ESQUISSE D'UNE GRAMMAIRE HPSG POUR L'ARABE

Mahmoud Fawzi Mammeri
Centre de Recherche Scientifique et Technique
pour le Développement de la Langue Arabe
f_mammeri@esc-alger.com

Hamid Azzoune
Institut d'Informatique
Faculté d'Informatique et d'Electronique (USTHB)
azzoune@yahoo.com

Zoheir Zemirli
Institut National d'Informatique (INI)
z_zemirli@ini.dz

Résumé

Le présent travail ébauche une nouvelle grammaire HPSG[1] pour l'arabe. Dans ce qui suit, nous allons décrire un fragment de cette grammaire. Il sera question, d'une part, de spécifier les différentes composantes de cette grammaire, à savoir ses schémas de dominance immédiate, ses principes et une réflexion sur son futur lexique. Il sera aussi question, d'autre part, de réfléchir, à chaque fois que le contexte l'exige, sur les futurs extensions et les problèmes qui leur sont liés. Les analyses que nous présenterons, s'inscrivent dans un projet, que nous avons initié au CRSTDLA[2], pour le développement d'une grammaire électronique de l'arabe, qui sera implémentée sur la plate-forme de développement de grammaire LKB[3].

Mots clés

Analyse syntaxique de l'arabe - HPSG - théorie néo-khalilienne - LKB - grammaire électronique - lexique.

[1] Head –driven Phrase Structure Grammar.
[2] Centre de Recherche Scientifique et Technique pour le Développement de la Langue Arabe.
[3] Linguistic Knowledge Builder.

الملخص

يرسم هذا العمل قواعد نحو جديدة للغة العربية من الشكل HPSG ثم يليه وصف لجزء من هذا النحو. ويهتم موضوعنا من جهة بصياغة المكونات المختلفة لهذا النحو من جهة، أي الرسوم التخطيطية للمكونات القريبة ومبادئ حسن التركيب والتكوين ونظرة في كيفية تكوين المعجم مستقبلا. من جهة أخرى، في التفكير كلما اقتضت الضرورة في التوسعات المستقبلية لهذا النحو والمشاكل المرتبطة به من جهة أخرى. إن التحاليل التي نقترحها هنا تندرج ضمن مشروع بدأنا العمل فيه بمركز البحث العلمي والتقني لتطوير اللغة العربية، من أجل تطوير قواعد لنحو إلكتروني للغة العربية، سوف يتم إنجازه على نظام LKB الخاص بكتابة قواعد النحو.

الكلمات المفاتيح

التحليل النحوي للغة العربية - HPSG - النظرية الخليلية الحديثة - LKB - قواعد النحو الإلكتروني - المعجم.

Abstract

The present work sketches a new HPSG grammar for Arabic. In what follows, we are going to describe a fragment of this grammar. It will be question, on the one hand, to specify the different components of this grammar, that is to say its dominance immediate schemata, its principles, and a thought on its future lexicon. It is also question, on the other hand, to think every time the context requires it, on the future extensions and the problems that are bound to it. The analyses that we will present, are be part of a project, which we initiated in the CRSTDLA for the development of an electronic grammar of Arabic, which will be implemented on the grammar development platform LKB.

Keywords

Arabic parsing - HPSG - neo-khalilian theory - LKB - computational grammar - lexicon.

1. Introduction et cadre théorique

Ce travail qui était censé, au départ, être une révision de la grammaire présentée dans [MAMMERI, 2003] (dorénavant [MAM, 2003]), ébauche une nouvelle grammaire HPSG pour l'arabe. Toutefois, comme il sera constaté, ne sont pas minimes les représentations ci-dessous présentées qui trouvent leur esquisse dans [MAM, 2003].

Il s'agit d'un travail d'analyse où l'on se propose dans une première étape d'analyser une phrase écrite dans la langue arabe pour reconnaître ses différents constituants syntaxiques. La grammaire de l'arabe qu'on se propose d'investiguer et de développer trouve ses origines dans l'analyse linguistique donnée par [HADJ SALAH, 1979], et son cadre formel dans la famille des grammaires dites d'unification.

Pour rappel, dans [MAM, 2003], l'objectif principal a été de proposer des descriptions des structures de la langue arabe qui, d'une part, tirent leurs origines de l'analyse linguistique donnée par [HADJ SALAH, 1979]. En effet, la mise en œuvre d'une analyse syntaxique de la langue arabe doit passer par l'utilisation de concepts et d'analyses linguistiques bien décrites. Cette théorie logico-mathématique définit les composantes syntaxiques d'une phrase de la langue arabe selon la notion de schèmes générateurs, et confère ainsi au module syntaxique des fondements très solides. D'autre part, lesdites descriptions ont été décrites, par la suite, dans le cadre formel du formalisme des Grammaires Syntagmatiques Guidées par les Têtes (Head-driven Phrase Structure Grammar, ou HPSG). Ces grammaires reposent sur l'idée de représenter tous les objets d'une grammaire donnée par des structures de traits typées (Typed Feature Structures, ou TFSs).

Les analyses qui seront présentées ici s'inscrivent dans un projet, que nous avons initié au CRSTDLA (Centre de Recherche Scientifique et Technique pour le Développement de la Langue Arabe), pour le développement d'une grammaire électronique pour le traitement automatique de l'arabe écrit. Cette grammaire sera implémentée sur la plate-forme LKB (Linguistic Knowledge Builder, Copestake, 2002).

2. Origine théorique

Notre présent travail repose sur l'étude du composant syntaxique de la langue arabe élaboré par [HADJ SALAH,1979] et connu sous l'abréviation TNK (pour théorie néo-khalilienne).

La TNK, dans sa description de la langue et pour aboutir à des fondements scientifiques, a développé certains concepts, tels que le *qiyās* (équivalence au sens mathématique), *aṣl/far'*, *mawḍi'* (position virtuelle), *ta'āqub* (alternance exclusive), *miṯāl* (schème générateur), *binā'/waṣl* (combinaison structurante/concaténation), *ḥarf*, *kalima*, tectonic, etc. Ces concepts ne seront pas tous développés ici, on se contentera de ceux qui ont un lien direct avec le présent travail, on trouvera les autres dans la littérature de la théorie.

