Conception d'un dictionnaire terminologique et phraséologique trilingue anglais/français-arabe dans le domaine de l'optique

Xavier LELUBRE

Université Lumière Lyon-2, France

• *Abstract* •

This is a project of a dictionary aimed at translators, essentially from French or English into Arabic language. This is why it associates to each concept its definition, the set of terms relative to it in each precited language On the other hand this dictionary will offer for each term a cluster of phraseologisms relative to it

It will also take under consideration the problems of synonymy, with a special consideration for Arabic optic terms

Introduction

Le point examiné ici est relatif aux choix et aux contraintes qui président à l'établissement d'un dictionnaire spécialisé trilingue français-anglais-arabe dans le domaine des interférences lumineuses, sous-domaine de l'optique.

Un travail de ce type doit prendre en considération le concept de *langue de spécialité*. Du point de vue qui nous intéresse, quatre composantes contribuent à l'établissement du texte de spécialité :

i. la composante terminologique : la terminologie de la spécialité concernée, c'est-à-dire les dénominations des unités référentielles constituant le domaine traité ;

ii. la composante phraséologique : les phraséologies relatives à cette terminologie, c'est-à-dire les contextes préférentiels d'occurrence des termes dans les textes de spécialité ;

iii. la composante brachygraphique : les signes brachygraphiques ;

iv. la composante discursive : les expressions propres aux types de discours intervenant dans les textes de la spécialité concernée.

Il convient aussi de prendre en compte d'une part les différents lieux des textes de spécialité (en physique, le texte de spécialité est susceptible de se dérouler selon différents moments tels que le corps du texte, y compris les notes de bas de page, le texte des figures (légendes sous la figure ; textes dans la figure) et illustrations, le texte des exercices, le texte des résumés, le texte des titres, les « textes brachygraphiques » (les formules), le texte de notices historiques, etc.) et d'autre part les différents niveaux de spécialisation (publications destinées aux spécialistes, manuels destinés aux apprenants, ouvrages de vulgarisation).

Un texte de spécialité n'est généralement pas homogène ; des discours d'autres spécialités viennent y contribuer (par ex., discours mathématique, notice historique dans un texte d'optique).

Quelle est l'incidence de ces différents aspects sur l'entreprise dictionnairique, c'est-à-dire quels sont donc les éléments de la langue de spécialité qui devront être représentés dans le dictionnaire, et comment vont-ils être représentés ? Elle dépend de toute évidence de la finalité qu'on lui assigne. Il s'agit ici d'un dictionnaire trilingue, et il est – tout naturellement – conçu comme un outil pour la traduction[1].

À titre d'exemple, voici un texte relatif aux interférences lumineuses[2] :

> « C'est dans la région M1 M2 où se superposent les faisceaux diffractés que l'on peut observer les franges d'interférences. En un point quelconque M de l'écran P, la différence de marche δ est (fig 2 b)
> (8) $\delta = [..]$;
> en considérant les triangles A1A2H et COM, et en posant [..], on a :
> (9) $\delta \# [..]$
> [..]
> En première approximation, les franges sont des droites parallèles et équidistantes. Elles sont dirigées perpendiculairement au plan de la figure 2 b. La formule (10) montre que les maximums de lumière, c'est-à-dire les franges brillantes, sont donnés par
> (11) $\delta / \lambda = [..]$.»

Pour traiter un texte de ce type, le traducteur a besoin des éléments suivants :

– **termes** (*faisceaux diffractés, franges d'interférences, franges, écran, différence de marche, triangle, droites parallèles, droites équidistantes, plan, maximum, lumière, franges brillantes*) ;

1 Remarquons au passage que le sens le plus fréquent de la traduction va du français ou de l'anglais vers l'arabe, ces deux langues étant aujourd'hui, comme on le sait, les deux principales langues – et de loin – de référence en ce qui concerne les domaines scientifiques, dans le Monde arabe. C'est la raison d'ailleurs pour laquelle un tel ouvrage se doit d'être au moins trilingue, avec le français et l'anglais comme « langues de référence » : ne tenir compte que de l'une de ces deux langues, à l'exclusion de l'autre, constituerait un vice rédhibitoire (ne serait-ce, nous y reviendrons, que parce que les terminologies scientifiques arabes contemporaines se constituent la plupart du temps sur la base des terminologies françaises et anglaises).

2 Maurice Françon, « interférences lumineuses », in *Encyclopaedia Universalis* (vol 9, 1968 : 2). C'est nous qui soulignons.

– connaissance de la façon dont ces termes fonctionnent en contexte, c'est-à-dire les **phraséologismes** au sein desquels ils interviennent (« se superposent les faisceaux diffractés », « observer les franges d'interférences », ...) ;

– signes **brachygraphiques** (d pour la différence de marche, l pour la longueur d'onde, M1M2, A1A1H, ...) ;

– **expressions des discours** organisant le texte (discours mathématique : « en un point quelconque de... », « en considérant les triangles », « en posant », « on a : », « la formule... montre que », « les maximums [..] sont donnés par : ». Physique : « c'est dans la région... que l'on peut observer... », « en première approximation ». Didactique/ terminologique : « les maximums de lumière, c'est-à-dire les franges brillantes ».

1. L'établissement du dictionnaire

L'étude terminologique préalable a porté sur la délimitation du domaine, la constitution d'un corpus, par dépouillement des textes en français, en anglais et surtout en arabe.

Nous avons pris comme base scientifique des textes, français et anglais, de manuels universitaires. Quant aux termes arabes, nous les avons pris essentiellement de manuels arabes qui ont été à notre disposition, tout en tenant compte des lexiques – lexiques bilingues ou trilingues –, sur lesquels bien souvent l'on peut faire de nombreuses réserves des points de vue lexicographique, terminologique et scientifique.

1.1. Organisation générale

Nous avons établi l'arbre (un arbre possible) du domaine et adopté à partir de cet arbre une classification systématique[3]. Chaque unité référentielle est repérée par un indice. Tout terme est repéré par un indice, qui est celui de l'unité référentielle qu'il dénomme, complétée par un code relatif à la langue :

xxFA5/F	frange centrale
xxFA5/E	central fringe
xxFA5/A11	/hudbat markazijjat/
xxFA5/A12	/hudb markazijj/

Les grandes divisions du sous-domaine des interférences lumineuses que nous avons dégagées sont les suivantes :

3. Il s'agit d'un système mixte de classification, hiérarchique et à facettes. Il utilise un ensemble de caractères alphanumériques (Lelubre, 1992 : 141-142).

Division A	(types d'interférences)
Division D	(nature ondulatoire de l'interférence)
Division F	(franges)
Division H	(éléments physiques qui interviennent)
Division K	(interférence en lumière polarisée)
Division U	(dispositifs d'obtention)
Division V	(interféromètres)
Division W	(pièces optiques utilisées pour les dispositifs interférentiels)
Division X	(applications)

1.2. Appareil lexicographique

Il est classique. En particulier, il convient de fournir à l'utilisateur les éléments non prédictibles, comme, pour l'indication du genre pour les substantifs français, de la forme au pluriel pour les substantifs arabes, etc.

Quant aux définitions, chacun des termes qui s'y trouve doit être à son tour une vedette du lexique[4].

2. Traitement des faits de variation

2.1. Prise en compte des phénomènes de synonymie

Les termes synonymes, pour une langue donnée, sont répertoriés par un indice placé après le code de la langue. Les faits de variations sont de nature et de degré variable (Lelubre, 1992 : 381-416) ; cela peut aller de la simple variante orthographique jusqu'à des termes exprimant pour la même unité référentielle des traits de substance différents. Nous en avons tenu compte de manière simplifiée, traitant de manière différente les seconds :

– indice à deux chiffres dans le cas de synonymie proche
xxVA3 /F11 interféromètre de Fabry-Pérot
xxVA3/F12 interféromètre de Pérot et Fabry

– variation de l'indice en premier rang pour une synonymie plus éloignée
xxFD9 /F1 facteur de visibilité /E1 visibility factor
xxFD9/F2 efficacité lumineuse /E2 luminous efficiency

Par ailleurs, il convient de noter que des termes ambigus ont des indices différents, apparaissant à des endroits différents de l'arbre du domaine.

4. En fait, si le domaine de spécialité traité par un dictionnaire est étroit, nombre de ces termes ne peuvent figurer comme vedette

2.2. Indication des « aires d'emploi » des termes

La question de la variation en ce qui concerne les termes arabes est particulièrement importante. L'un des problèmes épineux de la terminologie scientifique arabe est dû à ce fait de dispersion, dont les causes sont multiples mais dont deux sont à signaler : la multiplicité des foyers de création terminologique, joint au fait qu'aucun d'eux ne domine véritablement les autres, et l'existence de deux langues fournisseuses de terminologie, le français et l'anglais, dont l'on peut dire que chacune est pratiquée à l'exclusion de l'autre dans telle ou telle aire du Monde arabe.

De fait, on peut dire qu'il existe quelques grandes tendances : la terminologie majoritairement utilisée en Syrie – c'est le seul pays où l'enseignement des matières scientifiques se fait en arabe, y compris dans le supérieur – dont bon nombre d'enseignants ont travaillé dans d'autres pays arabes, la terminologie utilisée en Egypte – dont on connaît le poids dans le Monde arabe – et dont de nombreux enseignants ont eux aussi travaillé dans les autres pays arabes. De plus il existe aussi des organismes inter-arabes, comme l'Union des Physiciens, et le Bureau de Coordination pour l'arabisation, qui a organisé des congrès d'unification terminologique, qui produisent des listes de plusieurs centaines de termes « unifiés » en cette occasion, et publient des lexiques spécialisés – de valeur lexicographique et terminographique souvent discutable – c'est le cas pour le domaine de la physique. Une étude serait à mener sur l'influence réelle de ces efforts normatifs[5].

Un dictionnaire qui ne se veut pas normatif se doit de fournir à son utilisateur les variantes qui existent, pour chacune des trois langues, et de lui indiquer aussi les aires d'emploi de chacune d'entre elles. Cela est nécessaire quand on traduit vers l'arabe, car il va de soi que la terminologie adoptée doit être cohérente de ce point de vue.

Nous avons codé pour de nombreux termes arabes leur aire d'emploi (C pour les termes communs à tous les pays, qui ne posent de ce point de vue aucun problème, U pour les termes « unifiés », avec éventuellement des indices en cas de variantes officiellement admises, R pour des termes répandus, pouvant coexister avec des variantes, S pour les termes plus spécifiquement « syriens », E pour les termes « égyptiens »...). On peut de toutes façons indiquer aussi, pour chaque terme arabe, les références des textes où ils ont été trouvés (nous avons codé ces ouvrages en fonction des critères suivants : pays d'origine, type d'ouvrage, niveau, domaine et sous-domaine, et enfin, un numéro d'ordre ; par exemple : SUPO2 Syrie Universitaire Physique Optique 2. Il n'y figure pas d'indication d'ordre diachronique).

3. Traitement des phraséologismes

3.1. Caractérisation et modélisation des phraséologismes

On entend ici par *phraséologisme*, tout syntagme non terminologisé, comprenant au

5 La question de la *norme* se pose de manière différente selon les domaines. Pour ce qui est de la terminologie française de la physique, c'est essentiellement au sein du monde des spécialistes – enseignement supérieur, recherche, revues scientifiques – que la norme s'établit. Notons en particulier le rôle joué par les revues scientifiques de haut niveau, qui donnent le ton, non seulement en matière terminologique, mais encore qui imposent un style de rédaction donné. Rien de tel en arabe, même s'il existe des revues scientifiques arabes, mais qui sont des revues de vulgarisation, fussent-elles de bon niveau. Les chercheurs publient davantage en anglais ou en français.

moins un terme, d'étendue supérieure au terme et inférieure à la phrase, et par ailleurs de « fréquence non négligeable » dans les textes de spécialité.

Il s'agit, on le voit, d'une définition empirique, exprimant la notion intuitive de phraséologisme comme entourage préférentiel d'un ou de plusieurs termes. De fait, tout syntagme du type évoqué dans cette définition a vocation à constituer un phraséologisme. Quant au critère de fréquence, il relève d'un tout autre plan.

Nous travaillons actuellement sur la question des phraséologismes tels qu'ils se présentent dans les textes de physique, et nous sommes loin d'en avoir établi une liste qui ne pourrait être véritablement significative que si elle s'appuyait sur un dépouillement quantitativement important de textes de la spécialité étudiée. Si l'établissement de la terminologie d'un domaine relève d'une démarche fondamentalement onomasiologique, l'établissement de la phraséologie correspondante relève d'une démarche sémasiologique, pour laquelle le texte est le matériau de départ et les termes sont les signaux d'existence.

Nous donnons ci-dessous quelques phraséologismes – en français – relatifs au terme *frange (d'interférence) /hudbat/* (pl. */hudb/*) :

. nous obtenons des franges très étroites
. observe des franges d'interférence à la *surface* de la lame
. franges d'interférence produites par une *double source*
. nous pouvons mesurer l'*interfrange*
. on appelle *interfrange* la *distance* qui sépare deux franges voisines

L'on peut considérer les phraséologismes sous deux plans :

– le plan de la forme, le plan syntaxique : le phraséologisme est composé d'une *relation* mettant en rapport des *arguments*, que l'on peut représenter sous la forme d'un *vecteur phraséologique* :

. nous obtenons des franges très étroites
<obtenir 1 2 | nous; franges très étroites>
<obtenir 1 2 | [humain]; franges >

. franges étroites
<étroit 1 | franges>

Les phraséologismes sont susceptibles de s'enchâsser :

<obtenir 1 2 | nous; franges très étroites>
<obtenir 1 2 | nous; <étroit 1 | franges> >

La représentation phraséologique ci-dessus est susceptible de représenter des phraséologismes comme :

. on obtient des franges étroites
. l'obtention de franges étroites
. les franges étroites obtenues

. (pour) l'obtention de franges étroites
. des franges étroites s'obtiennent
. (on peut) obtenir des franges étroites

. des franges étroites
. des franges qui sont étroites
. ces franges sont étroites
. l'étroitesse des franges

Ces phraséologismes ont des structures syntaxiques de surface qui sont différentes, mais dont la représentation en termes de relation et arguments est la même[6].

À l'inverse la possibilité d'avoir diverses réalisations de la représentation phraséologique – en contexte, faisant donc intervenir d'autres lexies et d'autres contraintes syntaxiques – pourront être différentes selon la langue (ainsi en arabe, on passe facilement de la forme verbale conjuguée à l'« infinitif » correspondant, pouvant être utilisé de manière beaucoup plus souple qu'en français).

.* notre obtention de franges étroites
./HuSu:l-u-na: ' ala: hudb-in Dajjiqat-in/

– le plan du fond, d'ordre sémantique : le phraséologisme exprime un type de relation référentielle qu'une unité référentielle d'un certain type entretient avec d'autres, de nature généralement différente, et, si l'on peut dire, avec le reste de l'univers (physiciens compris !).

Argument 1	**Relation**	Argument 2	
nous	**obtenons**	des *franges*	très étroites
[humain] expérimentateur	réussir à atteindre (un résultat (*Petit Robert*)	phénomène physique observable, s'étendant dans l'espace de manière discontinue et dénombrable	qualification relative à une entité à deux dimensions

6 Le terme *frange* est ici employé au pluriel. L'ensemble des phraséologismes relatifs à un terme n'est pas forcément le même selon que le terme est employé au singulier ou au pluriel (quand le pluriel est possible), quand le terme est déterminé par l'article ou pas, etc. Un exemple, *source cohérente / source_s cohérente_s* (ces dénominations sont en fait métonymiques : ce sont les *rayonnements* – vus comme grandeurs sinusoïdales –, émis par ces sources qui sont cohérentes, c'est-à-dire qui ont des relations de phase constantes au cours du temps (Mathieu *et al.*, 1985 : 81)). Une source cohérente est telle que tous les rayons émis par cette source sont cohérents. Deux ou plusieurs sources cohérentes (cohérentes entre elles) sont telles que les rayons émis par elles sont cohérents. En définitive, si la cohérence concerne les rayons lumineux, dans l'emploi le plus fréquent, deux *sources cohérentes* sont cohérentes entre elles mais ne sont pas par elles-mêmes chacune une *source cohérente*.

		Argument 1	**Relation**
		des *franges*	très **étroites**
		phénomène physique observable, s'étendant dans l'espace de manière discontinue et dénombrable	qualification relative à une entité à deux dimensions

3.2. Traitement lexicographique des phraséologismes

Sur le plan lexicographique, la question se pose de savoir quels sont les phraséologismes qui doivent être insérés dans le dictionnaire spécialisé.

Par exemple, si nous considérons les deux termes suivants, dont l'un est générique et l'autre spécifique, l'ensemble des phraséologismes repérés pour chacun d'eux, *frange* et *frange centrale* – la *frange centrale* étant une *frange* particulière dans un *système de franges* – est-il le même ? Certainement pas : si les (ou des) *franges* et **la** *frange centrale* se déplacent, s'observent, sont situées sur une surface, par contre, on dénombre, on compte des *franges*, mais pas la *frange centrale*...

Trois critères semblent pouvoir être pris en considération pour le classement dans l'article du dictionnaire relatif à un terme, les phraséologismes qui lui sont associés :

– d'ordre formel : nombre d'arguments, type de constructions syntaxiques.

