Similarities and differences between GPSG and HPSG grammars applied to the arabic language

ACHIT Abdelmadjid(1) **AZZOUNE Hamid(2)**

1) Natural Language Processing Laboratory, Scientific and Technical Research Center for the Arabic Language Development CRSTDLA
BP 225 Rostomia-Hamadia Bouzaréah 16011 Algiers-Algeria
achit@wissal.dz

(2) Artificial Intelligence Research Laboratory LRIA, Computer Science Department, Houari Boumèdiene USTHB University
BP 32 USTHB 16123 Bab-Ezzouar Algiers, Algeria
hazzoune@wissal.dz

Abstract:

The development of an efficient lingware, for any language, requires first the modeling and the formalisation of the linguistic knowledge of it. In fact, a multitude of grammatical formalisms exist, however none of them had made the unanimity of specialists of the domain. This raised the problem of choosing the most appropriate formalism for representing linguistic constructions of that language. Thus, this paper is about the discussion of the aptitude of the Generalized Phrase Structure Grammar GPSG and the Head-Driven Phrase Structure Grammar HPSG to represent the different analysis' levels of the arabic language. So, a comparative study is carried out between these two formalisms, in order to evaluate their capacities to represent and give account of the different analysis' levels of the arabic by taking foundations on an arabic linguistic theory. This study aims to select between these two formalisms, on the very definite criteria , the one that will be adopted to act as basis to the conception of our NLP tools for the arabic language processing. This work succeeds to the fact that HPSG as a formalism representing the different analysis' levels of the arabic is more recommended than GPSG. Thus, it was established that HPSG proposes a richer and more organized lexicon than GPSG. Also, regarding the representation of arabic syntactic constructions, HPSG offers a more elegant representation, more general and more economic in term of cost: type and number of used rules.

Keywords: Natural language processing NLP, Unification grammar, HPSG, GPSG, syntactic analysis, arabic language processing, comparative syntax.

1. Introduction

The multiplicity of grammatical formalisms already existing and which continue to appear, lead us to seek among them after the one that provides the best linguistic representation: richest in information, most economic from the point of view that it doesn't contain redundant information and which uses a less number of rules and ensures the largest linguistic cover. This paper is concerned with a comparative survey of *Generalized Phrase Structure Grammars GPSG* with *Head Driven Phrase Structure Grammars HPSG* applied to the arabic language according to a selected set of criteria.

2. Presentation of GPSG [GAZ, 85],[TEND, 99],[VOL,88]

Generalized Phrase Structure Grammar GPSG, is the result of G.Gazdar and.Klein, G.Pullum and I.Sag researches at the begin of eighties [GAZ, 85]. GPSG is based on unification and is viewed as the resumption and the extension of Chomsky's initial research program started in the fifties. It's a phrasal model because, it founds its analysis on phrase structure rules and it's generalized because it doesn't describe all possible constructions using an infinite addition of rules but rather uses general principles.

GPSG		
GPSG linguistics features system		A syntactic category is defined as a feature structure which consists in a finite set of couples <attribute, value>. Syn-cat={<att1, v1>,<att2, v2>,<att3, v3>, ...}
GPSG syntactic rules system	**ID rules**	They express the immediate dominance relation between a phrase and its immediate constituents independently of their linear order. ID rule: cat0→ cat1, cazt2, cat3, ..., catn
	LP rules	They constrain the orderof the right member categories of ID rules. LP rule: cat0→ cat1, cat2, cat3, ..., catn
	Meta rules	They are mechanisms which produce a new ID rules from ID rules in entry. Meta-rule= {Model → Target; alpha→A → beta→B}
GPSG Feature Instantiation Principle FIP	**Head Feature convention HFC**	In a local tree, the mother category and its head daughter category share their head features.
	Foot Feature Principle FFP	Foot features instantiated by the mother category of a local tree must be identical to the unification of foot features instantiated by its daughters categories.
	Control Agreement Principle CAP	This principle assure the management of the agreement problems between the different categories to resolve this problem, GPSG uses the CAP which assign a semantic type to every category of rule.
Feature Cooccurrence Restrictions FCR		It defines a constrain on the cooccurrences of features in a feature structure, because not all possible combinations of feature make up a category. In GPSG, all legal category must satisfy all FCR . **Example:** FCR :[VFORM] ⊃ [-N, +V] This rule means that all category with a value for the VFROM feature must be [-N, +V].
Feature Specification default FSD		This mechanism enable us to specify values by default for certain features. **Example:** FSD: ~[Case Nominative] This rule means that [case nominative] doesn't appear on a category except when it is required or specified by a rule.