Les niveaux de la langue, proposés par la TNK, qui ont un rapport avec le présent travail sont le niveau intralexical (niveau de la lexie) et le niveau interlexical (niveau de la tectonie). Toutefois, nous ferons un bref passage sur le niveau lexical (niveau de la *kalima*), pour des raisons de lexique.

2.1. Différents niveaux

2.1.1. Niveau lexical

Le niveau lexical est basé sur la notion de *kalima*. Pour connaître ce qu'est la *kalima*, définissons d'abord la notion de *ḥarf*.

a- Notion de ḥarf

Le *ḥarf* est une unité qui se réalise dans le discours (kalām). On ne peut le prononcer isolément, mais concaténé à un autre *ḥarf*; il participe ainsi à la construction des unités du niveau supérieur que sont les *kalima*-s.

b- Notion de kalima

La *kalima* (ou segment signifiant minimal) est une séquence signifiante de *ḥurūf*. Les *kalima*-s relèvent de deux grandes classes : des *kalima*-s régulières, obéissant à des règles de formation (les verbes et les noms), et des *kalima*-s non régulières (exposants et certaines unités qui leur sont assimilées, et les noms propres et les noms communs).

La construction d'une *kalima* se fait à partir de ce qu'on appelle le *schème linéaire de la kalima*. Ce schème accepte en entrée ce que l'on nomme une *racine* pour générer en sortie la *kalima* qui lui correspond.

c- Notion de racine

Une *racine* est une séquence de consonnes primitives non ajoutées qui forment un ensemble ordonné. Les *kalima*-s, *maktab* (bureau), *kataba* (a écrit), *kātib* (écrivain), *kitāb* (livre), *kutub* (livres) etc., sont formées d'un ensemble fini de consonnes (k, t et b) présentes invariablement à la même position auxquelles sont rajoutés certains ajouts comme « ma » et « a » dans « **ma**k**t**a**b** », « a », « a » et « a » dans « k**a**t**a**b**a** », etc.

d- Schème linéaire de la kalima

Le *schème générateur de la kalima* se présente comme un ensemble de positions contiguës, qui sont de deux natures, des *positions fixes* (remplies au préalable), et d'autres *vides*, dans lesquelles peuvent être insérées des racines.

On se passera de plus de détails sur ce niveau qui réellement sort du cadre de notre travail. Nous n'en avons parlé que par souci de présenter le vocabulaire de la TNK qui nous servira de langage dans ce qui suit. Pour plus de détails, on pourra se référer à [MAM, 2003].

2.1.2. Niveau intralexical

Si nous regardons les séquences suivantes du point de vue construction :

(a) Kitāb
Un livre

(b) kitābu Zayd
Le livre de Zayd

(c) kitāba Zayd
Le livre de Zayd

(d) al-kitāb
Le livre

(e) kitābun ǧadīdun
Un nouveau livre
(f) bi-kitābi Zayd
Avec le livre de Zayd
(g) bi-l-kitābi al-mufīd
Avec le livre intéressant

on peut voir que ces différentes séquences sont identiques du fait qu'elles font référence à la même entité. Elles sont toutes bâties autour du nom ('ism) *kitāb* ; nom auquel on peut concaténer des ajouts à gauche et/ou à droite sans lui faire perdre son caractère de séquence insécable du point de vue de sa réalisation. Toutes les séquences (a-g) peuvent occuper la même *position* (mawḍi') dans une construction donnée. Elles sont dites *lexies nominales* (l'équivalent du *Groupe Nominal* dans la littérature des *Phrase Structure Grammar*, ou PSG). Toutes les lexies nominales se moulent sur un seul schème générateur générique dit « *schème générateur de la lexie nominale* ».

En se projetant au niveau supérieur, les lexies nominales sont combinables, soit entre elles soit avec d'autres segments signifiants, en unités syntaxiques, que nous développerons dans ce qui suit et qui feront l'objet du présent document.

2.1.3. Niveau interlexical

La grammaire qu'on présentera par la suite a été conçue pour traiter du *bina'* au niveau tectonique. D'après [HADJ SALAH,1979], pour générer des énoncés de la langue arabe, il ne suffit pas de juxtaposer un ensemble de lexies entre elles ou avec d'autres unités telles que « inna » ou « kāna » ; les lexies ne font que partie d'un système qui génère des unités supra lexicales. Ces unités sont appelées *tectonies syntaxiques*. On dit que ces structures syntaxiques sont générées par *bina'*.

Par la suite, et pour des raisons méthodologiques, nous nous restreignons aux lexies nominales dont le noyau (ou aṣl) est un nom unique (ou 'ism mufrad).
Si on regarde les séquences suivantes :

(1) Zaydun muntaliqun
Zayd part
(2) Inna Zaydan muntaliqun
Zayd est partant
(3) Kāna Zaydun muntaliqan
Zayd partait

on remarque que (2) et (3) dérivent de (1) par addition (ziyāda) de « inna » et de « kāna ».

L'addition de « *inna* » par exemple impose un changement au niveau de la désinence casuelle (*-u* en –a, i.e. la lexie Zaydun se transforme en la lexie Zaydan).

Si l'on met en correspondance les constructions précédentes comme suit :

R	T1	T2
Ø	Zayd**un**	muntaliq**un**
Inna	Zayd**an**	muntaliq**un**
Kāna	Zayd**un**	muntaliq**an**

on remarque que ce sont les items en première position qui déterminent les dé-sinences casuelles contenues dans les autres composantes du *binā'*; ils sont res-ponsables de la flexion (ou 'i'rāb) :

- Ø (ou ibtidā', ou aussi position zéro), *inna* et *kāna* sont dits régissants (ou 'āmil-s) et occupent la position (ou mawḍi') R.
- *Zaydun* et *Zaydan* sont dits termes régis en première position et occupent la position T1. Ces termes sont régis au nominatif (raf') par la classe de « kāna »[1], et à l'accusatif (naṣb) par la classe de « inna »[2].
- *Munṭaliqun* et *munṭaliqan* sont dits termes régis en seconde position et occupent la position T2. Ces termes sont régis à l'accusatif (naṣb) par la classe de « kāna », et au nominatif (raf') par la classe de « inna ».
- T1 et T2 sont tous deux régis au nominatif (raf') par Ø.