– d'ordre sémantique : les types de relation pouvant exister entre des types d'unités référentielles de nature donnée. On peut envisager une grille du type :

```
production/fabrication – lieu/déplacement – temps/évolution
- ...
```

– d'ordre statistique : fréquence des occurrences.

Si le critère statistique est pertinent pour le choix des phraséologismes retenus, l'ordre sémantique paraît s'imposer pour l'ordre de présentation des phraséologismes, du fait que le dictionnaire est trilingue.

Conclusion

Un dictionnaire comme celui esquissé ici concerne un domaine très étroit. Si l'utilisateur est préoccupé avant tout de choix de termes, de questions terminologiques, il doit trouver, dans un ouvrage de ce genre, les termes dont il a besoin, avec en particulier leurs variantes, avec leur aire d'emploi.

Mais l'aide au rédacteur ou au traducteur que doit apporter un dictionnaire termino-phraséologique plurilingue de ce genre est en relation avec les types de discours de spécialité qui viennent interférer dans les textes de la spécialité concernée. Or de ce point de vue l'intérêt d'un lexique qui ne traite que d'un domaine très étroit est très limité, du fait justement qu'il ne traite pas de la variété de ces discours de spécialité qui concourent, chacun avec sa terminologie et sa phraséologie au texte de la spécialité.

Ce travail ne peut être conçu que comme la partie d'un tout bien plus vaste, qui serait un dictionnaire termino-phraséologique de la physique.

Si la détermination des termes d'une langue relatifs à un domaine s'appuie tout naturellement sur la connaissance des unités référentielles qui le constituent, il n'en est pas de même pour les phraséologismes qui, eux, ne peuvent être mis au jour qu'à partir des textes de spécialité et donc de corpus significatifs quantitativement. Cette exigence implique le recours à une procédure de traitement automatique de ces textes, pour repérer les occurrences et les contextes de chaque terme, ce qui est nécessaire pour donner à l'expression « fréquence non négligeable » d'occurrence de phraséologismes évoquée plus haut une signification pertinente[7].

C'est à cette condition que pourra être achevé un véritable dictionnaire de spécialité relatif au domaine de l'optique.

Annexe 1

phraséologisme	Source	Argument 3	Argument 2	Argument 1	Relation
naHsulu`ala: hudb-in Dajjiqat-in zudd-an	EUPO1:365		hudb Dajjiqat zidd-an	/naHnu/	HaSala 1 `ala · 2
nous obtenons des franges très étroites			**franges très étroites**	**/nous/**	**obtenir 1 2**

Exemple de relevé de phraséologisme avec formalisation sous forme de vecteur phraséologique

7 Cela implique que les termes du domaine soient au préalable stockés dans une base de données informatique. Nous disposons de procédures de traitement automatique – essentiellement au niveau morphologique – de textes arabes non voyellés (voir, par ex. Dichy et *al*., 1989). Il reste à établir les procédures permettant la reconnaissance d'unités terminologiques complexes.

xxF	*Division F franges*								
xxF0	MKF 200 franges d'interférences «On nomme *système de franges d'interférences* une famille de courbes tracées sur une surface en tous points de laquelle se produisent des phénomènes d'interférence. Sur chaque courbe ou *frange*, la *différence de marche* entre les vibrations interférentes est constante et on peut caractériser chaque frange par un nombre. L'*ordre d'interférence* $\eta = \delta/\lambda$, λ est la longueur d'onde des ondes qui interfèrent. [illegible] «Le *facteur de contraste* (facteur de visibilité en C) qui a pour expression $V = (IM-Im)/(IM+Im)$» V Pitt 161 *franges* «Circular or rectilinear patterns of alternate light and dark or of colour produced by *interference* or *diffraction* of light»	xxF0/F	frange	xxF0/E	fringe	xxF0/A11	hudb /pl ʔahda b	S U	IUPO1 27 MLS1 MLS2 MSP3 195 SSP2 158 SUPO2 lex SUPO6 252 UL1opt
xxF0		xxF0/F	frange	xxF0/E	fringe	xxF0/A12	hudbat /pl hub	E	EUP1 630 EUPO1 364 LUP2 263 SUPO5 lex
xxF0		xxF0/F	frange	xxF0/E	fringe	xxF0/A2	$ari T/pl ʔa$riTat	¤	RUP2 119
xxF01		xxF01/F	frange_s d interférence	xxF01/E	interference fringe_s	xxF01/A11	ʔahda b (ʔa)t-tada.xul (pl)	S U	IUPO1 24 MSP3 197 SSP2 157 SaUPO1a 288
xxF01		xxF01/F	frange_s d interférence	xxF01/E	interference fringe_s	xxF01/A12	hudb (ʔa)t-tada xul	E	ELPN1 EUP1 629 LUP2 256 UL1opt ULP1b
xxF01		xxF01/F	frange_s d interférence	xxF01/E	interference fringe_s	xxF01/A13	ʔa$riTat (ʔa)t-tad xul (pl)	¤	RUP2 119

Génération automatique de néologismes arabes à partir des règles de formation de mots

Hussein HABAILI et Mohamed BEN AHMED

Laboratoire de Recherche en Informatique Arabisée et Documentique Intégrée, (RIADI), Tunis, Tunisie

• *Abstract* •

This paper enters within the theoretical framework of lexicalist hypothesis which preconizes an autonomous morphological component based on the lexicon, responsible of flexion, derivation and composition.

This research allowed to specify with details the nature of lexical rules permitting the derivation of words departing from basic words. Such relationship is achieved according to a set of word formation rules. These rules are characterized by their logical and mathemathical formalism. They can be computerized and naturally applied to Arabic morphological processes such as flexion, derivation and composition.

Indeed, their application in Arabic morphology is of a great importance for lexical neology, mainly while creating new words in technical and scientific domains.

A generator is designed to generate automatically new words for different grammatical categories (substantives, adjectives, action names, etc.) departing from a set of word formation rules set up for the standard Arabic.

Through a freak of word derivation and composition departing from basic words (by prefixation, suffixation and composition), the generator thus allows to update and enrich the Arabic dictionary with neologisms.

The generator is linked to a dictionary including basic words, roots, prefixes and suffixes with grammatical information associated with words and a set of word formation rules.

1. Introduction

L'hypothèse lexicaliste de Chomsky (1970) n'est pas l'unique proposition dans le cadre de la grammaire générative qui préconise la séparation des **règles de formation**

de mots (**RFM**) de la syntaxe. En effet, Kiefer (1973)[1] a proposé une composante morphologique autonome responsable de **la flexion**, de **la dérivation** et de **la composition** comme le montre le schéma suivant :

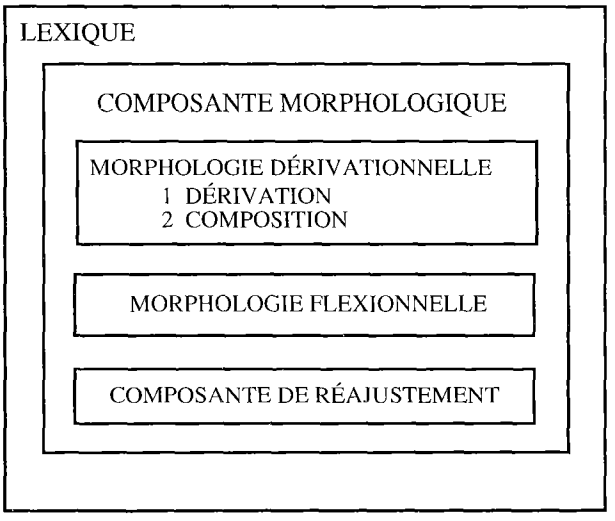


FIGURE 1 : La composante morphologique

Selon Kiefer la composante morphologique s'applique exactement après les transformations syntaxiques et avant la phonologie. Cet ordre est motivé, en un sens qu'il permet d'introduire les informations transformationnelles aux règles morphologiques.

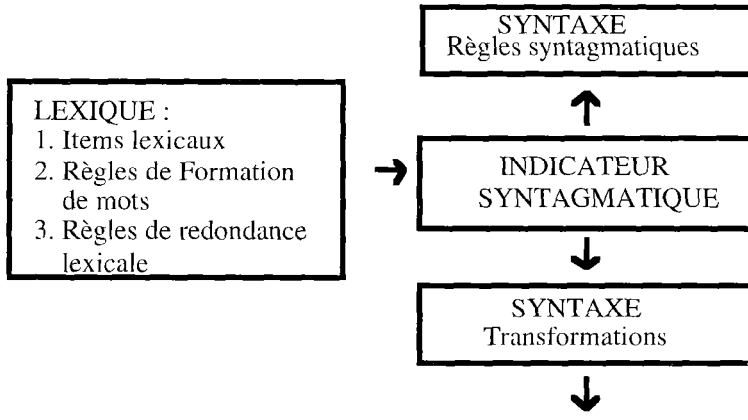
Ce point de vue est développé dans des études subséquentes dans le même cadre théorique de l'hypothèse lexicaliste originelle par certains linguistes (Siegel, 1974 ; Aronoff, 1976 ; Allen, 1978 ; Strauss, 1979, 1982 ; Habaili, 1990).

En outre, ces études ont permis de spécifier, d'une façon détaillée, la nature des règles lexicales qui permettent de dériver des mots à partir des mots de base. Cette relation s'accomplit en partie par un ensemble de règles de formation de mots : (WFR, cf. spécialement Aronoff, 1976), qui permettent d'attacher un affixe à la base.

Malgré les différences de détails de ces hypothèses toutes les études citées s'accordent sur un principe fondamental du lexicalisme selon lequel **la morphologie dérivationnelle** est construite au niveau du lexique.

L'organisation de la grammaire qui rendra compte de ce principe sera la suivante :

1 F. Kiefer, « Morphology in Generative Grammar », Gross et al. (Eds), *The Formal Analysis of Natural Languages*, The Hague, Mouton, 1973



Pour répondre à la question qui consiste à savoir si les règles de formation de mots s'appliquent dans le lexique, nous passons en revue les différentes positions de certains linguistes.

En 1965, N. Chomsky définit le lexique ainsi :

(A) lexicon... is simply an unordered list of all lexical formatives. More precisely, the lexicon is a set of lexical entries, each lexical entry being a pair (D,C), where D is a phonological distinctive feature matrix "spelling" a certain lexical formative and C is a collection of specified syntactic features (a complex symbol).

Dans son hypothèse prélexicaliste Chomsky considère les formants individuels comme provenant du lexique et insérés dans la syntaxe... Ces propriétés des formants lexicaux qui ne sont pas prévisibles sont considérées comme essentiellement idiosyncrasiques et doivent être associées aux formants dans le lexique.

Halle (1973) a observé que les mots flexionnels peuvent avoir une certaine idiosyncrasie. En effet, certaines formes flexionnelles ont une signification idiosyncrasique, d'autres sont des exceptions à des règles phonologiques régulières et d'autres, sont accidentellement absents des paradigmes. Pour Halle, de telles formes doivent entrer dans le lexique.

Allen (1978) a examiné des mots composés. Elle les considère comme ayant une signification idiosyncrasique. Elle arrive à la conclusion que ces mots composés qui ne sont pas sémantiquement prévisibles sont lexicalisés et leur sens n'est pas dérivable, mais sont listés simplement dans le lexique.

Ce faisant, la « **condition d'idiosyncrasie** » pour entrer dans le lexique dépasse les formants lexicaux pour inclure d'autres éléments qui peuvent figurer dans le lexique :

LEXIQUE

1. RACINES ET AFFIXES
2. MOTS DÉRIVÉS
3. MOTS FLEXIONNELS
4. MOTS COMPOSÉS
5. PHRASES SYNTAXIQUES

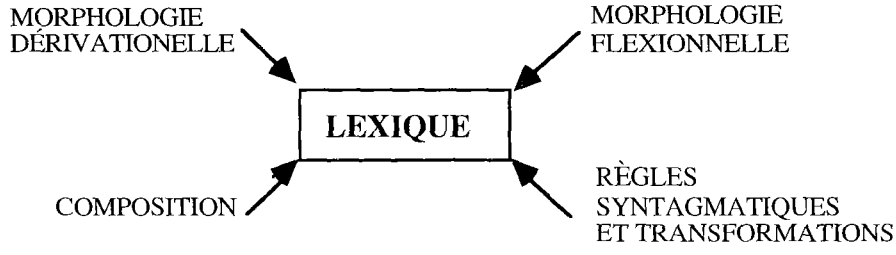
Ceci nous ramène à nous poser de nouveau la question qui consiste à déterminer la place des différents types de règles génératives dans la grammaire, en partant de la formulation originelle de l'hypothèse lexicaliste. En effet, nous pouvons nous demander pour quelle raison nous pouvons affirmer que les règles qui engendrent des entrées lexicales (mots dérivés...) sont considérées comme partie intégrante du lexique ; tandis que les règles relevant de la morphologie flexionnelle, de la composition et de la syntaxe ne le sont pas ?

Si les règles de la **morphologie dérivationnelle** sont situées dans le **lexique** pour la simple raison qu'elles sont capables d'engendrer des mots dérivés, la même raison est aussi valable pour les règles qui peuvent engendrer des entrées lexicales et pour cela ces dernières doivent figurer aussi dans le lexique.

Halle (1973) a raisonné en ce sens : il est arrivé à la conclusion selon laquelle les règles de la morphologie flexionnelle sont situées à la base du lexique, étant donné qu'il existe des mots flexionnels situés dans le lexique en raison de leur idiosyncrasie. Cependant, faut-il placer aussi les règles syntaxiques dans le lexique puisqu'il existe aussi des phrases syntaxiques idiosyncrasiques (les idiomes) ?

Si nous considérons toutes les règles génératives (structurales) de la grammaire comme faisant partie du lexique nous détruisons pour cela la séparation des règles de formation de mots de la syntaxe, sans oublier que l'essence même de l'hypothèse lexicaliste déjà atteinte consistait à déplacer les RFM de la composante syntaxique au lexique.

Pour résoudre ce problème, il faut considérer les règles de la morphologie dérivationnelle aussi régulières (sémantiquement et phonologiquement) que les autres règles génératives. Ce faisant, nous pouvons donc penser que chaque classe de règles peut engendrer des mots ou expressions qui peuvent devenir des items lexicaux :



En outre, pour marquer d'une façon précise la séparation de la morphologie de la syntaxe, il faut distinguer entre deux types de sous-grammaire connus sous les noms de *WORD-GRAMMAR* et de *SENTENCE-GRAMMAR*.

En effet, la première comprend la morphologie dérivationnelle, la morphologie flexionnelle et la composition et ne peut engendrer que des catégories lexicales (N, V, Adj...) ; tandis que la seconde qui comprend les structures de phrases et les transformations constitue un champ d'application des règles phonologiques.

2. Principes théoriques pour l'élaboration des règles de formation de mots

2.1. Les règles de formation de mots

La phonologie lexicaliste a pour objet l'établissement d'un principe de base pour l'explication morphologique en phonologie. Elle doit aussi déterminer le rôle des traits morphologiques et des structures de mots dans la forme et l'application des règles phonologiques.

La phonologie lexicaliste s'inscrit dans le cadre de l'hypothèse lexicaliste (Chomsky, 1970). Cependant, elle est à la recherche d'une morphologie autonome dont les structures de mots sont justifiées à la base à partir de la distinction des morphèmes.

La phonologie lexicaliste a donc un double objectif. Le premier consiste à élaborer et à justifier une théorie autonome et interne de formation des mots. Le second cherche à analyser et à expliquer comment les structures de mots engendrées par la théorie de formation de mots influencent la forme et l'application des règles phonologiques.

Considérons la représentation de la relation qui existe entre une suite terminale **t** et une suite non-terminale **A** :

(1)
A
|
t

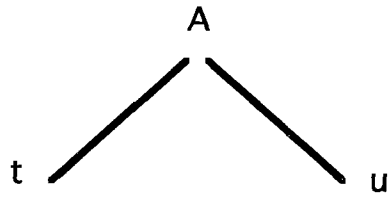
À partir de la représentation (1), nous pouvons formellement exprimer le fait que /kataba/ est un verbe, et que /baabun/ est un nom :

(2)

V	**N**
|	|
/kataba/	/baabun/

En supposant qu'il existe une possibilité logique qui consiste à ajouter une suite terminale **u** à droite de **t** dans (1), ceci nous donne les trois possibilités suivantes :

(3)

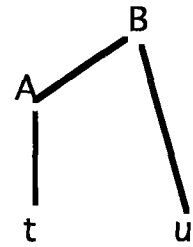


Dans cette représentation, la suite terminale **u** s'adjoint à **t** sans suite non terminale additionnelle. C'est ainsi que la structure de la représentation (3) est équivalente à **[t + u]A**, et la dérivation est :

[t]A → [t + u]A.

La deuxième possibilité est celle qui adjoint **u** à **[t]A** avec une construction simultanée d'un nouveau nœud **B** de la façon suivante :

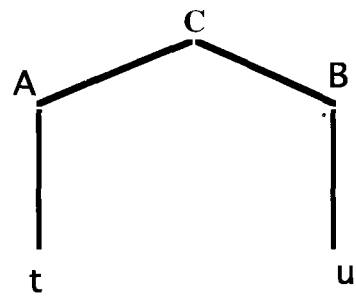
(4)



La dérivation de (4) est donc : **[t]A → [[t]A + u]B**. Si l'on compare les représentations (3) et (4) et leur dérivation, l'on constate que la dérivation de (4) diffère de celle de (3), étant donné que deux suites non terminales résultent d'un seul input initial.

La troisième possibilité est celle qui adjoint **u** à une suite non terminale qui possède à elle seule :

(5)



La dérivation de (5) devient alors :

[t]A → [[t]A + [u] B]C.

Grâce à cette distinction binaire entre le vocabulaire terminal et non terminal, les représentations (3), (4) et (5) constituent les seules possibilités où une suite terminale

u devient à droite d'une suite terminale **t** qui est un **A**. C'est ainsi que ces trois représentations constituent la division fondamentale de la formation des mots : **la flexion**, **la dérivation**, et **la composition**.

Un second principe de la théorie de la formation des mots est celui qui considère la flexion[2] comme n'altérant jamais les catégories lexicales à la base. Un nom fléchi demeure un nom et un verbe fléchi demeure un verbe, etc.

Par contre, la morphologie dérivationnelle peut en principe altérer les catégories lexicales à la base.

D'autre part, la représentation linéaire équivalente à (4) et (5) est (6) :

(6) a : [t + u] : Flexion
(6) b : [[t] + u] : Dérivation
(6) c : [[t] + [u] : Composition

Cette présentation linéaire montre clairement que les distinctions peuvent être faites par une inclusion sélective des crochets et par un seul opérateur de concaténation « + » dans la description structurale (**DS**) des règles phonologiques. C'est ainsi, que pour réécrire une règle phonologique qui s'appliquerait uniquement à une suite à l'intérieur du morphème + **u**, nous omettons le symbole de frontière de morphème « + » et les crochets « [ ] », de la description structurale de la règle :

(7) DS: t u
1 2

Cependant, une règle phonologique qui a comme déterminant un segment flexionnel doit être réécrite avec « + » dans la description structurale :

(8) DS: t + u
1 2 3

Par contre, une règle dont le déterminant est un segment dérivationnel doit être réécrite avec un seul crochet dans la description structurale :

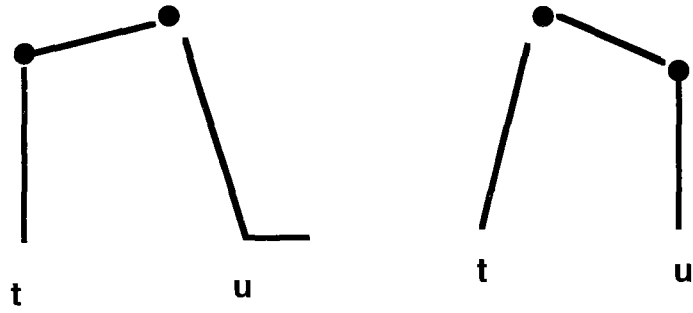
(9a) DS: t] u
1 2 3

(9b) DS: t [u
1 2 3

Il faut préciser que (9a) s'applique à des formes suffixales, tandis que (9b) s'applique à des formes préfixales :

2 Mark Aronoff définit la dérivation et la flexion en ces termes « there are traditionally two types of morphological phenomena, derivational and inflectional. The distinction is delicate, and sometimes elusive, but nonetheless important. Inflection is generally viewed as encompassing the "purely grammatical" markers, those for tense, aspect, person, number, gender, case, etc. Within a lexicalist, theory of syntax (cf. Chomsky, 1970), inflectional morphemes would be dominated by the node X, and perhaps higher nodes (cf. Siegel, 1974), while derivational morphemes would be dominated by the node X. Derivational morphology is thus restricted to the domain of lexical category » (Mark Arnoff, *Word Formation in Generative Grammar*, Cambridge, The MIT Press, 1976, p. 2)

(10a) (10b)



Étant donné les définitions formelles de la flexion, la dérivation et la composition, il n'existe pas d'autres façons pour interpréter la description structurale de la représentation (9).

Enfin, quand une règle phonologique s'applique à des mots composés, sa description structurale spécifiera les descriptions gauche et droite des crochets comme dans (11) :

(11) DS: t] [u
1 2 3 4

Ce système proposé présente des critères définitoires formels qui permettent de se référer à des classes naturelles qui ont les descriptions structurales suivantes :

(12) a : DS: t (]) + u
b : DS: t + ([) u
c : DS: t <]> + <[> u
d : DS: t] ([) u
e : DS: t (]) [*u*
f : DS: t (]) + ([) u
g : DS: t (]) (+) ([) u

Ces descriptions structurales peuvent être développées pour s'appliquer respectivement aux classes naturelles suivantes :

(12')a. Formes suffixées et fléchies :
t] + u
t + u
b. Formes préfixées et fléchies :
t + [u
t + u
c. Formes composées et fléchies :
t] + [u
t + u
d. Formes composées et suffixées :
t] [u
t] u

e. Formes composées et préfixées :
t] [u
t [u
f. Formes composées, dérivées et fléchies :
t] + [u
t + [u
t] + u
t + u
g. Formes composées, dérivées, fléchies et formes internes du morphème :
t] + [u
t + [u
t] + u
t + u
t u

Ces représentations développées montrent que dans (12'd) et (12'e) le symbole « + » des frontières de morphème est omis de leur description étant donné que sa présence à l'input est entièrement prédictible de l'occurrence obligatoire d'au moins un crochet.

Cependant, dans les autres descriptions structurales où les crochets sont tous optionnels, la référence à des segments flexionnels exige l'inclusion du symboles « + ». C'est ainsi, que pour faciliter l'écriture des règles phonologiques conditionnées par les mots structurés, nous adoptons la convention qui considère un terme non segmental (+, [, ou]) dont la présence est entièrement prédictible, comme redondant et par conséquent peut être omis de la description structurale de la règle.

3. Définition formelle des règles de formation de mots

Les abréviations ci-dessous ont les significations suivantes : **t** et **u** désignent des suites terminales, et **A**, **B**, **C**, représentent les catégories grammaticales :

Suffixation flexionnelle	:	[t]A → [t + u]A
Préfixation flexionnelle	:	[t]A → [u + t]A
Suffixation dérivationnelle	:	[t]A → [[t]A + u]B
Préfixation dérivationnelle	:	[t]A → [u + [t]A]B
Composition (par adjonction à droite)	:	[t]A → [[t]A + [u]B]C
Composition (par adjonction à gauche)	:	[t]A → [[u]B + [t]A]C
Dérivation nulle	:	[t]A → [[t]A]B

4. Les règles de formation de mots en arabe

4.1. La dérivation

Pour répondre aux exigences de la vie moderne, l'arabe standard forme un grand nombre d'adjectifs relatifs (**Adj.Rel.**) et de noms abstraits (**NAb.**) par suffixation de la semi-consonne /+ **j** / ou /+**ijj** /. Cette formation obéit à des règles de formation de

mots. En effet, l'arabe moderne forme la plupart de ces adjectifs à partir d'un substantif au singulier (après suppression éventuelle du suffixe final /+**at** /) ou au pluriel sans que sa structure soit modifiée :

a) Adjectifs relatifs :

L'arabe standard moderne fait un usage très fréquent de ces noms de « relation », la « ***nisba*** » des grammairiens arabes (ou relatifs) en / **+ii** /. Ils sont généralement construits à l'aide des « règles de formation de mots » (RFM) de suffixation dérivationnelle du type :

[t]A → [[t]A + u]B

ou **t** : est une suite terminale, **u** : est aussi une suite terminale et **A**, **B** sont des catégories :

RFM_1 :	[t]N →	[[t	]N	+ ii]Adj.	Rel.	
		[[ʒadal	]N	+ ii]Adj.	Rel.	« dialectique »
		[[ɻamal	]N	+ ii]Adj.	Rel.	« pratique »
		[[waðiif	]N	+ ii]Adj.	Rel.	« fonctionnel »
		[[raʔiis	]N	+ ii]Adj.	Rel.	« principal »

Quelques adjectifs relatifs sont construits sur un nom d'action (**N.Act.**) et le suffixe /**+ii**/ :

RFM_2 :	[t]N Act. →	[[t]	N Act.	+ ii]Adj.	Rel.	
		[[taqadum]	N Act.	+ ii]Adj.	Rel.	« progressiste »
		[[taɻaawun]	N Act.	+ ii]Adj.	Rel.	« coopératif »
		[[ʔinhizaam]	N Act.	+ ii]Adj.	Rel.	« défaitiste »

Certains adjectifs relatifs sont construits sur des mots composés (**MC**) et le suffixe / **+ii** / :

RFM_3 :	[t]MC →	[[t	]MC	+ ii]Adj.	Rel.	
		[[laaʃuɻuur	]MC	+ ii]Adj.	Rel.	« inconscient »
		[[raʔsmaal	]MC	+ ii]Adj.	Rel.	« capitaliste »

D'autres adjectifs relatifs sont construits sur des emprunts (**EM**) et le suffixe / + **ii** / :

RFM_4 :	[t]EM →	[[t	]EM	+ ii]Adj.	Rel.	
		[[kahrabaaʔ	]EM	+ ii]Adj.	Rel.	« électrique »
		[[sinimaaʔ	]EM	+ ii]Adj.	Rel.	« cinématographique »

Les noms propres (**NP**) peuvent aussi former des adjectifs relatifs :

RFM_5 :	[t]NP →	[[t	]NP	+ ii]Adj.	Rel.	
		[[naasir	]NP	+ ii]Adj.	Rel.	« nassérien »

Les adjectifs relatifs construits sur des néologismes (**NE**) du type :

RFM_6 : [t]NE →	[[t	]NE	+ ii]Adj. Rel.		
	[[ʔalmaan	]NE	+ ii]Adj. Rel.	« laïque »
	[[qablaan	]NE	+ ii]Adj. Rel.	« a priori »
	[[baʕdaan	]NE	+ ii]Adj. Rel.	« a posteriori »

Enfin, les adjectifs relatifs construits sur des particules (**P**) :

RFM_7 : [t]P →	[[t	]P	+ ii]Adj. Rel.		
	[[ɣajr	]P	+ ii]Adj. Rel.	« altruiste »

b) Les noms abstraits :

L'arabe standard moderne forme un grand nombre de noms abstraits (**NA**) qui correspondent généralement à des noms abstraits français en / **+té** / ou / **+ism** / à l'aide du suffixe / **+ijja** / et qui ont une grande variété d'origines.

Certains sont construits sur des participes passifs (**PP**) à l'aide des règles de formation de mots de suffixation dérivationnelle.

RFM_8 : [t]PP →	[[t	]PP	+ ijja]NA		
	[[masʔuul	]PP	+ ijja]NA	« responsabilité »
	[[maʃruuʕ	]PP	+ ijja]NA	« légitimité »
	[[mawḍuuʕ	]PP	+ ijja]NA	« objectivité »

D'autres noms abstraits sont construits sur des élatifs (**E**) :

RFM_9 : [t]E →	[[t	]E	+ ijja]NA		
	[[ʔaɣlab	]E	+ ijja]NA	« majorité »
	[[ʔaqall	]E	+ ijja]NA	« minorité »

Certains noms abstraits sont construits sur des particules (**P**) :

RFM_{10} : [t]P →	[[t	]P	+ ijja]NA		
	[[kamm	]P	+ ijja]NA	« quantité »
	[[kajf	]P	+ ijja]NA	« modalité »

Parfois les noms abstraits sont construits sur des mots composés (**MC**) :

RFM_{11} : [t]MC →	[[t	]MC	+ ijja]NA		
	[[laadiin	]MC	+ ijja]NA	« irreligiosité »
	[[laaʃajʔ	]MC	+ ijja]NA	« nulleté »
	[[laanihaaʔ	]MC	+ ijja]NA	« infinité »

Cependant, la majorité de ces noms abstraits (**NA**) est formée sur des noms substantifs (**N**) :

RFM_{12} : [t]N →	[[t	]N	+ ijja]NA		
	[[ʔinsaan	]N	+ ijja]NA	« humanité » (qualité)
	[[baʃar	]N	+ ijja]NA	« humanité » (genre)

[[ʒins	]N	+ ijja]NA	« nationalité »
[[ʔaxlaaq	]N	+ ijja]NA	« moralité »
[[ʃaxṣ	]N	+ ijja]NA	« personnalité »

D'autres noms abstraits sont formés plutôt sur des adjectifs (**Adj.**) :

RFM_{13} : [t]Adj. →	[[t	]Adj.	+ ijja]NA	
	[[ħurr	]Adj.	+ ijja]NA	« liberté »
	[[laaʔiik	]Adj.	+ ijja]NA	« laïcité » (fr.)
	[[ʕalmaa	]Adj.	+ ijja]NA	« laïcité »

Quant aux noms abstraits arabes qui correspondent aux noms abstraits français en **/+ism/**, ils ont aussi une grande variété d'origines, certains sont construits sur une particule comme (**P**) :

[[ɣajr]P + ijja]NA « altruisme »

D'autres sur un pronom (**PRO**), comme :

[[ʔanaa]PRO +[n]+ ijja]NA « égoïsme »

D'autres sont formés sur des emprunts (**EM**) :

RFM_{14} : [t]EM →	[[t	]EM	+ ijja]NA	
	[[qajṣar	]EM	+ ijja]NA	« césarisme »
	[[barlamaan	]EM	+ ijja]NA	« parlementarisme »
	[[maarks	]EM	+ ijja]NA	« marxisme »

Parfois aussi sur des mots composés (**MC**) :

RFM_{15} : [t]MC →	[[t	]MC	+ ijja]NA	
	[[raʔsmaal	]MC	+ ijja]NA	« capitalisme »

Cependant, la majorité des noms abstraits sont formés sur des noms (ou substantifs) :

RFM_{16} : [t]N →	[[t	]N	+ ijja]NA	
	[[ʔinsaan	]N	+ ijja]NA	« humanisme »
	[[taʔθiir	]N	+ ijja]NA	« impressionnisme »
	[[ħizb	]N	+ ijja]NA	« sectarisme, partisme »
	[[ðaat	]N	+ ijja]NA	« subjectivisme »
	[[ʔiʃtiraak	]N	+ ijja]NA	« socialisme »
	[[qawm	]N	+ ijja]NA	« nationalisme »
	[[waṭan	]N	+ ijja]NA	« patriotisme »

En guise de conclusion sur la dérivation, nous pouvons récapituler les règles de formation de mots de suffixation dérivationnelle du type :

[t]A → [[t]A + u]B :

RFM_1 :	[t]N	→	[[	t	]N	+ ii	]Adj.	Rel.
RFM_2 :	[t]N Act.	→	[[	t	]N Act.	+ ii	]Adj.	Rel.
RFM_3 :	[t]MC	→	[[	t	]MC	+ ii	]Adj.	Rel.
RFM_4 :	[t]EM	→	[[	t	]EM	+ ii	]Adj.	Rel.
RFM_5 :	[t]NP	→	[[	t	]NP	+ ii	]Adj.	Rel.
RFM_6 :	[t]NE	→	[[	t	]NE	+ ii	]Adj.	Rel.
RFM_7 :	[t]P	→	[[	t	]P	+ ii	]Adj.	Rel.
RFM_8 :	[t]PP	→	[[	t	]PP	+ ijja	]NA	
RFM_9 :	[t]E	→	[[	t	]E	+ ijja	]NA	
RFM_{10} :	[t]P	→	[[	t	]P	+ ijja	]NA	
RFM_{11} :	[t]MC	→	[[	t	]MC	+ ijja	]NA	
RFM_{12} :	[t]N	→	[[	t	]N	+ ijja	]NA	
RFM_{13} :	[t]Adj.	→	[[	t	]Adj.	+ ijja	]NA	
RFM_{14} :	[t]EM	→	[[	t	]EM	+ ijja	]NA	
RFM_{15} :	[t]MC	→	[[	t	]MC	+ ijja	]NA	
RFM_{16} :	[t]N	→	[[	t	]N	+ ijja	]NA	

4.2. La composition

La composition consiste à former un seul mot à partir de deux ou plusieurs mots réunis. Le véritable composé construit un mot nouveau (à sens nouveau) et l'on perd la conscience linguistique des composants.

Comme l'arabe standard moderne est défavorisé sur ce point, il lui est arrivé de former des mots composés à l'aide de deux mots brefs selon la règle de formation de mots composés (RFMC) :

RFMC1 : [t]A	→	[[	t	]A	+	[	u	]B	]MC	
		[[barr		]A	+	[maaʔii		]Adj. Rel.	]MC	« amphibie »
		[[ʒaw		]A	+	[maaʔii		]Adj. Rel.	]MC	« hydravion »

4.2.1. La préfixation

L'arabe standard moderne forme un certain nombre de mots composés par préfixation, selon les règles de formation de mots composés.

a) Les mots composés à préfixe / laa + / :

Le préfixe / laa +/ exprimant la négation permet la formation d'environ une cinquantaine de mots composés, surtout nominaux et rarement verbaux. Certains sont des termes abstraits ou philosophiques formés selon la règle de composition à gauche :

RFMC2 : [t]A	→	[[	u	]B	+ [	t	]A	]MC	
		[[laa		]Préf.	+ [mubaalaat		]N	]MC	« négligence »
		[[laa		]Préf.	+ [diinii		]Adj.R.	]MC	« irreligieux »
		[[laa		]Préf.	+ [ʔiraada		]N	]MC	« automatisme »
		[[laa		]Préf.	+ [ʃuʕuur		]N	]MC	« l'inconscient »

D'autres mots composés sont des termes politiques administratifs ou sociologiques et sont formés selon la même règle :

RFMC2' : [t]A →	[[u	]B	+	[t	]A	]MC	
	[[laa	]Préf.	+	[niðaam	]N	]MC	« anarchie »
	[[laa	]Préf.	+	[muħaaraba	]N	]MC	« non belligérance »
	[[laa	]Préf.	+	[tamarkuz	]N	]MC	« déconcentration »
	[[laa	]Préf.	+	[diimuqraaṭi	]N	]MC	« antidémocratie »

Certains mots composés sont des termes scientifiques ou techniques :

RFMC2" : [t]A →	[[u	]B	+	[t	]A	]MC	
	[[laa	]Préf.	+	[ʒafn	]N	]MC	« ablépharique »
	[[laa	]Préf.	+	[ʒinsii	]Adj.R.	]MC	« asexué »
	[[laa	]Préf.	+	[silkii	]Adj.R.	]MC	« sans fil »
	[[laa	]Préf.	+	[ʔinʃiṭaar	]N	]MC	« non fission »

b) Les mots composés à préfixe / ʃibhu + / :

Le préfixe /ʃibhu+/ entre en composition avec le sens de « semblable à », quasi, para, proto, etc. Ces mots composés sont formés à l'aide de la même règle de composition à gauche :

RFMC3 : [t]A →	[[u	]B	+	[t	]A	]MC	
	[[ʃibhu	]Préf.	+	[ʒaziira	]N	]MC	« presqu'île »
	[[ʃibhu	]Préf.	+	[ṭibii	]Adj.R.	]MC	« paramédical »

Ce préfixe se présente tantôt soudé au mot :

RFMC4 : [t]A →	[[u	]B	+	[t	]A	]MC	
	[[ʃib	]Préf.	+	[billawr	]N	]MC	« cristalloïdes »

c) Les mots composés à préfixe /ɣajru +/ :

Le préfixe / ɣajru+ / est assez productif en arabe standard moderne. Il correspond fréquemment aux préfixes français : **in+**, **a+**, **non**, **de+**, et parfois même, **extra**. Il forme des mots composés selon la règle de composition à gauche :

RFMC5 : [t]A →	[[u	]B	+	[t	]A	]MC	
	[[ɣajru	]Préf.	+	[ʔinsaanii	]N	]MC	« inhumain »
	[[ɣajru	]Préf.	+	[mutabalwar	]Adj.R.	]MC	« amorphe »
	[[ɣajru	]Préf.	+	[ʕaadii	]N	]MC	« extraordinaire »

d) Les mots composés avec / ʕadam + / :

Le préfixe / ʕadam+ / s'emploie en arabe standard moderne généralement devant un substantif en rapport d'annexion déterminé par l'article pour indiquer : l'absence, la privation.

Il correspond souvent aux préfixes français : **a+**, **in+**, **non**, **de+**, **me+**, ils forment

des mots composés selon la même règle de composition à gauche :

RFMC6 : [t]A →	[[	u	]B	+	[	t	]A	]MC	
	[[ʢadam		]Préf.	+	[muʔaaxaða		]N	]MC	« indulgence »
	[[ʢadam		]Préf.	+	[ħadd		]N	]MC	« non-détermination »
	[[ʢadam		]Préf.	+	[ʔirtijaaħ		]N	]MC	« mécontentement »
	[[ʢadam		]Préf.	+	[ʔittifaaq		]N	]MC	« désaccord »

e) Les mots composés avec /ḍidd +/ :

La préposition /ḍidd+/ rend l'idée de contre, /anti+/ en arabe standard moderne. Elle permet la formation de mots composés selon la règle de composition à gauche :

RFMC7 : [t]A →	[[	u	]B	+	[	t	]A	]MC	
	[[ḍidd		]Préf.	+	[ʔal ʃawsasa		]N	]MC	« contre espionnage »
	[[ḍidd		]Préf.	+	[ʔal dabbaat		]N	]MC	« antichars »

f) Les mots composés avec / suuʔ / :

Le préfixe / suuʔ +/ se présente comme le premier terme d'un rapport d'annexion. Il correspond en français à « mauvais », ou préfixes : **/mal+, in+, mès+ /** : il forme des mots composés à l'aide de la même règle de composition à gauche :

RFMC8 : [t]A →	[[	u	]B	+	[	t	]A	]MC	
	[[suuʔ		]Préf.	+	[ʔal baxt		]N	]MC	« malchance »
	[[suuʔ		]Préf.	+	[ʔal ħaal		]N	]MC	« mauvais état »
	[[suuʔ		]Préf.	+	[ʔal fahm		]N	]MC	« incompréhension »
	[[suuʔ		]Préf.	+	[ʔaltafaahum		]N	]MC	« mésentente »
	[[suuʔ		]Préf.	+	[ʔal tanḏiim		]N	]MC	« vices de structure »

Pour conclure sur ce point, nous pouvons dire qu'il existe deux types de composition en arabe standard. Une **composition à droite** du type :

[t]A → [[t]A + [u]A]MC

et une **composition à gauche** du type :

[t]A → [[u]B + [t]A]MC

qui se présentent de la façon suivante :

RFMC1 :	[t]	→	[[	t	]A	+	[u	]B	]MC	SUFF. DÉRIV.
RFMC2 :	[t]A	→	[[ɭaa		]B	+	[t	]A	]MC	PRÉF. DÉRIV.
RFMC3 :	[t]A	→	[[ʃibhu		]B	+	[t	]A	]MC	PRÉF. DÉRIV.
RFMC4 :	[t]A	→	[[ʃib		]B	+	[t	]A	]MC	PRÉF. DÉRIV.
RFMC5 :	[t]A	→	[[ɣajru		]B	+	[t	]A	]MC	PRÉF. DÉRIV.
RFMC6 :	[t]A	→	[[ʢadam		]B	+	[t	]A	]MC	PRÉF. DÉRIV.
RFMC7 :	[t]A	→	[[ḍidd		]B	+	[t	]A	]MC	PRÉF. DÉRIV.
RFMC8 :	[t]A	→	[[suuʔ		]B	+	[t	]A	]MC	PRÉF. DÉRIV.

5. Conclusion

Cette recherche, qui s'inscrit dans le cadre de l'hypothèse lexicaliste, préconise une composante morphologique autonome, dont le siège est le lexique, responsable de la flexion, de la dérivation et de la composition.

Cette étude a permis de spécifier d'une façon détaillée la nature des règles lexicales qui permettent de dériver des mots à partir des mots de base. Cette relation s'accomplit par un ensemble de règles de formation de mots. Ces règles se caractérisent par leur formalisme logico-mathématique. Elles sont informatisables et s'appliquent d'une façon naturelle à des processus morphologiques de l'arabe tels que la flexion, la dérivation et la composition.

En effet, leur application en morphologie arabe est d'une grande importance pour la néologie lexicale, surtout lors de la création de mots nouveaux dans les domaines scientifiques et techniques.

Un générateur est conçu pour engendrer automatiquement des mots nouveaux pour différentes catégories grammaticales (substantifs, adjectifs, noms d'action, etc.) à partir d'un ensemble de règles de formation de mots élaborées pour l'arabe standard.

Par un jeu de dérivation et de composition de mots à partir des mots de base (par préfixation, suffixation et composition), le générateur permet ainsi de mettre à jour et d'enrichir le dictionnaire arabe par des néologismes.

Le générateur est relié à un dictionnaire contenant les mots de base, racines, préfixes et suffixes, avec des informations grammaticales associées aux mots ; et à un ensemble de règles de formation de mots.

Lexicographie berbère. Construction des formes de mot et classification des entrées lexicales

Miloud TAIFI

Université de Fès, Maroc

À la mémoire de mon ami et collègue Kaddour CADI

État des lieux

La langue berbère occupe un vaste espace allant de l'oasis de Siwa en Égypte jusqu'en Afrique noire (Niger, Mali et Burkina Faso) en passant par le Maghreb qui constitue véritablement le fief du berbère, de par le nombre très important des populations berbérophones en Algérie et surtout au Maroc. La langue berbère est constituée de plusieurs dialectes ou supra-systèmes qui s'étendent sur des zones géographiques plus ou moins étanches : on dénombre ainsi le touareg (dans les régions sud-sahariennes algériennes, au Mali et au Niger), le tachelhiyt, le tamazight et le tirifiyt au Maroc, le kabyle, le hchaouit et le tamzaybit en Algérie (cf. Galand, 1988 : 207-242).

Les études sur le berbère sont très anciennes, mais les véritables traités de grammaire et les premiers recensements de vocabulaire datent de la deuxième moitié du XIXe siècle. Le premier lexique bilingue berbère/français (dialectes Algérie) est publié en 1844 par Venture de Paradis. Ont été édités ensuite plusieurs travaux de lexicologie et des inventaires de vocabulaire dont les plus importants, en nous astreignant aux dialectes marocains, sont les travaux de lexicologie de Destaing sur vocabulaire tachelhiyt en 1920 et de Laoust sur les mots et choses berbères, l'étude Loubignac sur le berbère des Zaïan et Ait-Sgougou en 1924 et celle de Mercier sur le dialecte des Ait-Izdeg en 1937.

Mais le premier véritable dictionnaire est sans doute celui de Charles de Foucauld : *Dictionnaire touareg-français* (*dialecte de l'Ahaggar*) publié en 1951. Est paru ensuite, en 1982, le *Dictionnaire kabyle-français* de Jean-Marie Dallet, à titre posthume. Le *Dictionnaire mozabite-français* de Jean Delheure vient en 1984, augmenter les travaux lexicographiques berbères. Le dernier travail dans le domaine est à ce jour, le *Dictionnaire tamazight-français* (*parlers du Maroc central*) que j'ai publié

moi-même en 1992. D'autres travaux lexicographiques de grande envergure sont actuellement en cours de réalisation dans le cadre de la préparation de thèses de doctorat.

Les quatre importants dictionnaires cités, recouvrant différents dialectes de la langue berbère constituent une somme considérable de données lexicales et assoient, du point de vue méthodologique, une tradition lexicographique. Ils ont en effet tous adopté, avec quelques options et amendements partiels, la classification par racines, sacrifiant ainsi aux exigences mêmes de la morphologie du berbère qui construit les formes de mot en associant les racines et les schèmes.

Les premières monographies de morphologie berbère ont abordé la construction des formes de mot à travers la morphologie des langues romanes, notamment le français. Plusieurs auteurs ont ainsi essayé de retrouver dans le berbère les procédés de dérivation affixale dont la segmentation isole les bases lexématiques et les morphèmes affixaux (préfixes, infixes et suffixes). Mais bien vite l'application des principes d'analyse valables pour les langues romanes s'avéra impropre à la langue berbère. On découvrit en effet que le berbère appartient à la famille chamito-sémitique et qu'il fallait par conséquent chercher du côté de la morphologie du sémitique.

L'appartenance et l'apparentement du berbère à la famille des langues chamito-sémitiques sont fondés sur plusieurs aspects communs et suffisants pour justifier, du point de vue linguistique, les rapprochements entre le berbère et le sémitique et, du point de vue méthodologique, l'application des mêmes paramètres d'analyse et de description (cf. Galand, 1979a : 463-478). D'où le transfert de la racine sémitique au domaine lexical berbère et l'adoption, dans la pratique lexicographique, de la classification par racines (cf. Cohen, 1993 : 161-175). Mais ce transfert, justifié par la même morphologie du berbère, fait apparaître d'innombrables problèmes tant théoriques que pratiques. C'est de ces problèmes que traitera cette communication.

Construction des formes de mot en berbère

Racines et schèmes

Les formes de mot en berbère sont toutes des formes construites par l'association de deux constituants formels : une racine et un schème. Le premier constituant représente le lexique, le second la morphologie ou plus exactement la grammaire. La racine est généralement définie comme un groupe de consonnes se présentant dans un ordre impératif et qui constitue l'invariant formel d'un paradigme lexical ; le schème comme une structure formelle comportant des éléments vocaliques et/ou consonantiques et assignant des places destinées à être occupées par les radicales de la racine. Le schème porte théoriquement un sens grammatical, puisqu'il catégorise les formes de mot construites en différentes parties de discours, comme le montre la figure suivante (cf. Chaker, 1984 : 136).

RACINES LEXICALES INDIFFÉRENCIÉES

⇓

FORMES DE MOT

⇓	⇓	⇓
Marques verbales	Marques nominales	Marques zéro
VERBES	NOMS	DÉTERMINANTS AUTONOMES
– formes simples	– substantifs	CONNECTEURS
– formes complexes	– adjectifs	– prépositions
	– numéraux	– coordonnants
	– pronoms	– conjonctions

Ainsi, par exemple, la racine MGR porteuse du sens « moisson », est commune à toutes les formes de mot suivantes attestées en berbère.

Formes verbales : *mger, mgir, megger, ttumger, ttumgir, ttumgar.*
Formes participiales : *imgern, imgirn, mgernin, ittumgern, ittumgirn, ttumgernin.*
Formes nominales : *amgar, imgarn, amggar, imeggarn, amg^{w}er, imeg^{w}ran, tamg^{w}ert, timegrin.*

Pour montrer cette communauté formelle, nous allons dégager le schème de chaque forme de mot en remplaçant chaque radicale de la racine par le symbole C que nous notons avec un trait souscrit à chaque fois que la radicale est tendue dans la forme de mot : C̲. Nous obtenons ainsi :

– CCC, CCIC, C̲C̲C, ttuCCC, ttuCCiC, ttuCCaC
– iCCCn, iCCiCn, CCCnin, ittuCCCn, ittuCCiCn, ttuCCCnin
– aCCaC, iCCaCn, aC̲C̲ar, iC̲C̲aCn, aCCC, iCCCan, taCCCt, tiCCCin.

Comme le montre cet exemple, la formation du mot se fait par dérivation associative qui consiste en l'insertion des radicales d'une racine dans les places vides du schème. Une telle insertion est régie par des règles morphologiques dépendant des associations phonétiques et/ou sémantiques permises par la langue. Pour les premières, certains voisinages de phonèmes sont neutralisés de par la nature articulatoire de ces derniers. Ainsi les suites consonantiques k/G, G/K, X/γ, γ/X, Q/γ, γ/Q sont rares, sinon exclues parce qu'elles sont imprononçables. Pour les secondes, c'est l'ordre linéaire des radicales qui confère à la racine son sens lexical : la racine MGR, notée ci-dessus, rend la notion de « moisson », invariant sémantique que se partagent toutes les formes de mot qui dérivent de la racine MGR.

Le changement de la racine par permutation des radicales engendre d'autres suites qui peuvent être soit attestées et constituer le chef de file d'un paradigme lexical, soit non attestées et donner lieu à des créations nouvelles (néologismes ou monstres linguistiques). La permutation de MGR fournit six racines trilitères dont une seule est non attestée (dans le dictionnaire de Taifi, 1992).

MGR : « notion de moisson »
MRG : « notion d'amour »
GMR : « deux notions » : 1) « chasse »; 2) « cheval »

GRM : « trois notions » : 1) « ronger, grignoter »; 2) « qui a une seule corne, qui est sans cornes (ovin) »; 3) « saint, marabout ».
RMG*
RGM : « notion d'insulte, de malédiction ».

L'ordre RMG n'est donc pas exploité par la morphologie du berbère (on notera du moins qu'en Kabyle RMG avec R emphatique rend le sens de « tonner » : Dallet, 1982 : 726). Les autres racines donnent lieu à des familles lexicales dont chacune est constituée d'un certain nombre de formes de mot ; chaque famille forme un champ morpho-sémantique. Ainsi la racine GMR dont le contenu lexical relève de la notion de « chasse » fournit le champ morpho-sémantique suivant comportant 17 formes de mot attestées :

verbes :	*gmer, gmir, gemmer, ttugmer, ttugmir, ttugmar*
participes :	*igmern, igmirn, gmernin, ittugmern, ittugmirn, ttugmernin*
noms :	*tagemrawt, tigemrawin, tanegmart, anegmar, inegmarn*

construites respectivement sur les schèmes suivants :

– CCC, CCIC, C<u>C</u>C, ttuCCC, ttuCCIC, ttuCCaC
– iCCCn, iCCiCn, CCCnin, ittuCCCn, ittuCCiCn, ttuCCCnin
– taCCCawt, tiCCCawin, tanCCaCt, anCCaC, inCCaCn

De même, le champ morpho-sémantique de la racine RGM relative au domaine notionnel de « insulte et malédiction » est composé de 20 formes de mot :

verbes :	*rgem, rgim, reggem, tturgem, tturgim, tturgam, mergam, ttemergam*
participes :	*irgemn, irgimn, rgemnin, itturegmen, itturgimn, tturegmnin, mergamnin, ttemergamnin*
noms :	*argam, irgamn, tareggimt, tirggam*

dont les schèmes se présentent ainsi :

– CCC, CCIC, C<u>C</u>C, ttuCCC, ttuCCiC, ttuCCaC, mCCaC, ttmCCCam
– iCCCn, iCCiCn, CCCnin, ittuCCCn, ittuCCiCn, ttuCCCnin, mCCaCnin, ttmCCaCnin
– aCCaC, iCCaCn, taC<u>C</u>iCt, tiC<u>C</u>aC

Les trois champs morpho-sémantiques présentés répondent à la définition des racines et des schèmes donnée par Cantineau (1950 : 74) et souvent citée pour décrire la construction lexicale dans les langues appartenant à la famille chamito-sémitique.

> Chaque mot a sa racine et son schème; on pourrait comparer le vocabulaire à un tissu dont la trame serait l'ensemble des racines attestées dans la langue et la chaîne l'ensemble des schèmes existants. Chaque point d'intersection de la chaîne, et de la trame, serait un mot, car tout mot est entièrement défini sans ambiguïté par sa racine et son schème, tout schème de son côté fournissant des mots à différentes racines et la plupart des racines fournissant des mots de différents schèmes.

Cette métaphore de tisserand n'explique cependant pas tout. Si le schème relève de la morphologie et constitue un cadre formel prêt à accueillir les radicales de la racine, celle-ci, par contre, n'est pas suffisamment et clairement définie. Premièrement, la racine est-elle exclusivement consonantique, ou bien y a-t-il lieu de considérer certains segments vocaliques comme radicales dans les cas où ils sont constants et ne subissent pas de changement ou d'effacement ? Deuxièmement, est-ce que la racine est tout simplement un groupe d'éléments commun à une série de formes de mot, ou bien est-ce un signifiant doté de signifié précis ? Autrement dit, la racine est-elle seulement une unité formelle, ou une unité formelle et sémantique. Apporter des réponses à ces questions par l'analyse lexicologique, est un préalable à toute pratique lexicographique berbère, et aussi, dans certaines mesures, à celle de l'arabe.

Racine : consonnes et voyelles

Si le critère qui préside à l'établissement d'une racine dans une famille lexicale est l'invariabilité de ses radicales dans tous les lexèmes construits, il n'y a pas lieu d'exclure les éléments vocaliques qui répondent à ce critère. Si certaines voyelles sont constantes, elles ne peuvent appartenir qu'à la racine et non aux schèmes. La définition donnée par Meillet (cité par Cohen, 1993 : 162) corrobore ce point de vue : « Un mot " appartient " à une racine, il fait partie d'un ensemble de mots ayant en commun un groupe de phonèmes auquel est associé un certain sens général. » Ainsi les radicales d'une racine, selon l'auteur, sont des phonèmes, ceux-ci pouvant être soit consonantiques ou vocaliques, l'essentiel étant leur régularité dans tous les mots appartenant à une même famille lexicale. Il faut ajouter que le critère de la constance n'est valide que si les voyelles occupent toujours la même place dans les schèmes.

Ceci amène Cohen (1993 : 162), commentant la citation de Meillet, à définir la racine ainsi : « la racine est une séquence ordonnée de phonèmes qui constitue la totalité des éléments communs à un ensemble dérivatif », en remarquant que si la racine a été toujours considérée exclusivement consonantique dans les langues chamito-sémitiques, ce n'est qu'un fait d'observation : les consonnes sont en effet (surtout en arabe classique, pris comme référence) beaucoup plus sujettes à la constance que ne le sont les voyelles ; mais ceci ne justifie pas l'exclusion des voyelles constantes de la racine. Cohen propose par conséquent, pour le berbère, de conférer aux voyelles régulières le statut de radicales et d'en tenir compte dans la pratique lexicographique. Donnons un exemple pour illustrer ce point de vue.

Soit les formes de mot suivantes :

aḍar	« pied »	schème : iCaC
iḍarn	« pieds »	schème : iCaC (n)
taḍartt	« petit pied, pieds d'enfant »	schème : (t)aCaC(tt)
tiḍarin	« petits pieds, pieds d'enfant »	schème : (t)iCaC(in)

Comme on le constate, ces formes se partagent la séquence -ḍar- qui sera ainsi la racine CVC = ḌAR. Les éléments mis entre parenthèses sont les marques du genre et du nombre. Par contre, dans le paradigme suivant, aucune voyelle n'est constante en occupant la même place dans les différents schèmes :

afus	« main »	schème : aCuC
ifassen	« mains »	schème : iCaC(n)

tafustt	« petite main, main d'enfant »	schème : (t)aCuC(t)
tifassin	« petites mains, mains d'enfant »	schème : (t)iCaC (in)

Dans ce cas, seules les consonnes constitueront la racine, à savoir CC : FS, puisque les lexèmes dérivés n'ont en commun aucune voyelle constante dans la même position.

L'intérêt de la promotion des voyelles constantes au statut de radicales, permettra, selon Cohen (1993 : 161-175), de pouvoir distinguer les racines homophones, surtout les monolitères et les bilitères, en réduisant leur nombre par l'isolement de celles qui comportent une ou plusieurs voyelles constantes. Ainsi, par exemple, au lieu de sept racines consonantiques homophones ḌR fournissant chacune un paradigme lexical (cf. Taifi, 1992 : 91-93), il n'y en aura que quatre, si l'on classe à part celles qui contiennent un élément vocalique régulier. Voici les données :

ḌR$_1$:
Sens général : « descendre, baisser (intransitif) ».
Formes de mot :
Verbes : *ḍer, tḍar/ḍḍar, ḍir, sḍer, sḍir, ttesḍar, ttusḍer, ttusḍir, ttusḍar.*
Participes : *iḍern, iḍirn, ḍernin, iḍḍarn, ḍḍarnin, isḍern, isḍirn, sḍirnin, ittesḍarn, ttesḍarnin, ittusḍern, ittusḍirn, ttusḍarnin.*
Noms : *taḍuri, taḍurin, asḍar, isḍarn.*

ḌR$_2$:
Sens général : « salir, souiller » .
Formes de mot :
Verbes : *aḍer, uḍer, uḍir,ttaḍer, tyiḍer, tyaḍar, myaḍar, temyaḍar.*
Participes : *yuḍern, yuḍirn, uḍernin, ityiḍern, tyiḍernin, ityaḍarn, tyaḍarnin, myaḍarnin.*
Noms : *iḍer.*

ḌR$_3$:
Sens général : « être sourd » .
Formes de mot :
Verbes : *ḍurḍer, tḍurḍur, ḍurḍir.*
Participes : *iḍurḍern, iḍurḍirn, ḍurḍernin, itḍurḍurn, tḍurḍurnin.*
Noms : *aḍerḍur, iḍerḍurn, taḍerḍurt, tiḍerḍurin, tiḍerḍert.*

ḌR$_4$:
Sens général : « nuire, faire mal »
Formes de mot :
Verbes : *ḍerra, tḍerra/ḍḍerra, tuḍerra, ttuḍerra, mḍerra, temḍerra.*
Participes : *iḍerran, ḍerranin, itḍerran, tḍerranin, ituḍerran, tuḍerranin, mḍerranin, temḍerranin.*
Noms : *aḍerra, lmaḍerra, lmaḍerrat, ḍḍarar.*

ḌR$_5$:
Sens général : « pied »
Formes de mot :
Noms : *aḍar, iḍarn, taḍartt, tiḍarin.*

ḌR$_6$:
Sens général : « maïs »
Formes de mot :
Noms : *ḍḍra, aḍḍra, aḍḍraten, taḍḍrat, taḍḍratin.*

Les racines 5 et 6 sont constituées de consonnes et de voyelles, celles-ci étant constantes, et se présentent ainsi : ḌAR, ḌRA. La différence de position (médiane # finale) permet donc de distinguer par la constance vocalique les deux racines à l'origine homophones. Ces racines ne seront plus alors considérées comme bilitères, mais s'ajouteront à l'ensemble des trilitères, puisqu'elles sont formées de trois radicales. Le traitement lexicographique placera ces nouvelles formes à leur place dans l'ordre alphabétique : ḌAR, ḌR (1, 2, 3, 4,), ḌRA.

Le second exemple illustre la même procédure ; cette fois, le paradigme des racines homophones est plus fourni : il comporte dix formes semblables :

LS$_1$:
Sens général : « se vêtir, s'habiller ».
Formes de mot :
Verbes : *lsi, lsa, lessa, lsi, ttulsa, ttulsi, ssels, sselsi, sselsa, mselsa, ttemselsa.*
Participes : *ilsan, ilsin, lsanin, ilessan, lessanin, ittulsan, ittulsin, ttulsanin, isselsen, isselsin, sselsanin, mselsanin, ttemselsan.*
Noms : *melsiwt, melsiwat, timelsit, timelsa, assels, isselsan, aselsu, iselsa.*

LS$_2$:
Sens général : « tondre ».
Formes de mot :
Verbes : *lles, telles, llis, ttulles, ttullas.*
Participes : *illsen, illisn, llesnin, ittellesn, ittullsen, ittullasn, ttullasnin.*
Noms : *talasa, talusi, ulus, ilis, ilisn, tilist, tilisin, amlas, imlasn, amlus, imlas, imlusen.*

LS$_3$:
Sens général : « souiller, salir ».
Formes de mot :
Verbes : *lles, telles, llis, ulus.*
Participes : *illesn, illisn, llesnin, itellesn, tellesnin.*
Noms : *ulus, ulusn.*

LS$_4$:
Sens général : « être obscur, sombre; faire noir ».
Formes de mot :
Verbes : *lles, telles, llis, ssuls, ssulus.*
Participes : *illesn, illisn, llesnin, issulsen, ssulsnin.*
Noms : *tallest, tillas.*

LS$_5$:
Sens général : « recommencer, refaire, répéter ».
Formes de mot :

Verbes : *als, ttals, ulis.*
Participes : *yulsen, yulisn, ulesnin, ittalsen, ttalesnin.*
Noms : *alas, ils.*

LS_6 :
Sens général : « être écarté de la succession du grand-père par ses oncles paternels, par suite de la mort de son père (petit-fils) ».
Formes de mot :
Verbes : *als, uls, ttels, ulis.*
Participes : *yulsen, yulisn, ulesnin, ittalsen, ttalesnin.*
Noms : *ulus, amalas, imalasn.*

LS_7 :
Sens général : « repas de la fin de la matinée ».
Formes de mot :
Noms : *allas, allasn.*

LS_8 :
Sens général : « frère/sœur du mari (pour la femme) ».
Formes de mot :
Noms : *alus, ilusn, talust, tilusin.*

LS_9 :
Sens général : « langue (organe et idiome) ».
Formes de mot :
Noms : *ils, alsiwn, tilset, tilsatin.*

LS_{10} :
Sens général : « mousse de savon ».
Formes de mot :
Noms : *alus, ilusa.*

Dans ce deuxième exemple, nous avons donc dix racines homophones LS. Mais si l'on tient compte des voyelles constantes, les racines 7, 8 et 9 doivent être exclues de cet ensemble : les formes de mot de LS_7 comportent, en effet, un /a/ constant, de même celles de LS_8 et de LS_{10} un /u/ régulier. Les trois nouvelles racines ainsi dégagées sont LAS, LUS et LUS devenant des trilitères, de par l'élément vocalique introduit. La nouvelle classification par ordre alphabétique sera : LAS, LS (1, 2, 3, 4, 5, 9) et LUS. Comme on l'aura remarqué, si l'homophonie de LS est ainsi réduite, il y a cependant création d'un autre cas de ressemblance formelle entre LUS < LS_8 et LUS < LS_9. On aura remarqué aussi que la voyelle a plus de chances de rester constante lorsque la racine ne fournit que très peu de formes de mot, c'est le cas effectivement des racines LAS (2 dérivés), LUS (4 dérivés) et LUS (2 dérivés).

Si la constance vocalique a, sans doute, un statut théorique qui exige l'insertion des voyelles dans l'armature des racines, répondant aux définitions de Meillet et de Cohen, il n'en demeure pas moins que son application dans la pratique lexicographique pose, du moins pour le berbère, plus de problèmes qu'elle n'en résout, car les voyelles sont plus alternantes que constantes. Et ceci pour plusieurs raisons :

1– Le nombre des voyelles en berbère est de trois phonèmes : /a/, /i/ et /u/. Ce sont celles-là qui constituent le triangle vocalique de base. Les trois voyelles connaissent cependant des allophones dus à certains environnements consonantiques emphatiques ou vélaires qui imposent une plus grande aperture vocalique. Ces allophones, n'apparaissent donc que lors de la construction des formes de mot ou des séquences syntagmatiques.

Sachant que plus un ensemble est réduit, plus les éléments qui le composent sont fréquemment utilisés, les trois voyelles basiques doivent donc nécessairement, et le plus souvent, alterner pour différencier les schèmes qui accueillent les racines. Une telle différenciation formelle des unités lexicales à travers les schèmes n'est possible en berbère que si les voyelles constitutives des schèmes alternent. C'est ce qui explique, sans doute, que la constance vocalique est très rare, sinon impossible, dans des paradigmes lexicaux comportant plusieurs formes de mot. C'est le cas notamment de ceux qui sont construits à partir des racines verbo-nominales : exemple de la racine FD (Taifi, 1992 : 103-104) qui fournit treize formes verbales (simples et complexes) et sept formes nominales (sans compter les formes participiales de chaque verbe) ; de même, la racine SY (Taifi, 1992 : 663-664) fournit vingt-trois formes verbales (simples et complexes) et huit formes nominales. Étant donné que chaque forme de mot doit s'articuler sur un schème spécifique, pour éviter une homophonie excessive, la morphologie berbère recourt aux voyelles, et comme celles-ci ne sont que trois, le jeu formel d'alternance, par des opérations de commutation (remplacement d'une voyelle par une autre) et de permutation (changement de position) est la seule voix de salut.

C'est ce rôle morphologique (laborieux sans conteste : elles ne sont que trois !) assuré par les voyelles qui a fait dire, avec raison, à André Basset (1929 : XXV) qu'en berbère, « la voyelle s'affirme par ailleurs comme un élément morphologique pour qu'on puisse lui attribuer pareille valeur même là où elle forme avec des éléments consonantiques, un ensemble invariable ». La valeur dont parle Basset est celle de l'appartenance de la voyelle à la racine. Il faut noter qu'une telle constatation n'exclut pas la constance vocalique. Mais celle-ci ne peut être observée que pour des racines à paradigme lexical réduit, notamment les racines exclusivement nominales ou celles, très peu nombreuses, qui fournissent des outils grammaticaux (connecteurs, conjonctions, prépositions...)

2 – L'alternance intervient, par ailleurs, dans la conjugaison des verbes selon les personnes et les valeurs aspectuelles (cf. Galand, 1984 : 304-315) et affecte essentiellement les initiales et les finales des formes verbales ; exemple : le verbe construit à partir de la racine F et signifiant « trouver » se conjugue ainsi :

Aoriste : *af-x, t-af-d, y-af, t-af, n-af, t-af-m, taf-mt, af-n, af-nt*
Accompli : *ufi-x, t-ufi-d, y-ufa, t-ufa, n-ufa, t-ufa-m, t-ufa-mt, ufa-n, ufa-nt*
Accompli négatif : *ufi-x, t-ufi-d, y-ufi, t-ufi, n-ufi, t-ufim, t-ufimt, ufi-n, ufi-nt*
Inaccompli : *ttafa-x, ttafa-d, ittafa, t-tafa, n-ttafa, ttafa-m, ttafa-mt, ttafa-n, ttafa-nt.*

Les voyelles assurent donc, dans la conjugaison, un rôle morphologique et leur alternance est un critère de distinction des thèmes verbaux. On aura ainsi, en éliminant les indices de personne (pronoms personnels) et le formant du schème de l'inaccompli -tt/t, quatre formes du même verbe : *af, ufi, ufa, afa*, dans lesquelles seule la consonne radicale F est constante : aC, uCi, uCa, aCa.

3 – L'alternance vocalique caractérise aussi les formes nominales à cause de l'opposition État libre/État d'annexion. : les noms au singulier ayant un /a/ à l'initiale et les noms féminins ayant (ta) subissent des changements dans certains contextes syntagmatiques : la voyelle /a/ est réalisée /u/ pour le singulier et elle est effacée pour le féminin. Ainsi, par exemple, le nom *argaz* « homme » est réalisé *urgaz* (a > u) lorsque le nom est postposé au verbe et a la fonction de complément explicatif : *argaz, i-rwel*, « l'homme, il s'est enfui » en contraste avec *i-rwel urgaz*, « il s'est enfui, l'homme ». De même *tameṭṭuṭṭ* « femme » devient *tmeṭṭuṭṭ* (a > Ø) dans le même contexte syntaxique : *tameṭṭuṭṭt, t-rwel*, « la femme, elle s'est enfuie » et *t-rwel tmeṭṭuṭṭ*, « elle s'est enfuie, la femme ».

4 – L'alternance vocalique est due aussi aux changements morphologiques relatifs à la catégorie du nombre, surtout pour les pluriels dits « internes » : exemple : singulier : *amazir* « campement » , pluriel : *imizar* (a > i, a > i, i > a) ; singulier : *adaku* « sandale », pluriel, *iduka* (a > i, a > u, u > a).

5 – Les semi-voyelles sont parfois réalisées en voyelles correspondantes (y >i, w > u,) ce qui complique davantage la reconstitution des voyelles radicales (cf. Taifi, 1990b : 219-232).

6 – L'alternance, sans être d'origine structurale, affecte les voyelles différemment selon les dialectes berbères, elle est, dans ce cas, une marque de l'habitus articulatoire d'un groupement géo-linguistique : le processus de dialectalisation de la langue berbère a eu comme conséquence l'accentuation des particularismes phonétiques concernant aussi bien les phonèmes consonantiques que vocaliques : exemple de différence dialectale due à l'alternance vocalique : le terme qui signifie « ficelle » est *ifili* dans un parler, mais *ifilu* dans un autre, la finale étant un /i/ dans l'un et un /u/ dans l'autre ; de même dans *abaxxu/abuxxu*, « insecte » et *ablullu/iblelli*, « papillon de nuit », il y a alternance vocalique dialectale.

En conclusion à cette présentation, il semble que l'introduction des radicales vocaliques dans la reconstitution des racines n'est pas efficace et rentable dans un travail lexicographique, dans la mesure où la constance vocalique ne concerne que très peu de racines et que son utilisation n'a pas, par conséquent, une grande incidence méthodologique quant à la différenciation des racines homophones.

Racine : tension et réduplication des radicales consonantiques

La lexicographie berbère adoptant la classification par racines est confrontée à deux autres phénomènes qui caractérisent certaines racines. Le premier concerne la tension des radicales et le second leur réduplication. Pour le premier cas, la question (pour le lexicographe) est la suivante : faut-il considérer le trait articulatoire de tension comme critère de distinction des racines et noter par conséquent les radicales tendues ? Ou bien considérer qu'une même consonne tendue (simple) ou non tendue est en fait une seule radicale. Illustrons le phénomène.

Soit les formes de mot : *abrid, iberdan* « chemin(s) » construites à partir de la racine BRD, et les formes de mot *aberrad, iberradn* « théière(s) » dérivées aussi de la

racine BRD, mais dont la deuxième radicale est tendue : BRRD. La tension est donc un trait différentiel qui distingue les deux racines, comme elle l'est aussi dans *aglu* « gésier » en opposition à *agella* « tenture ». Théoriquement, les deux racines (à radicale non tendue # à radicale tendue) ne peuvent être traitées comme homophones et doivent donc être classées séparément. La tension serait ainsi réductrice d'homophonie.

Les faits ne sont pas cependant aussi simples. Il faut noter qu'en berbère, la tension a un double rôle : elle est d'abord un trait phonologique distinguant des paires minimales : *ilis* « toison » # *illis* « sa fille », *kes* « paître » # *kkes* « enlever » ; mais elle est aussi un formant du schème et assure donc un rôle morphologique : l'inaccompli est rendu en berbère, soit par des schèmes comportant le formant consonantique t/tt à l'initiale : *afd* > *ttafd* « s'en aller », *af* > *ttafa* « trouver », *ddu* > *teddu* « partir » ..., soit par la tension de l'une des radicales de la racine : *rzu* > *rezzu* « chercher », *mger* > *megger* « moissonner », *γmes* > *γemmes* « couvrir » ...

Le même trait articulatoire est utilisé en outre dans la formation des dérivés nominaux : de nombreux noms d'agent et noms qualificatifs sont construits à partir des racines trilitères sur le schème aCC̲aC dans lequel la deuxième radicale est tendue sans qu'elle le soit forcément dans les autres formes de mot fournies par les mêmes racines (voir Taifi, 1989 : 872-926) : ainsi des racines MDY, NBḌ, ZDM, sont dérivés respectivement les verbes *mdey* « guetter », *nbeḍ*, « commander » et *zdem*, « chercher du bois », dans lesquels les radicales sont non tendues. Par contre, les noms d'agent correspondants ont la deuxième radicale tendue : *amedday*, *anebbaḍ*, *azeddam*. De même, le schème aCC̲aC constitue le cadre de formation pour certains noms qualificatifs : BXN > *bxin* « être noir et *abexxan* « noir » ; WSR > *wsir* « être vieux » et *awessar* « vieux » , LWγ > lwiγ, « être mou » et *alegg*^{w}*aγ* « mou » (*ww* > *gg*^{w}).

Ces exemples montrent donc que la tension relève plus de la morphologie, puisqu'elle est formant du schème, que du lexique. Un tel constat est suffisant pour soutenir que toutes les formes de mot avec une consonne tendue ne proviennent pas automatiquement de racines à radicale tendue. Ce point de vue est corroboré d'ailleurs par l'apparition aléatoire de la tension : dans des cas où elle ne joue aucun rôle morphologique, elle peut être en effet le résultat d'une assimilation phonétique : *tirnit* < *tirrit* « victoire », *anli* < *alli* « cerveau », ou tout simplement gratuite, imposée par l'arbitraire du signe, et n'affectant que certaines formes de mot d'une même famille lexicale : *lles* « tondre » mais *ilis* « toison » et *talasa* « action de tondre » ; *ffeγ* « sortir » mais *ufuγ* « action de sortir » .

La pertinence phonologique de la tension qui en aurait fait une caractéristique de la racine et, par conséquent, une propriété du lexique, est neutralisée par son rôle morphologique et grammatical plus dominant en berbère, et aussi, si l'on s'en tient seulement aux structures des mots, c'est-à-dire à l'arbitraire des signifiants, à l'aléatoire des régularités formelles des familles lexicales. Il est donc plus économique, dans un travail lexicographique, de ne pas tenir compte de ce trait dans l'établissement et la classification des racines dans un dictionnaire.

Le second fait lexical est la réduplication des radicales consonantiques, que nous illustrons par les exemples suivants : les formes de mot *adrar* « montagne » et *adlal* « longue tresse de cheveux » comportent respectivement trois consonnes DRR et DLL

qui constituent les racines à partir desquelles elles sont construites. On remarque que la deuxième et la troisième radicale sont de même nature consonantique, il y a donc réduplication d'une radicale. Faut-il donc, dans la classification par racines, tenir compte de la réduplication ou non ? Les formes de mot *adrar* et *adlal* seront-elles répertoriées respectivement sous les racines DRR et DLL ou bien tout simplement sous DR et DL. Dans le premier cas, les racines sont des trilitères et dans le second des bilitères.

La réduplication des radicales se présente dans deux cas de figure : la racine redupliquée reste invariable et commune à toutes les unités lexicales qu'elle informe. Ainsi la racine QQŠ, rendant le sens de « épier (à travers une ouverture) », fournit les formes verbales *qiqš, tqiqiš, tuqiqš, tuqiqiš, mqiqiš, ttemqiqiš*, les participes *iqiqšn, itqiqišn, tqiqišnin, ituqiqšn, ituqiqišn, mqiqišnin, ttemqiqišn*, et les formes nominales *aqiqš, iqiqšen*. Dans cette famille lexicale, la racine dont la première radicale est redupliquée, constitue la base commune à toutes les formes de mot construites.

Il nous semble, dans ce cas, en nous tenant seulement au niveau méthodologique, qu'il n'y a aucun intérêt pour le lexicographe, à distinguer, dans la procédure de classification, entre les racines redupliquées et celles qui ne le sont pas. Autrement dit, il serait plus économique, pour éviter une dispersion exagérée des formes, de répertorier la racine ci-dessus à l'adresse QŠ et non à QQŠ. Le même principe de classification peut être appliqué aussi pour des racines qui connaissent une double réduplication, procédé utilisé en berbère généralement dans un but expressif (cf. Azougarh, 1992 : 114-135). Les racines dγDγ et FLFL fournissent, la première, avec le sens de « être contusionné, meurtri », les verbes *deγdeγ, dγdiγ, tdeγdaγ, sdeγdeγ, sdeγdaγ, sdeγdiγ*, les participes *ideγdγen, deγdeγnin, isdeγdeγn, sdeγdeγnin* et les noms *adeγdeγ, asdeγdeγ* ; la seconde, avec le sens de « déborder (liquide qui bout) », les verbes *flufel, ttefluful*, les participes *ifluflen, flufelnin, itteflufuln, tteflufulin* et le nom *aflufel*. Ces racines illustrent la double réduplication $C_1C_2C_1C_2$ et seront donc classées comme bilitères C_1C_2 : Dγ et FL.

Une telle option doit être cependant amendée dans les cas où les radicales redupliquées ne sont pas limitrophes. Les racines quadrilitères suivantes comportent des radicales redupliquées : ŠNŠL ayant le sens de « secouer violemment » et fournissant les verbes *šenšel, šenšil, tšenšil, ttušenšel, ttušenšil, ttušenšal*, les participes *išenšeln, išenšiln, šenšelnin, šenšalnin, tšenšalnin*, et les noms *ašenšel, išenšeln* et SKSW avec le sens de « regarder, voir » et étant la racine des verbes *seksiw, sseksiw*, des participes *iseksiwn, seksiwnin, sseksiwnin* et du nom *aseksiw*. Les deux racines ont la structure $C_1C_2C_3C_4$ dans laquelle les deux radicales C_1 et C_3, quoique identiques, sont séparées cependant par une autre radicale et ne sont donc pas limitrophes. Pour ces cas, il y a nécessité procédurale, afin d'identifier ces racines, de les classer sous leur forme initiale, c'est-à-dire comme quatrilitères : ŠNŠL et SKSW et non pas comme des trilitères ŠNL* et SKW* car il n'y a aucun indice qui permettra d'identifier correctement les formes de mot construites à partir de leur racine d'origine.

Le deuxième cas de figure concerne la variabilité de la réduplication des radicales dans le paradigme lexical ; certaines formes comportent deux radicales identiques, et d'autres une seule, tendue ou non tendue. Examinons quelques exemples : soit la famille lexicale suivante dont l'invariant sémantique est la notion de « mastication » : les verbes *feẓẓ, feẓẓi, feẓẓa, feẓẓi, tefẓaẓ, ttufeẓẓi, ttufeẓẓa, ttufeẓẓi*, les par-

ticipes : *ifeẓẓen, ifeẓẓin, itefẓaẓn, tefẓaẓnin, ittufeẓẓen, ittufeẓẓin, ittufẓaẓn, ttufẓaẓnin*, les noms *afẓaẓ, ifẓaẓn, tuffiṭ, tuffaẓ, uffaẓ, uffaẓn*. À examiner ces différentes unités lexicales, l'on constate que la deuxième consonne radicale ẓ est réduplicquée dans quelques formes et elle est unique, tendue (*feẓẓ*) ou non tendue (*tuffiṭ*), dans d'autres. Pour ces cas, et ils sont nombreux dans le lexique berbère, faut-il relever la racine bilitère Fẓ ou la racine réduplicquée : la trilitère Fẓẓ. Il nous semble que, du point de vue méthodologique et dans un souci de simplification de la classification, il est préférable de reconstruire, pour ces cas, les racines à radicale unique, car la réduplication n'ayant pas de rôle morphologique régi par des règles de grammaire est aléatoire et fluctuante, dépendant de l'arbitraire des signes.

Nous avons examiné, dans ce qui précède, quelques faits formels de la racine en considérant la nature et les combinaisons des radicales. Nous avons proposé quelques options méthodologiques qui doivent présider à la classification par racines en lexicographie berbère. Nos propositions sont essentiellement dictées par le principe d'économie et de simplicité et ne sont pas toutes justifiées par une quelconque théorie de la racine.

Racine : forme et sens

La racine n'est pas cependant une simple forme basique d'un paradigme lexical et la considérer comme telle est sans intérêt pour le lexicographe désireux de fournir des informations sur l'organisation et les structurations morpho-sémantiques de la langue dont il confectionne le dictionnaire. La racine ne sera, dans cette perspective, qu'un simple indicateur d'ordre et de regroupement aberrants. Donnons un exemple pour montrer une telle aberration : les unités lexicales suivantes comportent toutes l'invariant consonantique BD pris comme racine (n'est fourni ici qu'un seul élément, avec son premier sens, de chaque champ morpho-sémantique) :

bedd : « se lever, se dresser »
bdu : « commencer, débuter »
abda : « toujours, tout le temps »
abadu : « canal d'irrigation »
biddi : « obligatoirement, nécessairement »
lebda : « feutre (étoffe) ».

Signalons d'abord que le (L) initial de *lebda* est un article défini de l'arabe que le berbère garde lorsqu'il emprunte à l'arabe, mais ce morphème perd sa fonction en berbère.

Il est tout à fait évident qu'il n'est pas raisonnable, du point de vue lexicographique, de classer les formes de mot données en exemple sous une même racine. Un tel regroupement transgresse la règle sémantique relative aux affinités de sens. En effet, aucune relation sémantique n'est possible à établir entre les différentes formes de mot construites à partir de la séquence commune BD, il y a donc lieu de les dégrouper bien qu'elles soient liées dans leur forme par la même charpente consonantique. On considérera ainsi que nous avons affaire non pas à une racine unique, mais à plusieurs BD_1, BD_2, BD_3, BD_4, BD_5, BD_6, qui sont homophones mais différenciées quant à leur sens général.

Le critère sémantique est donc important pour pouvoir donner à la racine un statut théorique et pratique acceptable en lexicographie berbère et pour qu'elle puisse être utilisée comme paramètre de classification. Le critère sémantique n'est cependant pas facile à circonscrire : selon quelle analyse peut-on différencier sémantiquement des racines homophones ? Autrement, qu'appelle-t-on un sens général ? Est-ce une notion suffisamment large pour intégrer diverses acceptions et nuances sémantiques ? Ou bien est-ce, au contraire, un sens unique ?

De telles questions redoutables n'ont, malheureusement, pas été l'objet de recherche en linguistique berbère. Mais, indépendamment de problèmes de sémantique théorique, le lexicographe est confronté à des options méthodologiques imposées par les lois du genre. Il nous semble qu'il y a deux positions extrêmes qui doivent être évitées si l'on veut que le travail lexicographique ne soit pas en contradiction avec les données de la langue. (cf. Taifi, 1988 : 15-26).

Premièrement la racine ne doit pas être considérée comme un simple ensemble invariable de consonnes commun à une série de formes de mot. Une telle option permettra certainement l'éradication de l'homophonie des racines, mais celles-ci n'auraient plus aucun statut théorique et aucune signification. Elles ne peuvent être utilisées comme principe de classification du lexique, car les regroupements des sous-entrées, c'est-à-dire des formes de mot, sous une racine, seront hétérogènes et ne constitueront plus des champs morpho-sémantiques. Or, un dictionnaire, qui est, sans doute d'abord, un travail sur le lexique, doit nécessairement tenir compte des structurations lexicales de la langue qu'il recense et présenter aux consultants de telles structurations. Définir la racine comme simple forme et l'utiliser comme telle dans la classification lexicographique aboutirait donc à établir des ensembles lexicaux composés d'éléments qui n'ont pas de relations sémantiques.

La seconde option extrême consiste à distinguer une racine pour chaque effet de sens délimité, c'est-à-dire à attribuer aux unités lexicales une stricte monosémie, en neutralisant les caractéristiques polysémiques du lexique. Ainsi, la forme de mot *ifri* ayant quatre significations apparentées : « grotte », « caverne », « gîte » et « terrier » sera classée sous quatre racines différentes FR_1, FR_2, FR_3 et FR_4 puisqu'à chaque sens unique doit correspondre, selon l'option de la monosémie stricte, une racine. Une telle procédure aboutirait, on s'en doute, à l'éclatement et à l'éparpillement du lexique en augmentant excessivement le nombre des racines.

La position intermédiaire, celle adoptée par les quatre dictionnaires précédemment cités, tient compte des deux aspects définitionnels de la racine : la forme et le sens. Mais dans la pratique, les critères de différenciation ne sont pas toujours suffisamment clairs pour permettre une rigoureuse distinction et circonscription des effets de sens. En fait, c'est l'éternel et fameux phénomène de l'opposition homophonie/polysémie qui resurgit à chaque fois, phénomène caractérisant toutes les langues du monde, mais plus épineux pour la lexicographie berbère à cause justement de l'organisation du lexique en racines et schèmes et aussi, dans l'état actuel de la linguistique berbère, de l'indigence des travaux en lexicologie et en sémantique.

Conclusion

Nous avons essayé, dans cette communication, d'exposer un certain nombre de problèmes de méthodologie en lexicographie berbère. Nous avons soutenu, en filigrane, que la dictionnairique berbère ne peut faire l'économie de la racine comme principe de classification, car ce principe est imposé par la morphologie de la langue. La construction des formes de mot se fait en effet par dérivation associative, insérant les racines dans des schèmes. La classification par racines bute, toutefois, sur des difficultés d'application et de procédure. Nous avons proposé des solutions à certaines de ces difficultés, solutions dictées par notre pratique. Mais il en reste d'autres, celles concernant, notamment, l'altération des racines et la dispersion des schèmes que nous avons décrites dans des articles précédents (cf. Taifi, 1990a : 219-232 et Taifi, 1990b : 92-110). L'objectif de cette contribution est de participer à l'élaboration de la métalexicographie berbère.

Informatisation du *Dictionnaire explicatif et combinatoire* : le projet NADIA-DEC

Gilles SÉRASSET

GETA-CLIPS-IMAG (UJF & CNRS), Grenoble, France

Présentation du projet

Le projet NADIA-DEC est basé sur les travaux de définition d'un système universel de bases de données lexicales multilingues au laboratoire GETA-CLIPS et sur les travaux de définition du *Dictionnaire explicatif et combinatoire du français contemporain* au laboratoire GRESLET.

Ce projet vise la création d'une version informatisée du *Dictionnaire explicatif et combinatoire du français contemporain* (DEC). Cette version devra contenir l'ensemble des informations présentes dans le DEC sous une forme aussi structurée que possible. Elle s'appuie donc sur le système de gestion de bases lexicales multilingues SUBLIM défini au GETA-CLIPS.

Ce projet répond à plusieurs motivations de la part de chacun des partenaires. D'une part, il permet de tester le système SUBLIM en l'utilisant pour un dictionnaire mettant en œuvre des structures complexes. D'autre part, les informations contenues dans le DEC présentent une richesse que l'on ne trouve dans aucun autre dictionnaire informatisé. Enfin, la mise en œuvre de ce projet suppose la création d'outils informatiques simplifiant la gestion d'un tel dictionnaire.

Pour atteindre ces différents objectifs, nous souhaitons non seulement informatiser une version du DEC, mais aussi informatiser sa chaîne de production afin de faire en sorte que le DEC existe d'abord sous forme informatique puis sous une forme imprimable. La méthodologie du projet est donnée par la figure 1.

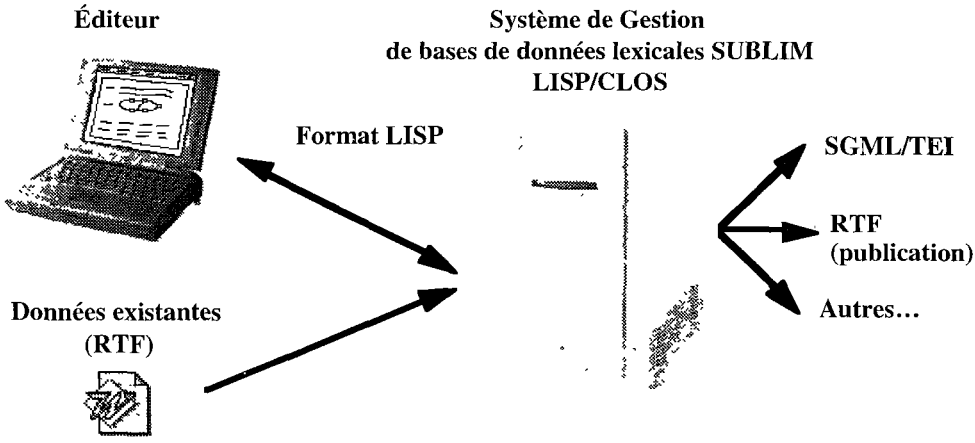


FIGURE 1 . Méthodologie du projet NADIA-DEC.

Ainsi, le travail à effectuer se décompose en différentes étapes :

- création d'un éditeur informatique spécialisé pour le DEC ;
- création (en utilisant le système SUBLIM) d'un système de gestion des données lexicales. Les vérifications des différentes contraintes sur les données seront vérifiées à ce niveau ;
- réalisation d'un mécanisme d'import des données existantes, actuellement au format RTF (*Rich Text Format*) ;
- réalisation d'un module d'export vers différents formats utilisables informatiquement (SGML/TEI, format LISP...) et pour la publication papier (RTF, MIF...).

Cet article présente la structure informatique utilisée à la fois pour l'éditeur et pour le système central de gestion des données lexicales.

Le projet SUBLIM

L'étude détaillée des travaux menés dans le cadre de la construction de systèmes de bases lexicales multilingues ou monolingues informatisées nous a permis de faire les constatations suivantes :

- de nombreux projets existent, qui présentent des particularités intéressantes. Il est donc vain de proposer un nouveau système qui ne puisse prendre en charge les méthodes de ces divers projets ;
- les projets multilingues existants figent en général l'approche lexicale que l'on peut utiliser pour réaliser le multilinguisme. Ainsi, le projet Esprit Multilex (par exemple) ne permet l'utilisation que d'une approche bilingue et ne peut donc satisfaire les lexicologues tentés par des approches interlingues ;
- les projets existants offrent parfois au lexicologue la possibilité de personnaliser les structures linguistiques utilisées dans le lexique, mais de manière insuffisante. Ainsi, si, dans le projet MULTILEX, il est possible de définir ses

propres traits linguistiques, il est impossible de s'affranchir de l'utilisation d'une structure de traits typés pour coder les informations lexicales.

Partant de ces constatations, Sérasset (1994) propose un système informatique de gestion de bases lexicales multilingues qui réponde à ces différentes critiques. Ainsi, ce système peut satisfaire les lexicologues souhaitant coder des informations lexicales complexes utilisant des structures informatiques diverses (arbres, automates, graphes, structures de traits, ensembles...) dans le cadre de bases lexicales monolingues ou multilingues, basées sur les approches par dictionnaires de transfert ou par dictionnaires interlingues.

Pour atteindre cet objectif, le système SUBLIM demande au lexicographe de définir la structure linguistique du dictionnaire qu'il souhaite informatiser par deux langages ayant des objectifs différents.

- Dalex (Définition de l'Architecture LEXicale) permet de définir les dictionnaires mis en œuvre dans la base lexicale et leurs liens. Il permet, entre autres, de définir l'approche lexicale utilisée (interlingue ou par transfert).
- Darling (Définition de l'ARchitecture LINGuistique) permet, pour chaque dictionnaire, de définir les structures informatiques utilisées dans les entrées. Le lexicographe peut utiliser de nombreuses structures prédéfinies comme les arbres, graphes, automates, structures de traits...

Nous ne donnerons pas dans cet article plus de détails sur ces langages. Le lecteur en trouvera néanmoins des exemples d'utilisation dans la suite de cet article.

Le *Dictionnaire explicatif et combinatoire du français contemporain*

Une unité de ce dictionnaire (lexème) est un sens de mot ou de locution. Cette unité lexicale est associée à une unité morphologique, à une définition, à d'éventuelles connotations, à un régime, à des exemples, et à des fonctions lexico-sémantiques. Nous lui affectons de plus un numéro de sens qui l'identifie parmi les différents sens d'une entrée.

Un lexème peut aisément être codé comme une structure de traits :

```
(def-linguistic-class lexème
  (feature-structure
    ((numéro           numéro)
     (UMorph           UMorph)
     (définition       définition)
     (connotations     connotations)
     (régime           régime)
     (exemples         exemples)
     (lexico-sem-fns   lex-sem-fns))
))
```

Une unité morphologique comprend une forme graphique et des informations morphologiques. Elle peut être reliée à plusieurs lexèmes qui lui sont associés de manière arborescente :

CŒUR, nom, masc.

I.1a.	Organe principal de la circulation sanguine d'une personne... [*le cœur de* ***Jean***]
1b.	Organe principal de la circulation sanguine d'un animal... [*le cœur de lion*]
2.	Produit alimentaire ... [*le cœur de veau*]
3.	Partie de la poitrine d'une personne ... [*Il a serré son fils sur son cœur*]
4a.	Organe imaginaire des sentiments ... [*Le cœur espère toujours*]
4b.	Organe imaginaire de l'intuition ... [*Son cœur le lui dit*]
5a.	... propriété de la personnalité ... [*un cœur de glace*]
5b.	Personne possédant le cœur I.5a [*Vous devez la vie à un noble cœur, à un homme vaillant*]
II.1a.	Partie principale d'une unité fonctionnelle... [*le cœur du bateau*]
1b.	Élément principal [*le cœur du problème*]
2a.	Partie centrale d'un espace... [*le cœur du royaume*]
3.	Objet... ayant la forme du cœur I.1a [*un cœur en papier*]
4.	Une des quatre couleurs 2 des cartes à jouer... [*l'as de cœur*]
III.	Organe imaginaire des nausées ... [*Cette senteur lui tournait le cœur*]

Aussi, nous définirons une unité morphologique comme un arbre portant des informations morphologiques à la racine et des lexèmes sur les feuilles.

```
(def-linguistic-class UMorph
  (tree :root     Morphological-information
        :leaves lexème))
```

L'information morphologique associée à la racine de cet arbre ne comporte qu'une graphie, une catégorie, un genre et un nombre.

```
(def-linguistic-class Morphological-information
  (feature-structure
     ((graph       string)
      (catégorie   cat)
      (genre       gnr)
      (nombre      nbr))))

(def-linguistic-class cat
  (one-of (nom verbe adjectif adverbe... )))
(def-linguistic-class gnr
  (one-of (masculin féminin)))
(def-linguistic-class nbr
  (one-of (singulier pluriel)))
```

Une définition du DEC n'est pas une simple chaîne de caractères :

I.1a. *Cœur de X* = Organe principal de la circulation sanguine d'une personne X qui se trouve dans la partie centrale du corps II.1d de X et qu'on représente symboliquement comme ayant la forme ♡.

Mis à part le fait que l'on y trouve une image, on peut remarquer qu'elle se compose de deux parties principales. La première (indiquée en italiques) présente un usage du lexème dans une locution où les différents arguments du prédicat représenté sont indiqués sous forme de variable. La seconde est une explicitation du sens du lexème, elle réutilise les variables de la première partie. On remarque aussi qu'elle fait référence à des lexèmes définis par ailleurs dans le dictionnaire (corps II.1d).

Nous simplifierons cette structure en la décomposant simplement en deux chaînes de caractères, l'une contenant la forme du prédicat, l'autre contenant sa définition :

```
(def-linguistic-class définition
  (feature-structure
    ((prédicat  string)
     (explicite string))))
```

Après cette partie de définition, on trouve éventuellement une partie consacrée aux connotations :

Connotations

1) Cœur I.1a est le siège des sentiments [voir CŒUR I.4a].

2) Cœur I.1a est le siège de l'intuition [voir CŒUR I.4b].

3) Cœur I.1a qui bat 1 représente la vie [voir les phrasèmes correspondants dans CŒUR I.1a].

Cette partie se présente comme une liste de connotations. Chacune est donnée sous forme de chaîne de caractères faisant référence à au moins un lexème. Il est donc intéressant, dans une version informatique de ce dictionnaire, de conserver à la fois la connotation sous forme de chaîne de caractères et sous forme d'un ensemble de liens vers d'autres lexèmes :

```
(def-linguistic-class connotations
  (set-of connotation))
(def-linguistic-class connotation
  (feature-structure
    ((texte    string)
     (réfère-à (set-of ((link :target lexème)))))))
```

À la suite de ces éventuelles connotations, on trouve le régime du prédicat. Ce régime donne les informations sur les différentes réalisations syntaxiques des arguments du prédicat. Le régime est donné sous forme de tableau dont les colonnes correspondent aux arguments et les lignes aux différentes réalisations. Certaines combinaisons ainsi établies étant non valides, on en reprend ensuite la liste, en indiquant leur impossibilité. On reprend aussi un certain nombre de ces combinaisons pour en donner des exemples (l'exemple suivant est tiré de l'entrée « enseigner 1 ») :

1. *X enseigne Y à Z* = X, censé avoir la qualification professionnelle dans le domaine Y, cause que Z apprenne III.1b Y en transmettant, méthodiquement et dans un cadre officiel, à Z des connaissances (portant sur) Y ou des techniques (portant sur) Y [≅ $CausConv_{21}$(*apprendre III.1b*)].

Régime

1 = X	2 = Y	3 = Z
1. N	1. N 2. *à* V_{inf}	1. *à* N 2. **rare** N

1) $C_{2\,2}$ sans $C_{3\,1}$
2) $C_2 + C_{3\,2}$ } : *impossible*

$C_1 + C_2$: *Pierre enseigne la grammaire <la couture >/ à faire cela*

$C_2 + C_2 + C_3$: *Pierre enseigne la grammaire à ses élèves*

La structure correspondant à cette partie est beaucoup plus compliquée que celle des parties précédentes. En effet, cette présentation n'est que le reflet, imprimable, d'une structure complexe où l'on retrouve l'ensemble des combinaisons possibles de réalisations d'arguments. On peut donc représenter cette partie de deux manières :

- en restant proche de sa forme papier. On a alors un tableau et une liste des combinaisons impossibles ;
- en représentant cette structure de manière plus abstraite. On peut ainsi la représenter par un automate dont chaque chemin forme une combinaison valide.

Si l'on choisit la seconde solution, le régime donné en exemple sera représenté par l'automate donné en figure 2.

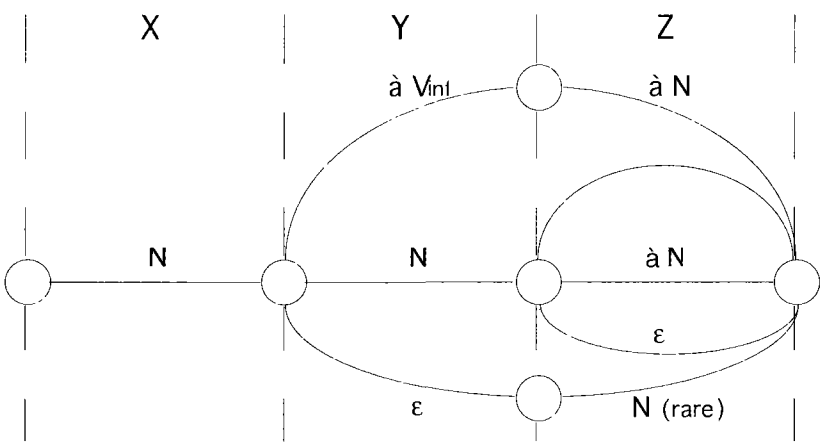


FIGURE 2 : Régime d'enseigner 1, sous forme d'automate.

Pour exprimer cette solution, nous utiliserons la structure logique d'automate prédéfinie dans le système SUBLIM où l'on peut définir des contraintes sur les classes

acceptables en décoration des différents éléments de l'automate (arcs, nœuds ou états, états initiaux, états finals).

Ainsi, la structure du régime s'exprimera sous forme d'une structure de traits regroupant l'automate des régimes, l'ordre d'apparition des arguments et l'ensemble des exemples :

```
(def-linguistic-class régime
  (feature-structure
    ((automate       automate-régime)
     (argument-order (list-of (string)))
     (exemples       exemples-régime))))
(def-linguistic-class automate-régime
  (automaton :arcs réalisation-argument))
(def-linguistic-class exemples-régime
  (set-of ((feature-structure
            ((réalisations (list-of (string)))
             (exemple      string))))))
```

La partie la plus importante de ce dictionnaire réside dans l'ensemble des fonctions lexicales du lexème. Leur meilleur définition est donnée, en première partie du DEC, par l'auteur, Igor Mel'čuk :

> Les fonctions lexicales (FL) présentent l'ensemble de la cooccurrence lexicale restreinte intéressant le lexème considéré. Elles constituent une innovation lexicographique qui permet de décrire d'une façon systématique un vaste ensemble de locutions plus ou moins figées qui ne sont quand même pas des expressions idiomatiques *stricto sensu*. Il s'agit, par exemple, des locutions comme une FERME intention, une résistance ACHARNÉE, un argument DE POIDS, un bruit INFERNAL, un désir ARDENT, une envie FOLLE, une règle STRICTE, une vérité INCONTESTABLE, où des adjectifs bien spécifiques doivent être employés avec les différents noms pour exprimer la même idée d'intensification. Comme autre exemple de locution de ce type, on peut citer les expressions DONNER une leçon, FAIRE un pas, COMMETTRE un crime, PORTER une accusation, etc., où des verbes sémantiquement vides (ou presque vides) différents doivent être choisis en fonction du nom d'action pour lier le nom d'agent en tant que sujet grammatical au nom d'action en tant que complément d'objet direct.

L'écriture générale d'une FL est de la forme : **f**(X) = Y, où **f** est la FL, X est son argument (un lexème ou bien une locution), et Y est la valeur de la FL **f** pour cet argument, c'est-à-dire l'ensemble des expressions linguistiques qui peuvent exprimer le sens ou le rôle syntaxique donné (noté par **f**) auprès de X.

Comme ce dictionnaire est imprimé, les expressions linguistiques sont données sous une forme linéaire :

MÉPRIS, nom, masc.

I. Attitude émotionnelle défavorable... [*le mépris pour ce corrupteur*] [...]

Fonctions lexicales

$Caus_3Func_1$	: engendrer [ART ~ chez N] [*La familiarité engendre le mépris*]
$Caus_{(3)}Func_1$	: apprendre, inculquer [ART ~ à N] [*Jean inculque à ses étudiants le mépris de l'hypocrisie; Son attitude partiale envers ses employés apprend à ces derniers le mépris de leur chef*]
$Caus_{(2/3)}Func_1$	: inspirer [ART ~ à N] [*Cet événement inspire aux travailleurs le mépris de leur patron; L'argent inspirait à ce philosophe un tel mépris qu'il a donné son héritage à son frère; L'hypocrisie de Jean leur inspirait un profond mépris*]

Mais la structure interne de ces expressions linguistiques est un arbre syntaxique donnant la construction de ces expressions linguistiques et de l'argument X pour réaliser la fonction **f**.

Ainsi, la structure interne de $Caus_3Func_1$(Mépris I) est l'arbre donné ci-dessous :

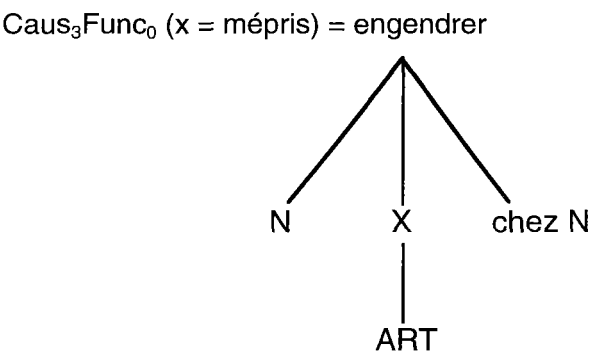


FIGURE 3 : Structure interne d'une expression linguistique, valeur de fonction lexicale.

Une fonction lexicale représente donc un lien entre un lexème et une expression linguistique complexe comportant d'autres lexèmes. Aussi, la valeur de ces fonctions lexicales peut être représentée comme un ensemble d'arbres dont certains nœuds sont des variables, et d'autres sont des lexèmes.

Il faut aussi représenter les fonctions lexicales. En effet, s'il y a un nombre limité de fonctions lexicales de base, on trouvera des fonctions composées dans les différents articles de dictionnaire.

Prenons un exemple : les fonctions **Oper**$_1$, **Oper**$_2$... ont pour valeur les verbes sémantiquement vides qui prennent le nom du premier, deuxième... actant comme sujet grammatical et C_0 (leur argument) comme complément d'objet principal :

Oper$_1$*(attention)* = *faire*
Oper$_2$*(attention)* = *attirer*
Oper$_1$*(conseil)* = *donner*
Oper$_2$*(conseil)* = *recevoir*
Oper$_1$*(aide)* = *prêter, accorder*
Oper$_2$*(aide)* = *recevoir*

La fonction **Caus** représente la notion : « faire en sorte que quelque chose ait lieu ». Elle s'emploie le plus souvent en combinaison avec d'autres fonctions. Ainsi, si **Oper**$_1$*(désespoir)* = *éprouver, ressentir, avoir*, **CausOper**$_1$*(désespoir)* représente « faire en sorte que quelqu'un éprouve du désespoir ». Donc **CausOper**$_1$*(désespoir)* = *pousser, réduire* [*qqn au désespoir*], *jeter* [*qqn dans le désespoir*], *frapper* [*qqn de désespoir*].

Il n'est donc pas possible de représenter chaque fonction lexicale comme un attribut dans une structure, puisque la possibilité de composition entraîne toute une combinatoire des fonctions lexicales. Nous les représenterons donc par la structure logique de base *function* prédéfinie dans le système SUBLIM, où l'on peut contraindre les classes acceptables pour les différents éléments de fonction (nom, arguments, valeur). Ainsi, la structure correspondant aux fonctions lexicales peut s'exprimer comme suit :

```
(def-linguistic-class lex-sem-fns
  (set-of (lex-sem-fn)))
(def-linguistic-class lex-sem-fn
  (function   :label nom-FL
              :arguments FL-arg
              :value expression-linguistique))
```

Pour représenter la composition de fonctions, on peut autoriser l'utilisation d'une fonction lexicale en argument d'une fonction lexicale. Néanmoins, la valeur de la fonction intermédiaire (si elle existe) n'est pas nécessairement pertinente. Seuls les noms des fonctions composées sont porteurs d'information. Aussi, le plus simple est d'autoriser une valeur complexe comme nom de la fonction. Nous définirons donc un label de fonction comme une liste (ordonnée) de noms de fonctions de base.

```
(def-linguistic-class nom-FL
  (list-of (nom-FL-base)))
```

Le nom d'une fonction de base est donné par un identificateur de la fonction (une chaîne de caractères) et par un éventuel numéro de l'actant sur lequel elle opère :

```
(def-linguistic-class nom-FL-base
  (feature-structure
    ((fonction string)
     (actant   integer))))
```

L'argument de la fonction est un lexème. Le fait d'indiquer cet argument est re-

dondant puisque cette fonction est définie à l'intérieur de la structure du lexème (il sera néanmoins représenté afin de simplifier les mécanismes d'extraction d'information).

```
(def-linguistic-class FL-arg lexème)
```

L'expression linguistique valeur de la fonction est représentée sous forme d'arbre (comme nous l'avons indiqué plus haut). Les nœuds de cet arbre sont soit des lexèmes, soit des variables. Pour simplifier, nous noterons les variables comme des chaînes de caractères.

```
(def-linguistic-class expression-linguistique
  (tree :nodes (one-of (lexème string))))
```

La définition de la structure linguistique du DEC, même simplifiée, illustre parfaitement le besoin ressenti par les linguistes de pouvoir mélanger différentes structures logiques dans une seule et même structure linguistique. Le fait de proposer différentes structures logiques permet au linguiste de manipuler des concepts proches de ceux utilisés dans sa théorie. Cela permet de simplifier le travail du linguiste en lui permettant de rester à un niveau d'abstraction très utile lorsqu'il souhaite implémenter une théorie complexe.

Conclusion et perspectives

L'étape de formalisation de la structure du DEC est nécessaire pour son informatisation. On remarque que ce dictionnaire met en œuvre de nombreuses structures informatiques de base (arbres, listes, structures de traits, etc.) et nécessite donc un environnement permettant de gérer un tel ensemble de structures hétérogènes. Cette constatation nous conforte dans l'idée que les futurs systèmes de gestion de bases lexicales devront être à même de gérer de telles structures.

Le projet NADIA-DEC va permettre différentes études :

- multilinguisation du DEC : le DEC est un ensemble de dictionnaires monolingues, mais il se base sur une théorie qui s'applique à l'ensemble des langues étudiées jusqu'alors. Il semble donc souhaitable de mettre en correspondance les différents lexèmes des différentes langues disponibles, afin d'étudier les problèmes éventuels d'une multilinguisation du DEC ;
- vérification automatique de cohérence : pour la sortie de chaque volume du DEC, les auteurs doivent vérifier de nombreuses contraintes de cohérence sur les différents articles. Certaines de ces contraintes (comme la vérification de la numérotation des lexèmes, de certaines contraintes de formatage...) peuvent être prises en charge par le gestionnaire de bases lexicales. Nous étudierons les contraintes à vérifier et un environnement permettant de les spécifier et de les vérifier ;
- récupération des informations existantes : l'ensemble des informations déjà publiées n'est disponible que sous forme de fichiers de traitement de texte. Nous souhaitons étudier les problèmes d'import de ces entrées existantes, en les généralisant afin de définir un environnement d'import de différentes sources existantes dans une base lexicale.

En résumé, le projet NADIA-DEC ne vise pas seulement le développement d'une version informatique du DEC, mais permet aussi, de mettre en pratique, sur un cas réel et exigeant les idées de généricité développées au GETA dans le cadre du système SUBLIM. Il permettra de plus de développer d'autres aspects de recherche dans le domaine de la lexicographie (multilinguisation du DEC) et dans le domaine de l'informatique linguistique (développement d'environnements de manipulation de données linguistiques, etc.).

Pour terminer, je tiens à remercier l'AUPELF-UREF pour son soutien à cette action de recherche partagée entre le GETA-CLIPS (institut IMAG, Université Joseph Fourier - Grenoble I) et le GRESLET (Département de linguistique et de traduction, Université de Montréal).

La formalisation des collocations pour le traitement automatique du langage naturel : le modèle des fonctions lexicales syntagmatiques[1]

Agnès TUTIN

URA SILEX, Université de Lille III, Villeneuve d'Ascq, France

Introduction

Le modèle des Fonctions Lexicales, qui appartient au *Dictionnaire explicatif et combinatoire* (DEC), constitue un formalisme séduisant pour le traitement des collocations :

- Les Fonctions Lexicales (FLs désormais) font partie d'une théorie linguistique complète dont le DEC est un composant essentiel (Mel'čuk, 1981; Mel'čuk et Polguère, 1987). La théorie Sens-Texte apparaît bien adaptée au traitement automatique du langage et les FLs ont déjà été exploitées pour des applications dans le domaine de la traduction assistée par ordinateur (Heylen et Maxwell, 1994) et en génération de textes (Iordanskaja, Kim et Polguère, 1995).

- Le modèle des FLs fournit une formalisation à la fois sémantique et syntaxique : les collocations sont décrites par le biais d'une étiquette syntactico-sémantique, par exemple la FL **Magn** signifie ‹très›, ‹intensément› et elle produit des modifieurs (des adjectifs ou des adverbes selon la partie du discours du mot d'entrée).

- Le modèle est « génératif » : les FLs sont combinables (par exemple, **AntiMagn** ‹peu intensément› est construite à partir de **Anti** et de **Magn**), même si la syntaxe et la sémantique des règles de combinaison apparaissent parfois confuses.

- Le modèle a déjà été utilisé sur un très large échantillon de données : les trois volumes du *DECFC* (Mel'čuk *et al.*,1984, 1988, 1992).

1. Un grand merci à Alain Polguère qui a relu attentivement cette version et m'a bien entendu signalé quelques incohérences.

Dans cette étude, nous essayons d'approfondir le formalisme des Fonctions Lexicales syntagmatiques, dans la perspective d'adapter ce modèle à des applications de traitement automatique en français, complétant ainsi le défrichage entamé par Alonso Ramos (1993). Pour cette tâche, nous nous basons sur les trois volumes du *Dictionnaire explicatif et combinatoire du français contemporain* (Mel'čuk *et al.*, 1984, 1988, 1992), le DEC français, qui constituent déjà une riche base de recherche. Notre étude a pour objet de déterminer dans quelle mesure le formalisme actuel doit être remanié et approfondi et elle vise à proposer des modes de description adaptés pour le TALN.

1. Les fonctions lexicales syntagmatiques du DEC

1.1. Une brève définition de la notion de collocation[2]

Conformément aux perspectives du DEC, nous définissons une collocation comme un **syntagme semi-figé** ou semi-phrasème (par exemple, *gros fumeur, prendre un bain,...*) dans lequel le sens d'une unité lexicale A (la base) est transparent, alors que le sens de l'autre unité lexicale B (le collocatif) est déterminé par la présence de A. Les collocations s'opposent, d'une part, aux phrasèmes (par exemple, *prendre le taureau par les cornes*) pour lesquelles le sens du tout n'est pas déductible (ou difficilement) du sens des parties, et d'autre part, aux syntagmes standard (*un chanteur gros, un taureau méchant*) dont l'interprétation découle du sens des composants. En d'autres termes, la façon dont un mot est lexicalisé (le *collocatif*) dépend du mot auquel il est rattaché sémantiquement (la *base*). Ainsi, la lexie utilisée pour exprimer l'intensité dans le contexte de *fumeur* est *gros*. Dans cette approche, la collocation n'est pas définie par une fréquence statistique ; la récurrence de certains couples de mots est simplement une conséquence du phénomène sémantique de la cooccurrence restreinte[3].

1.2. Différents types de Fonctions Lexicales

Les collocations ont été formellement décrites d'un point de vue syntaxique et sémantique au moyen des FLs et plus exactement, par le biais des FLs syntagmatiques. Une FL est une fonction, au sens mathématique, qui associe un ensemble d'unités lexicales à une unité lexicale. Dans le cas des FLs syntagmatiques, le mot-clé peut être considéré comme la base, et la(les) valeur(s), le(s) collocatif(s). Les FLs les plus productives et les plus universelles ont été étiquetées et une cinquantaine de FLs standard a été proposée. Par exemple, la FL **Magn** (sens d'intensité) est appliquée à *fumeur* pour produire *gros*, alors que la FL $\mathbf{Oper_1}$ est appliquée à *bain* pour produire *prendre*. Cela est codé de la façon suivante :

Magn(*fumeur*) = *gros*
$\mathbf{Oper_1}$(*bain*) = *prendre*

2 Pour une discussion approfondie sur la notion de collocation, voir Heid (1994).
3. Cf. Smadja et McKeown (1991).

Toutes les FLs ne sont pas utilisées pour décrire des relations de collocation. En utilisant des critères de cooccurrence et des critères sémantiques, trois types de FLs peuvent être dégagés :

- **les FLs syntagmatiques** comme **Magn** ou $\mathbf{Oper_i}$ sont utilisées pour formaliser les collocations ;

- **les FLs paradigmatiques** sont utilisées pour associer une unité lexicale qui partage avec d'autres unités lexicales un composant de sens non trivial. La(es) valeur(s) et le mot-clé ne forment habituellement pas un syntagme. Ainsi, $\mathbf{S_{loc}}$ (nom typique pour le lieu) et $\mathbf{S_3}$ (nom typique pour le troisième actant) sont des relations paradigmatiques : $\mathbf{S_{loc}}$(*patiner*) = *patinoire* et $\mathbf{S_3}$(*conférence*) = *assistance, audience* ;

- **les FLs mixtes** partagent les caractéristiques des deux types précédents[4]. Elles sont utilisées pour associer un mot-clé avec un ensemble de lexèmes qui partagent une composante sémantique, mais la valeur et le mot-clé peuvent aussi se combiner pour former un syntagme, par exemple **Mult**(*bateau*) = *flotte* ('nom standard pour un ensemble de') ou **Cap**(*lycée*) = *censeur* ('la tête de').

1.3. Les combinaisons de FLs

De plus, les FLs peuvent être combinées. Trois types de combinaisons peuvent être réalisées.

- **Les FLs composées** sont combinées de façon similaire aux compositions de fonctions mathématiques. La valeur finale est produite par le biais des valeurs intermédiaires des fonctions intermédiaires. Par exemple, pour calculer la valeur de $\mathbf{S_0}$(**Gener**(*étuver*)), on cherchera d'abord la valeur de **Gener**(*étuver*) = *cuire*, puis de $\mathbf{S_0}$(*cuire*) = *cuisson*. La composition est bien entendu non commutative[5].

  Pour l'étude des collocations, les FLs composées ne paraissent pas pertinentes[6]. Tout d'abord, elles sont avant tout utilisées pour décrire des fonctions paradigmatiques. Par ailleurs, la question de l'encodage des FLs composées se résout d'elle-même, puisque les compositions peuvent être calculées de façon régulière à partir des FLs simples.

- **Les configurations de FLs** ont un mot-clé en commun. Les FLs sont liées par un signe «+ », qui est ici commutatif.

  Par exemple, **Magn + Func**$_0$(*tempête*) = *faire rage* signifie que *faire rage* est la valeur quand *tempête* est le sujet grammatical et que la tempête est intense.

---

4. Plus exactement, les FLs considérées comme mixtes sont en fait des FLs pour lesquelles dans le cas de certaines valeurs (*essaim* pour **Mult**(*abeille*), par exemple), on peut avoir une cooccurrence facultative du mot-clé et de la valeur. Cela n'est pas le cas pour l'ensemble des valeurs produites pour une fonction mixte donnée, mais pour un sous-ensemble restreint. De ce fait, la notion de fonction mixte n'est pas opérationnelle car elle conduit à des interprétations erronées dans le décodage. Dans un contexte donné, il faut en effet pouvoir décider de façon claire si l'interprétation de la FL est syntagmatique ou paradigmatique.

5. **Gener**($\mathbf{S_0}$(*étuver*)) ne produit aucune valeur parce que $\mathbf{S_0}$(*étuver*) ne produit lui-même pas de valeur.

6. Notre intention n'est pas de dire qu'elles ne sont d'aucune utilité pour le TALN. Au contraire, nous sommes persuadée qu'elles peuvent être utilisées avec profit pour générer des expressions de référence appropriées (Alonso Ramos, Tutin et Lapalme, 1995).

Les configurations devraient aussi être prises en compte pour les relations de collocations, mais elles apparaissent relativement rares, et elles ne seront pas davantage examinées ici.

- **Les fonctions complexes** sont des combinaisons qui ne peuvent pas être décomposées. La valeur n'est pas obtenue par une combinaison régulière, à la différence des compositions. Par exemple, la valeur de **AntiMagn**(*prix*) [= *modique*] ne peut pas être déduite de **Magn**(*prix*) [= *élevé*]. *Modique* en effet n'est pas dans l'absolu un antonyme de *élevé*, mais seulement lorsqu'il apparaît en cooccurrence avec *prix*. Quand une combinaison est syntagmatique, il s'agit généralement d'une FL complexe, alors que les fonctions complexes peuvent elles aussi être paradigmatiques.

  Une fonction complexe n'est donc pas une composition : en fait, une FL complexe est une FL résultant elle-même d'une combinaison. Dans l'exemple de **AntiMagn, Magn** peut être considérée comme une fonction, alors que **Anti** peut être considérée comme une « fonctionnelle », cette dernière étant définie comme une fonction de second ordre s'appliquant elle-même à des fonctions.

  Nous pouvons ainsi définir une **fonctionnelle lexicale** comme une fonction qui s'applique à une fonction simple ou complexe pour produire une fonction complexe (le mécanisme est récursif). La FL complexe s'applique elle-même à une unité lexicale pour produire des unités lexicales.

  De nombreuses fonctions sont à la fois des FLs et des fonctionnelles lexicales, mais lorsque ce cas se produit, elles ne partagent pas nécessairement exactement les mêmes caractéristiques. Par exemple, **Magn** est une FL syntagmatique simple (**Magn**(*pluie*) = *torrentielle*) et est aussi une fonctionnelle utilisée pour produire des FLs complexes où elle exprime l'intensification de la fonction sur laquelle elle porte (**Magn** signifie 'très, intensément', alors que **MagnManif** signifie 'se manifester avec intensité' et doit être considéré comme un tout[7]).

Pour l'implémentation des collocations en TALN, la sémantique et la syntaxe des fonctions complexes doivent être définies de façon détaillée. À l'heure actuelle, un certain flou règne chez les rédacteurs du DEC, dépositaires cependant d'une expertise précieuse en matière de FLs. La méconnaissance des heuristiques employées et des règles de combinaison constitue une entrave à la formalisation et l'implémentation des FLs en TALN.

Une première étape dans cette direction serait d'établir une description détaillée des FLs simples et des fonctionnelles. