (segmentation) ce système réalise l'analyse morpho-lexicale des éléments du texte.

Le sous-système « signes et règles HPSG » : Ce sous-système est représenté par la signature HPSG retenue, la hiérarchie de types ainsi que les règles à appliquer. Il est composé de trois fichiers : le fichier « Types », le fichier « Règles » et le fichier « Lexique ».

Dans le fichier « Types », on définit la hiérarchie des types utilisés pour décrire les traits. Rappelons que cette hiérarchie joue un rôle primordial en HPSG puisqu'elle constitue elle-même un ensemble de contraintes sur les structures de traits.

Dans le fichier « Lexique », on stocke toutes les entrées lexicales (verbes, noms, adjectifs, particules) déjà rencontrées ou analysées.

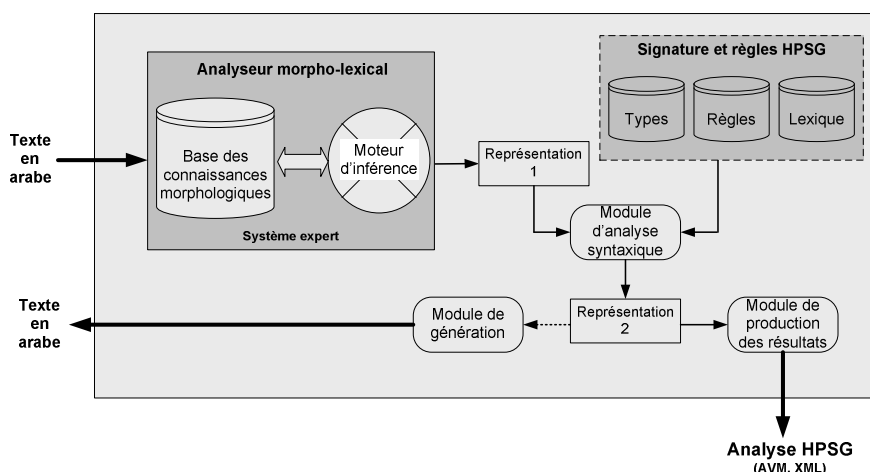


Figure 2. Architecture générale de PHARAS

Chaque entrée est représentée par une structure de traits (voir l'exemple du verbe « fahima »³ en figure 3). Rappelons que les des caractéristiques essentielles de HPSG, à l'instar des formalismes dits fortement lexicalisés, est que la plupart des contraintes sur la syntaxe, voire la même la sémantique, sont présentes dans les entrées lexicales elles-mêmes.

³ a compris

Phon	فهم	فهم:= word & [PHON < 'فهم'>, SS.LOC[CAT.Tete[MAJ فعل, VFORM 2, TEMPS الماضي , ROOT ر-ح-م, MOD مبني على الفتح , CAT.Valence <>, CAT.S-ARG<>], CONT [Index[الفاعل:[PERSON الغائب, NUMB مفرد, GENRE مذكر , Restr <>]]]]
Class	فعل	
Voix	معلوم	
Temps	الماضي	
Mode	مبني على الفتح	
Root	ف-ه-م	
VForm	متعدي 2	
Personne	الغائب	
Nombre	مفرد	
Genre	مذكر	

Figure 3. Exemple de structure de traits d'une entrée lexicale (verbe « fahima »)

Dans le Fichier « Règles », on formalise l'ensemble des règles syntaxiques (schémas) à traiter de la langue arabe standard. Rappelons que dans le formalisme HPSG, les règles elles-mêmes sont décrites par des structures de traits. L'existence de ce fichier est primordiale pour l'étape d'unification.

3.3 Module d'analyse morpho-lexicale

Ce sous-système a été réalisé en utilisant l'approche « système expert » (Loukam et al., 2007). Cette approche est intéressante à plusieurs points de vue, elle permet notamment de rendre le système ouvert et paramétrable, et ce en raison de la possibilité d'ajouter de nouvelles connaissances morpho-lexicales sous forme de règles et de faits.

3.4 Module d'analyse syntaxique

L'analyse syntaxique se base principalement sur le processus d'unification. Il opère à partir de l'ensemble des structures de traits (AVM) des entrées lexicales analysées et produites à l'étape précédente ainsi que sur les règles syntaxiques (schémas).

L'Algorithme d'unification se déroule sommairement en trois étapes :

- indexation de toutes les AVM de la phrase à analyser :
- application des règles syntaxiques existant dans le fichier "Règles".
- construction incrémentale de structures de traits « compatibles » par enrichissement.

Le processus d'unification s'arrête après le traitement de tous les éléments de la phrase et peut donner naissance à la structure globale (si le texte analysé ne comporte aucune erreur) ou encore à des fragments de structures (en cas de texte non reconnu en totalité par la grammaire).

En cas d'échec de l'unification, le système affiche un diagnostic des règles qui n'ont pas été satisfaites.

3.5 Module de production des résultats

Le module de production des résultats : Ce module s'occupe de restituer le résultat de l'analyse à l'utilisateur sous l'un des formats suivants : matrice attributs-valeurs (AVM) ou fichier XML.

3.6 Module de génération :

Ce module est prévu dans le but de faire une « vérification » de l'analyse obtenue.

3.7 Modules supplémentaires :

Plusieurs autres modules sont prévus pour une utilisation efficace de la plateforme. Nous pouvons citer notamment :

Les modules « interfaces » : ces modules sont prévus dans le but de faciliter l'utilisation et le paramétrage du système : affichage graphique de la hiérarchie des types, mise à jour du fichier des règles, des types, ...etc.

Un module « trace » : ce module est prévu dans le but de mémoriser, et de restituer en cas de besoin, les étapes d'application du processus d'analyse et éventuellement les erreurs rencontrées (cas de phrases partiellement correctes).

4. Etat des lieux et perspectives :

Dans cet article, nous avons présenté notre projet PHARAS, une plateforme d'analyse basée sur le formalisme HPSG pour l'arabe standard. Il s'agit d'un outil intégrant la chaîne totale de traitement d'un texte écrit en arabe standard, voyellé ou non.

Notre système se démarque des rares autres travaux se faisant sur l'arabe standard et utilisant HPSG sur plusieurs aspects :

Il intègre un analyseur morpho-lexical donnant ses résultats sous la forme de structures de traits, alors que les autres systèmes en sont dépourvus et supposent que le texte entré est déjà étiqueté grammaticalement (Bahou et al., 2006).

Il intègre un module de génération qui permet de valider l'analyse obtenue, en reconstituant à partir de la structure profonde le texte brut introduit au préalable.

Plusieurs perspectives s'offrent à notre travail, nous pouvons citer, entre autres :

Sur le plan de la modélisation : une multitude de travaux de modélisation peuvent être entrepris pour élargir la couverture des phénomènes linguistiques traités (les phrases passives, interrogatives, relatives, coordination, ... etc).

Sur le plan de l'implémentation, nous proposons la mise en œuvre des autres modules de la plateforme.

5. Bibliographie

- Abeillé A., *Les grammaires d'unification*, Paris, Lavoisier Editions, 2007.
- Bahou Y., Hadrich Belguith L., Aloulou C., Ben Hamadou A., « Adaptation et implémentation des grammaires HPSG pour l'analyse de textes arabes non voyellés », *Actes du 15e congrès francophone AFRIF-AFIA Reconnaissance des Formes et Intelligence Artificielle RFIA'2006*, 25/26/27 Janvier 2006, Tours/France.
- Copestake A., *Implementing Typed Feature Structure Grammars*, CSLI Publications, Stanford University, 2002.
- Loukam M., « PHARAS : Une plateforme d'analyse basée sur le formalisme HPSG pour l'arabe standard », *Actes du premier séminaire sur le langage naturel et l'intelligence artificielle LANIA'2007*, 20/21 Novembre 2007, Chlef/Algérie, p 31-40.
- Loukam M., Abbache A., Laskri M.T., « Un analyseur morpho-lexical à base de système expert en vue d'une analyse en HPSG », *Actes de la conférence Internationale sur le traitement automatique de la langue arabe CITALA'07*, 18/19 Juin 2007, Rabat/Maroc, p 159-166.
- Loukam M., Laskri M.T., « Vers la modélisation de la grammaire de l'arabe standard basée sur le formalisme HPSG », *Actes des journées de l'école doctorale JED'2007*, Université Badji Mokhtar, 27/28 Mai 2007, Annaba/Algérie.
- Miyao Y., Tsujii J., « Probabilistic Disambiguation Models for Wide-Coverage HPSG Parsing », In *Proceedings of ACL-2005*, 2005, p. 83-90.
- Müller S., *The Babel-System : a parser for an HPSG fragment of German*, 2001, Language Technology Lab, Université de Berlin.
- Ninomiya T., Matsuzaki T., Tsuruoka Y., Miyao Y. Tsujii J., « Extremely Lexicalized Models for Accurate and Fast HPSG Parsing ». In *Proceedings of EMNLP 2006*.
- Pollard C, Sag I., *Head-driven Phrase Structure Grammar*. The Ohio State University and Stanford University, 1994.
- Tseng J., *Implémentation HPSG avec LKB: La Matrix et la Grenouille*, Séminaire HPSG UFRL – Paris 7, 14 décembre 2006.