3. Le modèle HPSG

Le modèle des HPSG a été conçu au début des années quatre-vingts par Carl Pollard et Ivan Sag. Ce modèle constitue un formalisme sans aucun doute actuel, résultat d'une évolution qui s'est produite au cours des trois dernières décennies ; il a essayé de s'approprier et d'unifier les héritages des différents formalismes qui l'ont précédé.

3.1. Généralités

HPSG est dite théorie monostratale (utilise un seul niveau de représentation dans lequel c'est la notion de partage d'information qui permet d'analyser les relations entre les éléments d'une structure linguistique ; elle ne recourt donc pas aux transformations) non dérivationnelle (dans la mesure où des règles de réécriture ne suffisent pas pour produire l'ensemble des structures possibles d'une langue). En terme d'analyse, elle se fixe comme objectif la constitution d'un système de contraintes sur les structures qu'elle construit. La satisfaction de ce système de contraintes est alors synonyme de bonne formation des structures. Ainsi, le modèle HPSG révise la notion de grammaire et la voit comme étant un système de contraintes sur les structures de traits, qu'il faut satisfaire. La bonne formation des objets linguistiques est alors un problème de satisfaction de contraintes. En terme de constructions (les structures qu'elle construit), HPSG articule les différentes connaissances linguistiques (phonétique, lexicale, syntaxique, sémantique et pragmatique), qui sont par essence hétérogènes, dans une seule représentation homogène. On parle alors d'intégration de connaissances. En terme formel, HPSG se base sur la notion de *structures de traits typées* (Typed Feature Structures, ou TFS) ; la notion de typage y joue un rôle majeur, à la fois pour ce qui concerne la représentation des connaissances et pour les mécanismes d'analyse.

[1] Classe de « kāna » : « sāra » = devenir, « laysa » = ne pas être, « mādāma » = tant que, etc. Ils sont dits verbes exponentiels (ve).

[2] Classe de « inna » : « ka'anna » marque la comparaison, « layta » le souhait, « la'alla » l'attente, etc. Ils sont dits exposants non verbaux (e).

3.2. Bases formelles

La structure de base de la représentation syntaxique en HPSG est la structure de traits typée. Pour faciliter la compréhension de son fonctionnement et de ses propriétés, nous allons présenter, en premier, la notion de structure de traits et ses caractéristiques.

3.2.1. Structures de traits

En général, pour représenter un objet (linguistique ou autre) on a besoin de spécifier ses attributs. On définit alors des ensembles de traits appelés *structures de traits* (Feature Structures, ou FS). Un trait est un couple *attribut-valeur* (Atribut-Value), les valeurs pouvant être des symboles atomiques ou des traits. Les traits à valeur non atomique conduisent à des structures de traits présentant des enchâssements.

La notation la plus pratique consiste à présenter les structures de traits sous forme de *matrice attribut valeur* (Attribute Value Matrix, ou AVM). Chaque ligne représente un trait (Feature), donc une paire attribut-valeur. Implicitement, des traits appartenant à la même structure sont coordonnés ; l'ordre des traits est sans importance (i.e. pas d'ordonnancement). Une condition de cohérence : une structure bien formée ne doit pas contenir deux fois le même attribut (au même niveau d'enchâssement) avec deux valeurs différentes.

Les caractéristiques essentielles d'une structure de traits sont :

- les éléments de la structure sont atomiques ou complexes (i.e. la valeur d'un attribut peut être une structure de traits, on dit qu'il y a *récursivité*) ;
- la structure interne d'un élément est définie par ses attributs et ses valeurs ;
- les valeurs peuvent être *partagées*[3] (on dit que les structures sont *réentrantes*).

3.2.2. Extension et unification

On définit d'abord une relation d'ordre partiel entre structures de traits dites d'*extension*. De façon intuitive, une structure plus spécifique est une extension d'une structure plus générale (ou moins spécifique). Cette relation d'ordre sert à définir l'opération d'*unification* de la façon suivante :

> L'unification de deux structures de traits A et B est la structure minimale qui est à la fois une extension de A et de B, si elle existe. Si elle n'existe pas, on dit que l'unification « *échoue* ».

De façon informelle, l'unification vérifie la compatibilité[4] entre deux structures de traits et produit une structure résultante qui est la plus petite structure qui contient toute l'information contenue dans la première structure et toute l'information contenue dans la deuxième structure.

---

[3] Il est important de faire la distinction entre *structures identiques* et *structures à valeurs partagées* (ou *réentrantes*) ; les secondes sont identiques et le seront quelles que soient leurs modifications ultérieures, mais pas les premières.

[4] En effet, l'unification de deux AVM A et B n'est possible que ssi A et B peuvent décrire le même objet (i.e. des informations compatibles).

3.2.3. Descriptions partielles (sous-spécification)

La structure de traits offre une grande souplesse de description, en ce qu'elle permet de manipuler des représentations *sous-spécifiées* (ou *partielles*) et *extensibles*. Ces représentations sous-spécifiées sont dites *descriptions*. Ce sont des structures de traits mais qui ne sont pas nécessairement complètes. Seules les traits concernés par la contrainte exprimée sont généralement présents dans la description. Et ces descriptions se complexifient (i.e. il y a enrichissement d'information) par unification. La présentation d'une grammaire se fait donc essentiellement par la présentation de descriptions, donc de structures incomplètes.

3.2.4. Structures de traits typées

La notion de *typage* joue donc un rôle fondamental dans la représentation des connaissances. L'intérêt d'utiliser un typage des structures de traits tient essentiellement dans la vérification de la cohérence des grammaires. Il est en effet impossible, dans ce modèle, d'introduire des traits non appropriés aux types.

La solution proposée dans le cadre du modèle HPSG consiste à adopter des structures de traits qui soient *typées* : tous les traits appartenant à une structure de traits doivent être appropriés au type de la structure. Le type d'un objet détermine donc les attributs appropriés pour l'objet. Ces attributs ont, à leur tour, pour valeur des objets typés. On définit, alors, une hiérarchie de types (en *types* et *sous-types*) à l'aide d'une relation d'ordre partiel. Les types les plus généraux sont les plus bas dans la relation.

3.3. La structure du signe en HPSG

Ce qui différencie HPSG des autres modèles[5] est sa volonté de donner des descriptions uniformes des différentes dimensions du langage. Cette uniformité de la modélisation se manifeste en ce que le modèle de toute unité est construit sur le même patron quelle que soit sa taille. En d'autres termes, utiliser les structures de traits comme cadre unique pour représenter des informations linguistiques de nature hétérogène. Ainsi, un mot (c'est-à-dire une unité du lexique) est représenté de la même manière qu'un syntagme ou qu'une phrase, voire un discours, tous ces objets étant des signes, qui ne sont à leur tour que des structures de traits typées. Les règles de grammaire, les principes généraux et les grammaires elles mêmes ne sont que des structures de traits typées.

Certes, un modèle mathématique qui autorise de telles représentations est nécessairement très peu contraint. Mais il ne peut en être autrement si on veut, par un formalisme unique, représenter des dimensions distinctes et hétérogènes.

En HPSG, les objets linguistiques sont de type *signe* (sign). Un *signe* associe une forme (phonologique) à un ensemble de propriétés grammaticales (syntaxiques, sémantiques, ...). Le type *signe* a pour sous-types *mot* (*word*) et *syntagme* (*phrase*). Les sous-types d'un super-type héritent des contraintes qui s'appliquent au super-type ; ainsi, les *mots* et les *syntagmes* héritent des traits associés aux *signes*. En plus, les signes lexicaux (word) ont une structure morphologique, alors que les syntagmes (phrase) sont

[5] Linguistique, bien sûr. Puisqu'il s'inspire largement des systèmes informatiques de représentation des connaissances, définis pour représenter n'importe quelle forme de connaissance.

décomposables en constituants (i.e. se caractérisent par la présence de traits supplémentaires encodant la structure en constituants).

3.4. Typage des objets linguistiques et héritage

Comme il a été signalé auparavant, la notion de typage est fondamentale en HPSG. Tous les objets linguistiques (*signes*) sont représentés sous forme de structures de traits typées rassemblant informations phonologiques, syntaxiques, sémantiques et pragmatiques. Le typage permet de préciser l'ensemble des attributs appropriés et de contraindre leurs valeurs ; les types sont organisés selon une hiérarchie. Les structures ainsi construites peuvent contenir des sous-structures récursives. L'organisation hiérarchique des traits et leur typage permettent d'éviter une prolifération anarchique et l'éclatement du nombre des catégories. Cette organisation permet aussi d'utiliser un principe d'héritage ; chaque sous-type hérite de toutes les spécifications du type supérieur.

Le nombre de type à distinguer ne peut dépendre que d'observations linguistiques empiriques. Une hiérarchie de types (avec le détail des subdivisions) a été proposée pour l'anglais dans (Pollard & Sag, 1987) et (Pollard & Sag, 1994). Une autre pour le français fait l'objet de travaux récents, notamment, [Abeillé, 1995] et [Tseng, 2003]. Pour l'arabe, aucune classification n'a été proposée jusqu'à l'heure actuelle.

3.5. Les principes de bonne formation

En HPSG, les principes de bonne formation des objets linguistiques résultent de plusieurs composantes. La première, réside dans le typage des structures syntagmatiques (cf. supra). La deuxième, les schémas de dominance immédiate (Immediate Dominance Schemata, ou ID Schemata), qui viennent renforcer le typage des structures syntagmatiques, définissent la structuration interne des types de signes syntagmatiques. Ils assurent cette tâche en décrivant l'organisation de la constituance des syntagmes (définissent les contraintes de bonne formation pour les relations hiérarchiques). Ces schémas sont en petit nombre. Il y a, ainsi, six schémas qui exposent les types de relations entre une racine et ses nœuds filles (Sag & Wassow, 1999)[6]. La troisième, sont les principes de partage et/ou de propagation des traits, qui valent souvent pour plusieurs schémas DI à la fois. Ils encodent les contraintes que certains types de structures doivent satisfaire. Par exemple, le *Principe de Valence* rend compte de la *sous-catégorisation*. Le principe indique comment retirer des constituants réalisés à partir d'une liste contenue dans un trait de valence. De tels principes peuvent vérifier une propriété donnée sur une structure totalement connue, comme ils peuvent, également, compléter des structures qui ne sont que particllement connues.

3.6. Les règles lexicales

En HPSG, les règles lexicales expriment des relations régulières entre différentes descriptions d'entrées lexicales, c'est-à-dire des relations entre structures de traits

[6] [Pollard & Sag, 94] définissait cinq schémas DI.

typées. On choisit, ainsi, de traiter un certain nombre de phénomènes syntaxiques[7], pour lesquels la théorie chomskyenne utilisait des transformations[8] (par des règles lexicales, qui font référence aux traits fonctionnels et non aux structures de constituants).

La notion de règle lexicale en HPSG est inspirée du mécanisme qu'on trouve en morphologie, où des règles de flexion et de dérivation peuvent relier entre elles différentes formes lexicales. D'ailleurs, les règles lexicales permettent la flexion des mots sans qu'il ne soit nécessaire de spécifier dans le lexique chacune des formes fléchies (il s'agit de règles de génération de formes de pluriel, féminines, etc.).

3.7. Le lexique

Le développement d'un lexique sert à encoder le sens et à guider l'analyse syntaxique. C'est dans le lexique que peuvent se trouver selon les formalismes les informations sur la morphologie, la sémantique, une partie de la syntaxe et aussi la phonétique.

En HPSG, le lexique est très riche. Les représentations lexicales contiennent un nombre important d'informations de différentes natures (lexicale, syntaxique, sémantique, pragmatique, ...).

4. La TNK et la grammaire HPSG

Dans ce qui suit, nous allons décrire un fragment substantiel de la grammaire de l'arabe à l'intérieur d'un système entièrement cohérent et ce grâce au choix théorique qui s'est porté sur la TNK et les HPSG. Pour cela nous allons essayer de représenter les constructions de deux niveaux d'analyse : celles du niveau intralexical (i.e. la LN) et celles du niveau interlexical (i.e. la tectonic), en exprimant à chaque fois le schème générateur propre à chaque niveau à l'aide de la grammaire HPSG.

L'un des points attrayants de la vision TNK en ce qui concerne la construction des phrases arabes est sa capacité de donner les structures qu'elles expriment sous une représentation arborescente très simple à lire et à décrire. Nous allons exposer dans ce qui suit, les représentations que nous avons mis au point pour expliciter les structures syntaxiques de l'arabe, énumérées et exprimées par [HADJ-SALAH, 1979].