Table 1: Summary of the main concepts of GPSG

3. Presentation of HPSG [ANN, 2000],[ACH, 2003],[BLA,95],[P&S, 94]

Head-Driven Phrase Structure Grammar HPSG is a non derivational monostratal linguistic theory. It was developed in the mid of eighties by Carl Pollard and Ivan Sag as an alternative to the transformational grammar. This formalism belongs to the unification grammars family and has been greatly influenced by a number of linguistic theories such as the GPSG, LFG, P&P, Categorial grammar, Functional grammar. Also, computer science with related areas such as logic, set theory influenced HPSG. Since 1994, there is an annual HPSG conference every year.

HPSG	
Main characteristics of HPSG	1. Integration of heterogeneous linguistics information (phonological, syntactic, semantic and pragmatic) in the same homogeneous representation which is the feature structures. 2. monostratal theory (one level of analysis) 3. non derivational 4. constraint-based 5.unification-based 6. surface oriented grammatical architecture 7. highly lexicalised 8. declarative 9. It uses typed feature structures 10. it uses a sharing values mechanism within the feature structures. 11. it uses a multiple inheritance for the features
Representation of linguistics signs	In HPSG, everything is modelled in feature structures. Thus, they are used to represent lexical entries, grammar rules and grammar principles **Feature structures:** A feature structure is encoded by an Attribute-Value-Matrixes (AVMs). Feature structures properties: - feature structure is a sign - feature structures are of a certain type - types are organized in hierarchies - type definition says what features are appropriate for a structure of the defined type. -they are recursive or re-entrants structures In HPSG, we distinguish two sorts of signs: word and phrase.

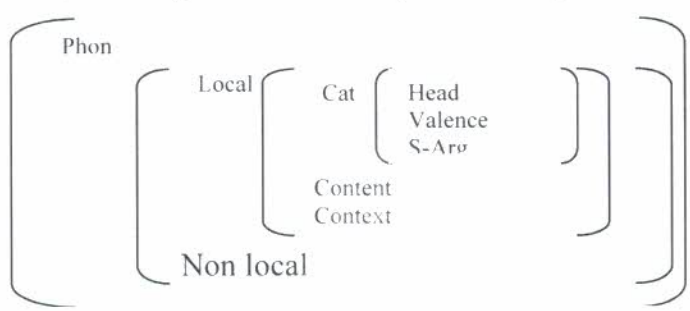

Fig 1: feature structures of a word sign

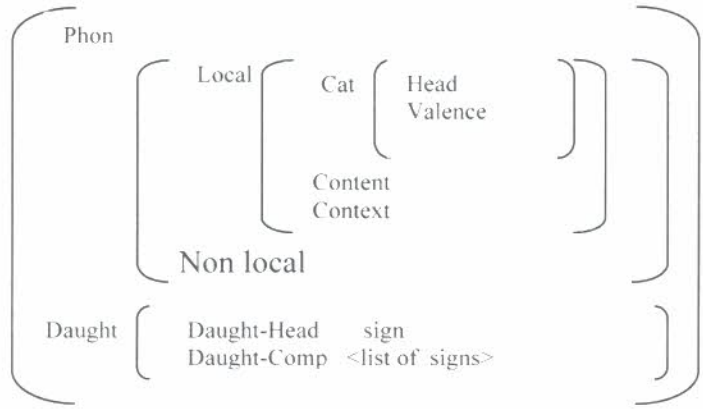

Fig 2: feature structure of a phrase sign

Type hierarchy of a sign

Sign			
Word		**Phrase**	
Lexical	**Functional**	**Headed str**	**No-headed structure**
Noun	Determinant	Head-compr	...
Verb	Marker	Head-Sub	
Adjectiv		Head-sub-comp	
Adverb		Head-mark	
Preposition		Head-adj	
		Head-spec	

Fig 3: type hierarchy of a sign

HPSG lexicon		The lexical entries are very rich and very complex. According to [Carnie, 99], there are a lot of information stored in the lexical entries of the HPSG grammars. In most cases, these information are redundant or predicative. Thus, HPSG uses type hierarchies and lexical rules (plural rule's, the passive form rule's, etc) to reduce the quantity of redundant / predicative information found in the lexical entries.
Signs combinations in HPSG	**ID Schemas**	Head-Subject schema Head-Specifier schema Head-Complement schema Head-Marker schema Head-Adjunct schema Head-Filler schema Head-Subject-Complement schema
	Well formation principles	Head Features principle Valence principle Non local feature principle Marker principle Specifier principle Weak coordination principle Semantic principle

Table 2: Summary of the main concepts of HPSG

4. Overview on the Neo-Khalilien linguistic theory

4.1. Brief historic

The Neo-khalilien linguistic theory NKT has been developed by the Professor HADJ-SALAH Salah[Hadj,79]. It got its origin and foundations from the ancient works of arabic grammarians: Al-Khalil Ibn Ahmed El-Farahidi and his disciple Sibawayh. What is to be noted in the work of those famous arabic grammarians is their use of the logical-mathematical concepts in order to describe and explain the different linguistic phenomena. The NKT is a linguistic theory based on syntax. It is mainly centred around the concept of generator pattern (considered as a *production / recognition model*) that intervenes in all arabic language analysis levels.

4.2. Analysis level of the arabic language according to the NKT

4.2.1. Kalima

It can consist in either a noun (common or proper) or a verb or a functional word.
We distinguishes two classes of kalima: regular (common noun, verbs,..) and et non-regular (functional words, proper nouns and some common nouns.)

4.2.2. Lexie

All isolable sequence that admits or no additions by simple concatenation without loosing its character of indivisible sequence from the point of view of its realization.

Example: لباس جديد، سيارة عمر القديمة، ذهب، قد ذهبوا،ما قد ذهبن are lexies.

4.2.2.1. Lexie's generator pattern

It is constituted by a core and positions (مواضع) structured so that they respect intralexicale relations existing within a lexie. Incrementations by ante-position or post-position in relation to the core on the horizontal axis are possible.

4.2.2.2. Type of lexie

According to the type of the core, we distinguish two sorts of lexies: the nominal lexie and the verbal lexie.

a. Nominal lexie

In this type of lexie, the core is a noun and the generator pattern that governs this lexie is of the following form:

Charachteriser / adjunct	Tanwin/adnominal complement	flexion	Core	determinant	Genetiv particule
			كتاب		
			كتاب	ال	
مفيد	ن	ُ/	كتاب	–	
هذا	زيدٍ	ُ/	كتاب	–	
المفيد	زيدٍ	ُ/	كتاب	–	
المفيد	–		كتاب	ال	ب
الذي هو هنا	زيدٍ		كتاب	–	ب
←	←	←	↔	→	→
3	2	1	0	1	2

Fig 4: Nominal lexie's generator pattern

b. Verbal lexie

The core of a verbal lexie is binary. It is constituted by a verb + affix pronoun. We distinguish, in the arabic language, three generator patterns which correspond to three verbal modes (accomplished, unaccomplished, imperative).

affixed pronoun	core	Exposant	Converter
←	↔	→	→
1	0	1	2

Fig 5: Generator pattern of the accomplished mode

4.2.3. Tectonie

It is defined as sequence of lexies. The generator pattern describing a tectonie, more abstract than the one of the lexie, is represented by an ordered couple (governor, governed items) equivalent to the following formula: **[[R→ T1]+/-T2]+/-D**

R: *governor, it exercises influence on the flexion and on the meaning of the subsequent elements.*

Ti: *governed item, it can be either a lexie or a tectonie. It's an item governed by a governor which affect its flexion. The item T1 can't never be placed before the R.*

D: *syntactic determinant or peripheral unit.*

Example:

D	T2	T1	R
البَارحة	زيدًا	مُحمدُ	لقِيَ

Fig 6: example of a tectonie

5. Comparison of the two formalisms

Linguistic features system
Syntactic rules system
- ID rules
- LP rules
- Meta-rules

Feature instantiation principles FIPs
- HFC
- FFP
- CAP

FCR
FSD
Operation: Unification

[I] Comparison independently from any langugae

Before the modelisation

Linguistic features system
Immediate domination schema system
Head-Subject / Head-Complement / Head-Subject-Complement / Head-Specifier / Head-Marker / Head-Adjunct/ Head-Filler
Well formation principles
Head Features principle/ Valence principle / Non local feature principle/ Marker principle/ Specifier principle/ Weak coordination principle
Semantic principle

Operation: Unification

Fig 7:GPSG and HPSG grammatical systems to be compared

- Linguistics Features system
Features set={features already existing }∪{ features introduced to give account appropriately about the arabic lexicon and constructions}
- Syntactic rules system
- ID rules for the arabic constructions
- LP rules for the arabic constructions
- Meta-rules proprer to the arabic

- Feature instantiation principles FIP
We keep the same set of FIP
- FCR proper to the arabic constructions
- FSD proper to the arabic constructions
- Operation: unification

[II] Comparison in relation to their application on the arabic language

After the modelisation

- Linguistics Features system
Features set={features already existing }∪{ features introduced to give account appropriately about the arabic lexicon and constructions}
- ID Schemas System
{ID schemas already existing}∪{ID schemas proper to the arabic constructions}

- Well formation Principles
{Principles already existing}∪{Principles proper to the arabic constructions}
- **Operation:** d'unification

5.1. Comparison between GPSG and HPSG independently from any language [ACH, 2003]

5.1.1. Similarities between GPSG and HPSG

1. avoidance of transformations
2. conception of the syntactic categories as feature structures
3. use of the unification operation
4. adoption of the ID / LP formalism with some little differences
5. the lexicalization and the declarative properties are present in the two formalisms but with high degree in HPSG
6. equivalents of certain concepts of GPSG in HPSG:

GPSG	HPSG
Subcat feature	Valence feature or the three features (SUBJ, COMPS, SPR)
HFC principle	HFP principle
FCR	typed feature structures
FFP	Non local Features principle's

5.1.2. Differences between GPSG and HPSG

A. Case of phrase structure rules
Criteria (A.1) : character of generality and the number of used rules

GPSG	HPSG
GPSG makes use of a lot ID rules with the aim of giving an account of the different possible syntactic constructions. Among these rules, there are some lacking all kinds of generality, as it is illustrated in the following example. **Example:** VP→ V VP→ V NP VP→ V NP NP a. several VP rules for describing the VP phrase. b. the number of arguments depends directly on the type of the verb, It losses therefore generality.	Like GPSG, HPSG adopts the ID/LP formalism but with a slight difference in that it abandons completely the notion of phrase structure rule while generalizing the representation of the hierarchical information. Thus, HPSG proposes to use schema of rules that define the great range of phrase structure rules following their general structure: the type of the head, of complements, of the root, etc. However, HPSG makes use of a limited number of rules (ID Schemas) that have the characteristic to be general.

Criteria (A.2) : Prediction

GPSG	HPSG
GPSG fails to make predictions. In fact, each time, it comes to face a new problem; it proposes to introduce a new ID/LP rules.	In HPSG, we use a less number of rules (ID schemas) which have the characteristic of being general. Hence, the problem of prediction is almost solved.

B. Case of the lexicon

Criteria (B.1) : Representation of the lexicon

GPSG	HPSG
- Regarding the number of linguistic features and their varieties, GPSG feature structures, are less rich compared to those existing in HPSG. - Feature structures correspond to a non structured set of couples <attribute, value>.**Example:** Noun={<N,+>,<V, -, <bar,0>, <Plu,+>} - the notion of rule is not integrated in feature structures.	- HPSG lexical entries are richer than those of GPSG (in number of features and their level of adherence). - Feature structures are represented in the form of matrix, that permits the overlapping of structures and organize information in a better way. - feature structures are linguistic signs and the grammar declares a priori what signs are well formed. Therefore, the notion of rule is integrated to the feature structures.

Criteria (B.2) : Lexicalisation

HPSG is more lexicalised (strongly lexicalised grammar) than GPSG because of the presence of a great number of information (linguistic features) at the level of the lexicon concerning the upper levels (syntactic and semantic).

C. Case of linguistic sign representations

Criteria (C.1) : Sign representation uniformity

GPSG	HPSG
- We distinguish three sorts of linguistic signs: the lexical categories, the phrase and the sentence. - The syntactic rules as well as the syntactic and lexical structures are not represented in a uniform manner by feature structures. - The different analysis levels are represented differently and the Bar=0,1,2,3 feature (the level of analysis according to Uszkoreit) gives an account of the current level.	- We distinguish two sorts of signs: word and phrase. - There is a uniformity in the representation either the sign or the analysis level, like FUG, the syntactic rules as well as the syntactic and lexical structures are represented in an uniform manner by feature structures.

Criteria (C.2) : Typed feature structures

GPSG	HPSG
There is no typed feature or typed feature structures and there is no hierarchy of types.	The linguistic signs are typed and their types are organized in hierarchical way

Criteria (C.3) : Treatment of the subcategorization

GPSG	HPSG
The subcategorization concerns only the verbs, the subcat feature is used to indicate subcategorised units.	In HPSG, the subcategorization doesn't concern only the verbs but also nouns have elements of subcategorization such as determinants.

Criteria (C.4) : The linear order of the phrase constituents
The representation of the order of the constituent is encoded in GPSG by the LP-rules whereas in HPSG, the Phon feature together with the operation of concatenation ensures this role.

5.2. Comparison between GPSG and HPSG applied to the Arabic language

In this section, we will proceed by analysis level:

5.2.1. Lexicon

In the representation of the arabic lexicon, we have introduced some features.We list a few of them in the following table:

Feature	Meaning
CAD	Adnominal complement
TAN	Tanwin
Root	Root of the arabic word
Pattern	Pattern of the arabic word
Sifa	Charactriser or adjunct
Def	Boolean feature : Defined
Aff	Boolean feature: Affixed
Art	Boolean feature: Article
Exp	Exposant
Conv	Converter
Mode	Mode {accomplished, unaccomplished, imperativ}
Pr	Pronoun

Fig 8: list of introduced features

a. GPSG arabic lexicon representation
The lexical entries are represented by a feature structures. The introduction of a new feature within the feature structures is done according to a linguistic justification.
Exemple: A GPSG feature structures of a noun is of the form:

Noun={< N,+ >,< V, - >,< BAR, 0 >, < CAS, >, < PERS, >,< PLU, >,<GEN, >,
<TAN, >, <ART, >,<PR, >,<RACINE, >, <PATTERN, >, <CAD, >,<DEF, >}

$$\begin{cases} CAS \in \{nominatif, accusatif, génitif, datif\}, \\ PERS \in \{1,2,3\}, PLU \in \{+,-\} \text{ and } GEN \in \{Masculine , Feminine\}. \end{cases}$$

In order to ensure the exclusion of co-occurrence of root and pattern feature in the case of a pronoun we define and introduce these FCRs rules:

FCR: [+PR]⊃~[PATTERN]

FCR: [+PR]⊃~[ROOT]

b. HPSG Arabic lexicon representation

Example: Feature structures of an arabic noun كتاب

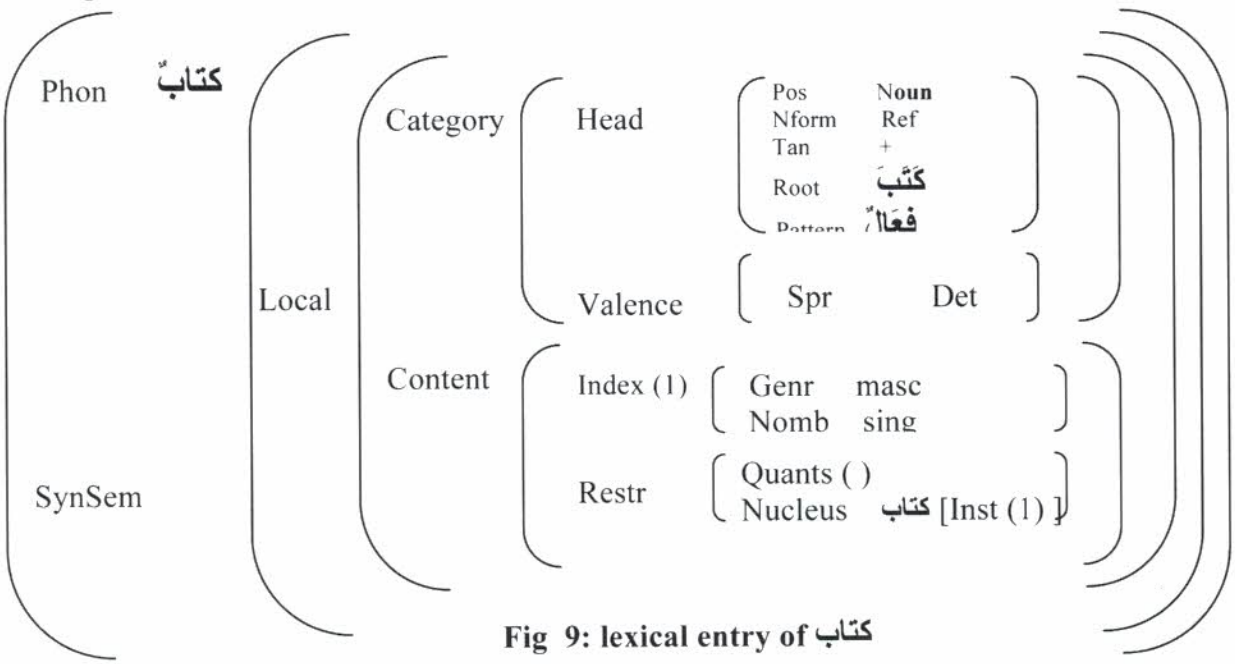

Fig 9: lexical entry of كتاب

Result 1: It is quiet clear that the HPSG feature structures are richer and more representative than those of GPSG because of the diversity of linguistic information stored in them. Also, they are organized under respect of the type hierarchy. That is to say that HPSG feature structures encapsulate more information concerning the different analysis' levels (morpho - lexical, syntactic, semantic and pragmatic levels) according to a pre-established type hierarchy. And in the light of what has been said, it would be useful to recommend the adoption of the HPSG lexicon representation.

5.2.2. Lexie

In the case of the lexie, the problematic comes from a central question: which of this two formalisms (GPSG and HPSG) is the one that ensures the best optimum linguistic cover of the lexie notion in the NKT. Let us consider the case of a simple nominal lexie, the core of which is a common noun.

a. Case of GPSG

The nominal lexie is specified by an ID/LP rules set with some feature co-occurrence restrictions FCRs.

ID rules: N2→N1, (A2) N2→N0[+PRO, -Aff], (A2) N1 → N0[-Cad], N1[Gen] N1→ N0[+Cad, -PRO], (A2) N1→ N2[+TAN] N1→ N2[+ART] N2[+ART]→Art,N0[-Cad], (A2)	***LP rules:*** a. [BAR 0] < [BAR 1] b. [BAR 1] < [BAR 2] c. Art< [BAR 0] ***FCRs rules:*** a. [+TAN] ⊃ [-DEF]

	b. [+PR] ⊃ [+DEF] c. [+ART] ⊃ [+DEF] d. ~[-CAD, +DEF] e. [+PRO, -AFF] ⊃ ~[GEN]

This description inherits inconveniences of the representation by such rules. Indeed, we must associate a set of ID, LP, FCRs, FSDs rules to each type of lexie. Thus, they don't have the character to be general.

b. Case of HPSG:

A maximal projection of the nominal lexie in HPSG needs the use of these following ID schemas of these following principles of well formation :

ID Schemas	Principles
H-C H-A H-PAFFC H-SP	Head Feature Principle HFP Valence principle Specifier principle Semantic principle

Example:

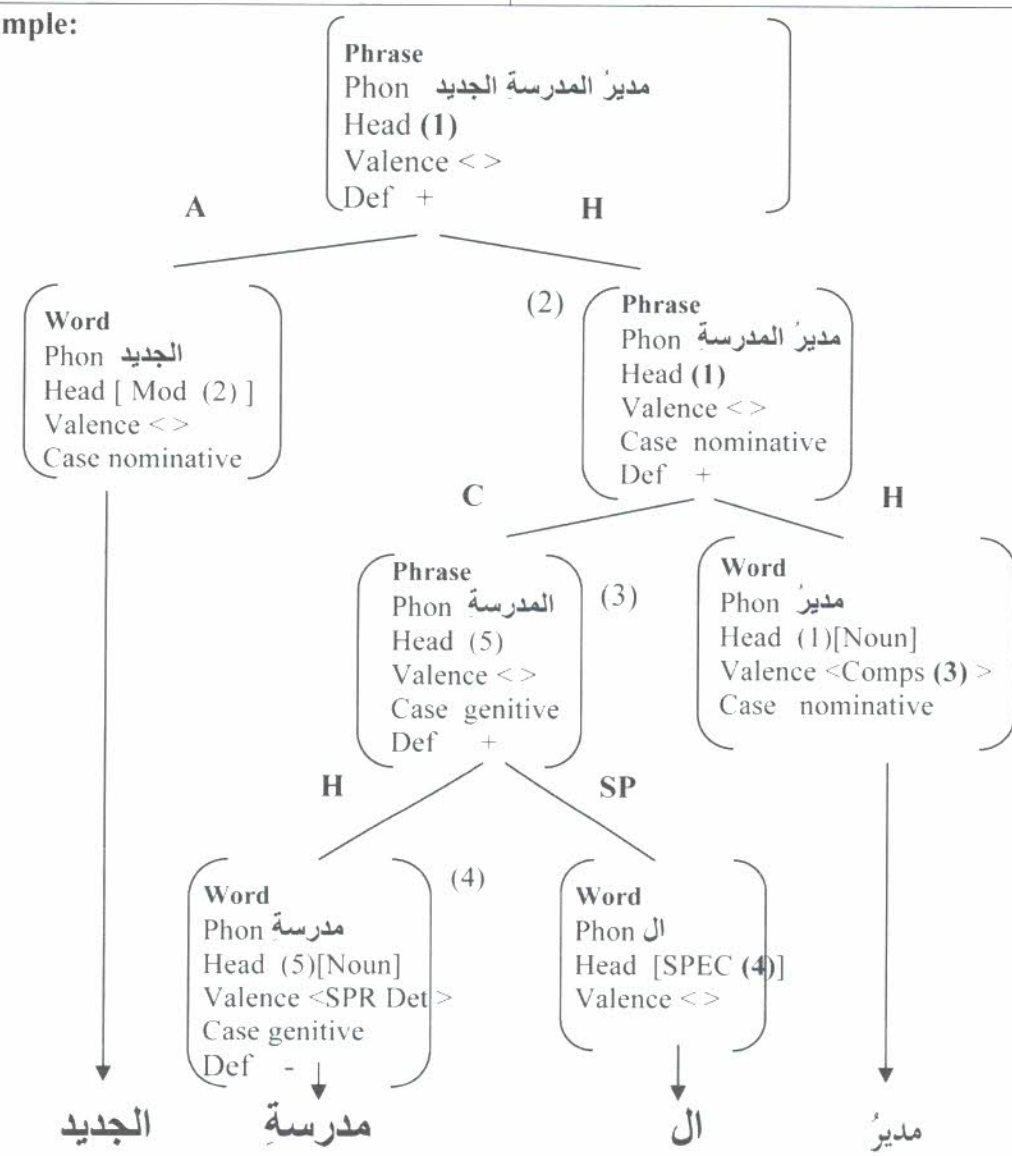

Fig 10: Représentation of nominal lexie مديرُ المدرسةِ الجديد

Result 2:

It was established according to this study that whatever the type of the nominal lexie to modelise, it is the same set of ID schema and principles that would be used. This fact represents a major advantage in relation to what suggests GPSG as far as the different types of nominal lexies are concerned. Hence, we got this partial result:

GPSG : different type of lexie ⟶ different ID rules
HPSG : différent type of lexie ⟶ same ID schémas

The number of used rules to ensure the description of the different types of lexie in the case of GPSG is more important than the one in HPSG representations, so we come to the conclusion that the HPSG representation is the most economic, the most compact and the most elegant.

5.2.3.Tectonie

The problematic in the case of tectonie comes out of the following question: which of this two formalisms (GPSG and HPSG),ensures a large optimum linguistic cover of the tectonie or of the arabic syntactic constructions according to the TNK?. Let us a consider the case of a nominal syntactic constructions with exponential verb as governor R=كان

a. case of GPSG:

V3→ V2[Subcat k], N2[+Def], A2[-Def, Acc]
V3→ V2[Subcat k], N2, P2
V3→ V2[Subcat k], N2, N2[+Adv]
V2<N2[Nominative]

b. Case of HPSG: There is an introduction of this feature: type-item ∈{R,T1,T2, D} and the application of the following ID schemas and principles:

ID Schemas	Principles
H-Comps H-Comps-Subj H-Subject H-Adjunct H-Specifier H-Modifier	Head feature principle HFP Valence principle Specifier principle Semantic principle

As a result, the representation of the syntactic construction is more concise in HPSG than in GPSG and we also note that:

GPSG : different constructions ⟶ different ID/LP rules
HPSG : différent constructions ⟶ same ID schémas

Result 3

As a result, we can say that the HPSG representations are the most adapted and the most economic to describe the different types of the arabic syntactic constructions according to the Neo-khalilien linguistic theory.

6. Conclusion

It has been clearly showed through this comparative survey that the HPSG grammars offer more advantages than the GPSG grammars in relation to several criteria. So it has been established that:

- HPSG is more declarative and more lexicalised than GPSG
- HPSG uses ID schemas instead of classical phrase structure rules that have the tendency to be not general.
- HPSG is more constrained. It operates a rigorous control on the syntactic construction formation by the principles of well formation and the ID schemas and also thanks to the typed features.

-HPSG formalism ensures the best linguistic cover of the arabic analysis levels according to the Neo-Khalilienne theory:

- *Lexical level: HPSG proposes a lexicon that is richer and more organized than the one suggested by GPSG.*
- *-Lexie/Tectonie level: HPSG proposes a more elegant and economical representation, more general.*

At the end, we can say that HPSG is the best adapted to a computer implementation because of these attractive characteristics and because of its large linguistic cover. The HPSG grammars are more advisable and more recommended than GPSG in the representation of Arabic analysis levels.

7. References

[Ann, 2000] Anna Kupsc,"An HPSG Grammar of Polish clitics" Doctoral Dissertation, University Paris 7, 2000.

[ACH, 2003] ACHIT Abdelmadjid, «Comparaison des grammaires syntagmatiques généralisées GPSG avec les grammaires syntagmatiques guidées par les têtes HPSG » Troisièmes journées scientifiques des jeunes chercheurs en génie électrique et informatique GEI 2003, MAHDIA, du 16 au 19 Mars 2003, Tunisie.

[BLA, 95] Philippe Blache, "Une introduction à HPSG", 2LC-CNRS, 1995.

[CAR, 99] Andrew Carnie, "Alternatives theories: HPSG " Department of Linguistics University of Arizona, 1999.

[GAZ, 85] GAZDAR Gérald, "Genaralized Phrase Structure Grammar", Oxford University, 1985.

[HADJ, 79] HADJ SALAH Abderrahman "Linguistique arabe et linguistique générale, Essai de méthodologie et d'épisthomologie du ilm al-Arabiyya". 02 volumes, 1979.

[P&S, 94] POLLARD Carl & Ivan SAG, "Head Driven Phrase Structure Grammar" University of Chicago Press and Stanford: CSLI Publications, 1994.

[TEND, 99] Hafida TENDJAOUI, « vers une compréhension automatique de l'arabe basée sur le modèle GPSG et le modèle linguistique Néo-khalilien » mémoire de Magister, CRSTDLA- ENS d'Alger, 1999.

[VOL, 88] Martin Volk:"Parsing German with GPSG: The problem of separable-prefix verbs". ACMC Research Report 01-0026. The University of Georgia. Athens, 1988.