4.1. Représentation au niveau interlexical (syntaxique)

Dans cette section, nous allons voir comment le formalisme des HPSG permet de représenter les structures proposées par la TNK pour un traitement automatique de la langue arabe. Nous présenterons, à chaque fois que c'est nécessaire, un ou plusieurs exemples qui éclairciront les analyses qu'on a développé.

4.1.1. Représentation des constructions nominales

Une grammaire pour les CN est loin d'être contourné pour le moment. Nous commencerons, dans cet article, à décrire les brins d'un noyau de base qui servira à

[7] E.g. la construction infinitive *Jean veut dormir*, est dérivée de *Jean veut que Jean dorme*, par une règle d'effacement, la passive *Jean est soigné par Marie* est dérivée à partir de la phrase active *Marie soigne Jean* par les règles de déplacement et d'insertion.

[8] Dans le modèle de la Théorie Standard, il s'agit des règles transformationnelles qui sont utilisées pour déplacer, effacer ou insérer des éléments.

cerner et dégager les différentes façades de notre future grammaire. Nous expliquerons dans ce qui suit les cas de *inna* et de *kāna.*

1er cas : CN ayant comme régissant un exposant non verbal (classe de *inna*)

C'est l'ensemble des constructions de la forme :

$CN_{Re} = \{R_e, T1, T2\}$

Re peut régir un amalgame de lexies et de tectonies[9], mais pour simplifier nous dirons que R_e peut régir deux unités, la première une LN à l'accusatif, et la seconde qui peut être soit une LN au nominatif soit une Tectonie syntaxique au nominatif (N.B. Cette contrainte sous-entend que l'ordre des termes régis est figé, i.e. R_e suivi de T1 suivi de T2).

C'est la principale contrainte qui caractérise la classe de *inna* ; bien que d'autres pistes de contraintes sont toujours à explorer. Pour le moment, cette contrainte suffit pour une modélisation claire et extensible.

Cette information pertinente doit, bien sûr, figurer dans l'entrée lexicale prévu pour l'exposant *inna* et dans les exposants de sa classe. Cette entrée lexicale aura le squelette suivant :

$$\begin{pmatrix} \underline{Classe\text{-}inna} \\ \text{HEAD} \quad ① \ [\text{ORTH "inna"}] \\ \text{VALENCE} \ [\ \text{COMPS} < ②, ③ > \ ] \end{pmatrix}$$

Passons maintenant aux relations syntaxiques (de dépendances) que peuvent entretenir les différents constituants d'une tectonie syntaxique.

Pour rendre compte du présent cas de figure, nous avons développé un schéma de règles (schéma DI) qui stipule que :

Schéma DI **Head-Complements-R_eT1T2-inna**[10] :

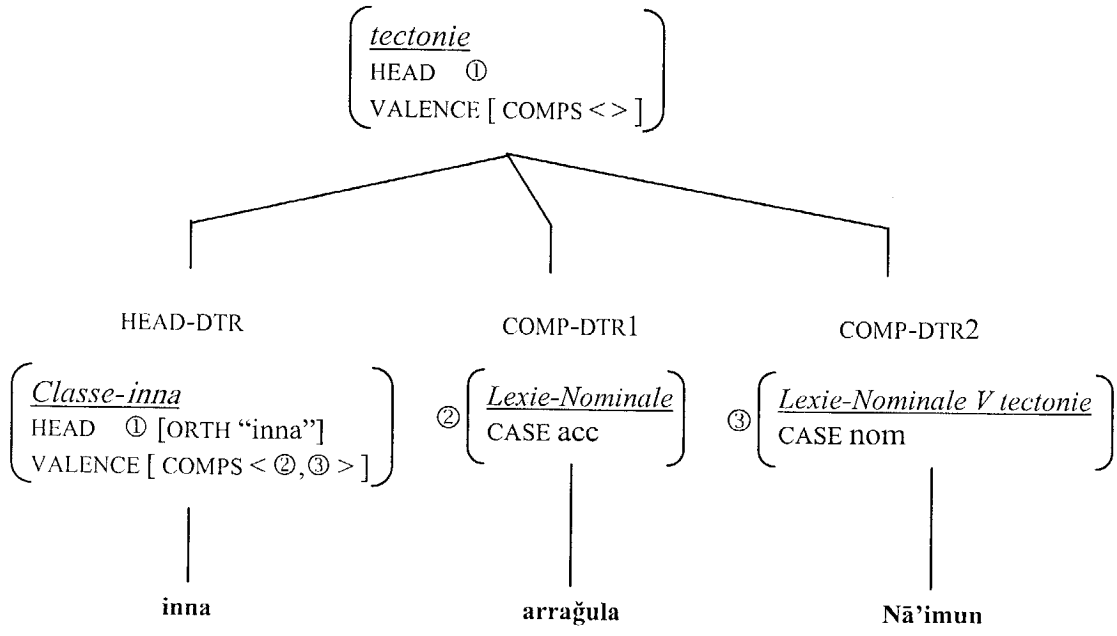
La racine est un signe saturé.
Le fils-tête est un signe lexical classe de *inna.*
Le fils-comps est la liste[11] T1[acc], T2[nom].

Pour décrire plus visiblement le schéma de DI sus-énoncé, nous prenons comme exemple la construction :

(1) inna arraǧula nā'imun.

Et nous opterons pour le modèle en constituants qui permet d'aboutir à l'arbre de constituants qui suit :

[9] Cela peut faire l'objet d'une exploration.
[10] Ce schéma est une version adaptée qui remplace partiellement celle établit par [ps 94] pour l'anglais (*head-complement-phrase*).
[11] Une liste est par essence ordonnée.



Nous n'avons jusqu'à maintenant examiné la structure syntagmatique de la Construction Nominale à régissant R_e que sous l'angle de son schéma de DI. On doit à ce niveau incorporer deux principes essentiels de partage de traits, en l'occurrence le ***Principe des Traits de Tête*** **(Head Feature Principle, ou HFP)** qui permet la propagation de certaines propriétés du fils-tête à la racine, et le ***Principe de Valence*** **(Valence Principle, ou ValP)** qui permet de contrôler étape par étape la saturation d'un signe syntagmatique et qui agit en interaction avec le HFP. Ces principes universels, qui ont été repris de [P&S 94], ont un rôle très important lors des unifications des descriptions pour aboutir à la structure finale de la construction. En termes très simples, ces deux principes peuvent s'énoncer comme suit :

Le HFP :

La valeur du trait de tête de la catégorie racine est égale à la valeur du trait tête du fils-tête.

Autrement dit :

La valeur de l'attribut tête du signe lexical doit être partagée par l'ensemble des syntagmes qui le contiennent.

Le ValP :

La valeur de chaque trait de valence du syntagme est égale à la valeur de chaque trait de valence correspondant du fils-tête moins les valeurs des compléments qu'il a déjà instancié.

A ce niveau de l'analyse, on peut voir en toute transparence l'interaction des différentes composantes grammaticales des HPSG. Nous allons ainsi examiner la représentation arborescente ci-dessus et illustrer les mécanismes y opérant pour analyser la construction (1) :

- les nœuds terminaux correspondent aux entrées lexicales ;
- la tête lexicale de la construction est l'exposant *inna*. La description des compléments qu'il sous-catégorise est donnée dans l'entrée lexicale du signe

(fils-tête). A partir de là, l'exposant doit instancier ses compléments pour que l'ensemble devienne une tectonie saturée. Cette étape nécessite l'introduction du schéma de DI **Head-Complements- R_cT1T2-inna** ;

- une nouvelle structure étant montée, aussi faudrait-il instancier les valeurs des traits de la tectonie ainsi formée (i.e. compléter ce qui est manquant). A ce niveau, les deux principes qu'on vient d'énoncer interviennent. En premier, le **ValP** touche aux traits de valence, ici seul COMPS est concerné. La valeur COMPS de la racine est égale à la liste COMPS de l'exposant moins les descriptions des compléments déjà instanciés, ce qui donne la liste vide. Ensuite, le **HFP** permet de remonter la valeur de l'attribut HEAD de la tête lexicale (i.e. l'exposant) à la tectonie supérieure. Sur l'arborescence la valeur ① remonte vers la tectonie. On dit de la structure ainsi obtenue qu'elle est saturée, sa liste de valence est vide.

2è cas : CN à régissant un verbe exponentiel (classe de *kāna*)

C'est l'ensemble des constructions de la forme :

$CN_{Rve} = \{R_{ve}, T1, T2\}$

L'analyse de ce type de construction a permis de dégager la similitude existante entre la classe *kāna* et la classe *inna* à un niveau très intéressant. De quoi il s'agit ? En fait, on peut sans trop tarder expliciter le schéma de DI suivant pour prendre en compte la classe de *kāna* dans notre grammaire :

Schéma DI **Head-Complements-R_{ve}T1T2-kāna**[12] :

La racine est un signe saturé.
Le fils-tête est un signe lexical classe de *kāna*.
Le fils-comps est la liste[13] T1[nom], T2[acc].

Mais un retour en arrière vers le schéma DI de la classe de *inna*, nous pousse à nous poser la question suivante : pourquoi ne pas avoir un schéma de règle plus générale pour les deux classes ? La réponse à cette question est positive, et cela bien sûr grâce à l'une des caractéristiques les plus frappantes du formalisme des grammaires HPSG, qu'est la lexicalisation. En fait, l'approche lexicale des HPSG conduit à un nombre très restreint de règles syntagmatiques extrêmement schématiques accompagnées par un lexique très riche et complexe. Alors, pourquoi ne pas délocaliser les contraintes concernant le cas des termes régis vers les entrées lexicales des régissants correspondants ? C'est ce que nous avons adopté pour notre grammaire. Nous aurons alors pour nos deux classes de régissants qui précèdent quelque chose comme :

Schéma DI **Head-Complements-ReT1T2**[14]:

La racine est un signe saturé.
Le fils-tête est un signe lexical classe de *inna* ou classe de *kāna*

[12] Ce schéma est une version adaptée qui remplace partiellement celle établie par [p&s 94] pour l'anglais (*head-complement-phrase*).
[13] Une liste est ordonnée.
[14] Ce schéma est une version adaptée qui remplace partiellement celle établie par [p&s 94] pour l'anglais (*head-complement-phrase*).

Le fils-comps est la liste[15] T1, T2.

Avec une réadaptation des entrées lexicales de *inna* et de *kāna* comme suit :
Entrée lexicale de *inna* :

$$\begin{pmatrix} \underline{Classe\text{-}inna} \\ \text{HEAD} \quad ① \text{ [ORTH "inna"]} \\ \text{VALENCE [COMPS} < ② \text{ [acc]}, ③ \text{ [nom]} >] \end{pmatrix}$$

Entrée lexicale de *kāna* :

$$\begin{pmatrix} \underline{Classe\text{-}\ kāna} \\ \text{HEAD} \quad ① \text{ [ORTH "}kāna\text{"]} \\ \text{VALENCE [COMPS} < ② \text{ [nom]}, ③ \text{ [acc]} >] \end{pmatrix}$$

Représentation des constructions verbales
C'est l'ensemble des constructions de la forme :

$$CV = CV_{Rvne\text{-}itr} \cup CV_{Rvne\text{-}tr} \cup CV_{Rvne\text{-}dtr}$$

Avec :

- $CV_{Rvne\text{-}itr} = \{R_{vne\text{-}itr}, T1\}$
- $CV_{Rvne\text{-}tr} = \{R_{vne\text{-}tr}, T1, T2\}$
- $CV_{Rvne\text{-}dtr} = \{R_{vne\text{-}dtr}, T1, T2, T3\}$

La CV est la combinaison d'une tête, qui est une lexie verbale (LV), suivie de 1, 2 ou 3 compléments. Dans ce qui suit, nous allons simplifier le composant LV et nous nous limitons à une LV restreinte à son noyau. Et là aussi, nous nous limitons à la sous-composante *fī'l* (i.e. j'éliminerais dans mon analyse la sous-composante *ḍamīr* du noyau de la LV)[16].

Le schéma de règle (schéma de DI) que nous avons développé ci-dessus a toutes les chances d'être généralisé et appliqué aux CV-s.

Intuitivement, nous aimerions avoir une règle exprimant simplement qu'une tectonie se compose, ainsi, d'une tête lexicale suivie par n'importe quelles autres unités syntaxiques (des lexies, des segments signifiants, des tectonies) que la tête lexicale exige. Nous pourrions alors renvoyer au lexique (de la manière dont nous avons fait ci-dessus) la question de spécifier pour chaque tête lexicale quels éléments celle-ci exige (ou cooccurre avec).

Dans notre présente grammaire, l'entrée lexicale pour un verbe intransitif tel que *qāma* devrait stipuler que ce dernier sous-catégorise (régit) un seul terme régi qui n'est autre que T1[17]. Cela se fait grâce au trait de valence COMPS qui est une liste de structures de traits. Intuitivement, cette liste spécifie une séquence de catégories correspondant aux termes régis (compléments) dont la tête se combine avec. Ainsi, on pourra synthétiser toutes les différentes règles pour le développement d'une cons-

[15] Une liste est ordonnée.
[16] Pour mémoire, le noyau de la LV est le couple (fi'l, ḍamīr).
[17] Une question peut se poser à ce niveau : pourquoi parler en terme de T1 bien qu'il s'agisse du fā'il du verbe, ici ? Nous ne voulons pas répondre à cette question à ce stade de l'analyse, et nous continuerons d'utiliser le terme T1 à la place du fā'il, par souci de cohérence au niveau de la TNK, jusqu'à ce que nous entamerons la composante sémantique.

truction (nominale et verbale), tout simplement en réarrangeant la règle précédemment révisée en :
Schéma DI **Head-Complements-R$_c$Ti** :

$$\begin{bmatrix} \underline{\textit{tectonie}} \\ \text{HEAD } ① \\ \text{VALENCE [COMPS < >]} \end{bmatrix} \rightarrow \begin{bmatrix} \underline{\textit{régissant}} \\ \text{HEAD } ① \\ \text{VALENCE [COMPS < ②,③, ...,Ⓟ >]} \end{bmatrix} ② \ ③ \ \ldots \ Ⓟ$$

En paraphrasant :

> La racine est un signe saturé.
> Le fils-tête est un signe lexical[18].

En bref, la liste COMPS d'une entrée lexicale spécifie des conditions de cooccurrence d'unités syntaxiques ; et la liste COMPS d'un nœud tectonique est vide. Ainsi, un régissant R doit avoir des termes régis qui correspondent (ou plutôt répondent) à toutes les TFSs dans sa valeur COMPS, et la tectonic qu'il tête a la liste vide comme sa valeur COMPS et donc ne peut pas se combiner avec des termes régis. La règle Head-Complément, comme exposée, exige la réalisation de tous les termes régis par la tête lexicale.

Nous illustrons cela avec la tectonie *ḍahaba karīmun ila almadrasati*. Le verbe *ḍahaba* a besoin aussi bien d'un T1 que d'un T2. Ainsi, la valeur de son COMPS est <T1,T2>. Les requis T1 et T2 devront être tous les deux des sœurs (on pense en termes de construction de l'arbre ascendant) du verbe régissant *ḍahaba*. Comme ci-dessus, en tout les 3 combinés pour former une tectonie dont les besoins en termes régis ont été réalisés (remplis).

Comme il est apparaît clairement d'après cet exemple, nous supposons que les éléments dans les valeurs de COMPS apparaissent dans le même ordre que dans la tectonie. Nous continuerons à user de cette supposition, bien qu'en fin de compte un traitement plus raffiné de l'ordre linéaire des unités syntaxiques (termes régis) dans la tectonic serait nécessaire. En fait, l'ordre des mots en HPSG est réglé en dehors des règles de grammaires (et c'est ainsi d'ailleurs que les règles de grammaires sont appelées ***Immediate Dominance Schemata***) qui précisent uniquement les relations de dominance immédiate.

4.2. Représentation du niveau intralexical : Représentation de la Lexie Nominale (LN)

Pour des raisons de simplification de l'analyse, par accroissement en complexité, et pour une fluidité du document que nous sommes en train de présenter, nous avons préféré présenter des termes régis réduits à des *'ism* (noms). Le *'ism*[19] a pour positions premières les positions où sont régis les items (Ti). Réciproquement, un item régi par un

[18] Qui peut être de la classe de inna, de la classe de kāna, un verbe intransitif, transitif ou bi-transitif.
[19] Le 'ism englobe, dans le naḥw (la syntaxe de l'arabe), le nom commun et le nom propre, les pronoms, les adverbes, ainsi que tous les éléments qui n'entrent pas dans la classe du verbe (fi'l) ni dans celle des exposants (ḥurūf alma'ānī).

autre ne peut être dans son *aṣl* qu'un *'ism*. Cela dit, il existe des structures syntaxiques où les termes régis sont constitués par des LN ou des tectonies aussi bien à l'intralexical qu'à l'interlexical. On pourra voir l'exemple :

(2) Inna aṭṭifla almuǧtahida yadrusu al-alsuniyyata.
L'enfant sérieux étudie la linguistique.

Nous allons dans ce qui suit voir comment va se faire l'analyse dans le cas général, ensuite nous développerons les schémas de DI qui prennent en charge les différentes combinaisons. Une fois cela terminé, nous verrons dans quelle mesure nous pouvons faire des généralisations plus intéressantes.

A ce niveau de l'analyse, on va voir la LN non comme une liste qui consiste en un noyau, éventuellement précédé et/ou suivi par des positions (mawḍi'-s) [HADJ SALAH, 1979], mais comme étant une liste[20] formée d'une première sous-liste contenant les positions qui précèdent le noyau, le noyau lui-même, et une deuxième sous-liste contenant les positions qui suivent. Cette révision de la notion de LN permet une meilleure modularisation de l'information et par là même la possibilité d'exprimer des propriétés très générales. Il devient en effet possible de distinguer clairement celles qu'on pourrait considérer comme des positions compléments et celles qu'on peut considérer comme des positions spécificateurs[21].

Dans l'exemple (2) ci-dessus, le régissant *inna* exige les deux compléments T1 et T2 pour générer une tectonie saturée. Mais *inna*, avant de se satisfaire de son premier complément T1, impose d'en faire l'analyse préalable (i.e. être une LN saturée, dans notre cas). Donc, le noyau de la LN (représentant T1) doit satisfaire ses besoins en compléments en premier. Cela ne suffit pas, puisque ce même noyau doit satisfaire ses besoins en spécificateurs. C'est à ce moment là seulement que la LN devient saturée. Ce qui a été dit pour T1 sera redit pour T2, qui est ici une tectonie dont le régissant est un $V_{ne\text{-}tr}$; pour pouvoir analyser la tectonie (inna,T1,T2), on doit aussi attendre la saturation de T2 (yadrusu al-alsuniyyata).

Dans ce qui suit, on ne s'intéressera qu'au premier terme régi, qui représente la LN, objet de notre présente analyse.

A la manière de [Sag & Wassow, 2003], on proposera deux schémas de DI, l'un concernant les compléments, l'autre, pour les spécificateurs.

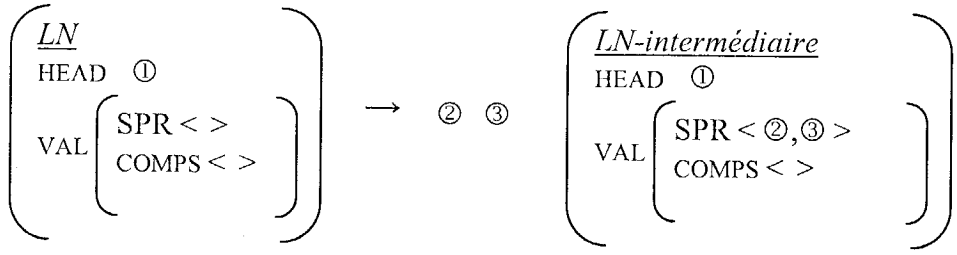
Schéma DI **Head-Complements-LN** :

$$\begin{bmatrix} \underline{\textit{LN-intermédiaire}} \\ \text{HEAD} \quad ① \\ \text{VAL} \begin{bmatrix} \text{SPR} < ②,③ > \\ \text{COMPS} < > \end{bmatrix} \end{bmatrix} \rightarrow \begin{bmatrix} \underline{\textit{Noyau-LN}} \\ \text{HEAD} \quad ① \\ \text{VAL} \begin{bmatrix} \text{SPR} < ②,③ > \\ \text{COMPS} < ④,⑤,⑥ > \end{bmatrix} \end{bmatrix} \quad ④ \quad ⑤ \quad ⑥$$

[20] Une liste est par essence ordonnée.

[21] Selon l'analyse de [HADJ-SALAH, 1979], tous les éléments, à l'exception du noyau, occupant des positions au niveau du schème générateur de la LN, sont des déterminants. Cette vision va à l'encontre d'une généralisation au niveau des schémas de DI qu'on se propose d'établir.

Schéma DI **Head-Specifiers-Rule** :



Cette manière de voir la LN reste encore problématique. En effet, une tête lexicale exige ses compléments pour former un syntagme saturé, alors que dans notre cas, le noyau de la LN, qui est supposé être la tête, n'est pas toujours en mesure d'exiger quoi que ce soit, ce qui peut compliquer les choses au niveau de l'implémentation. Pour le moment, nous gardons à l'esprit deux pistes de solutions possibles, qu'il va falloir explorer par la suite. La première, au niveau du lexique, par une multiplication des entrées lexicales. Une deuxième, non moins intéressante, qui consiste à proposer un nouveau principe qui gère la saturation de la LN.

Conclusion et perspectives

Un des avantages de l'utilisation du formalisme des grammaires HPSG réside dans le fait qu'actuellement beaucoup d'équipes de recherche travaillent sur l'implémentation de ce formalisme, ce qui est d'un apport considérable pour les travaux portant sur la langue arabe.

Comme il a sans doute été remarqué, nous avons repoussé à des travaux futurs des cas d'analyses très délicates, qui nécessitent d'ailleurs plus d'espace que ce document ne le permet. Nous citerons, à titre d'exemples seulement, deux problèmes très intéressants. Le premier concerne les alternances exclusives qui existent entre un sous-ensemble des éléments de la LN. Ces alternances ne peuvent être explicitées au niveau des schémas de DI. On s'orientera vers le système de typage, puisqu'au niveau du lexique, HPSG ne présente pas d'opérateur (ou exclusif) qui nous permettra de modéliser les alternances exclusives. Le deuxième problème concerne les différents cas d'accord qui peuvent intervenir aussi bien au niveau intralexical qu'au niveau interlexical.

Il est à noter que ce noyau de grammaire pourra faire l'objet de révision et de remaniement selon les résultats obtenus après implémentation (feedback). En effet, la projection du fragment de grammaire présenté dans ce document dans une plate-forme informatique telle que la LKB, permettra de déceler les insuffisances et les révisions qui s'imposent.

BIBLIOGRAPHIE

Abeillé, A., (1993), *Les nouvelles syntaxes : grammaires d'unification et analyse du français*, Paris, Armand Colin.

Carpenter, B., (1992), *The Logic of Typed Feature Structures*, Cambridge University Press.

Copestake, A., (2002), *Implementing Typed Feature Structure Grammars*, CSLI Publications, Stanford, Ca.

Gazdar, G., Klein, E., Pullum G. K., & Ivan A. Sag, (1985), *Generalized Phrase Structure Grammar*, Cambridge, MA : Harvard University Press.

Hadj-Salah, A., (1979), *Linguistique arabe et linguistique générale : Essai de méthodologie et d'épistémologie du 'ilm al-'arabiyya*, Thèse de Doctorat, Paris, Sorbone.

Mammeri, M. F., (2003), *Une approche pour l'analyse syntaxique de l'arabe basée sur la théorie néo-khalilienne et la Grammaire Syntagmatique Guidée par les Têtes.* Mémoire de Magister, CRSTDLA-ENSLSH, Alger.

Pollard, C. & Sag I.A., (1994), *Head-Driven Phrase Structure Grammar*, Chicago, University of Chicago Press.

___, (1987), *Information-based Syntax and Semantics*, volume 1: Fundamentals, CSLI Publications, Stanford, Ca.

Sag, A. I., Bender, E. Wasow, T., (2003), *Syntactic Theory : A Formal Introduction, 2nd Edition*, CSLI Publication, Stanford University.

Shieber, S.M., (1990), *Les grammaires basées sur l'unification*, In Miller P. & Torris T. (eds) *Formalismes syntaxiques pour le traitement automatique du langage naturel*, Hermès, pp. 27-85.

___, (1986), *An Introduction to Unification-Based Approaches to Grammars*, CSLI Publications, Stanford, Ca.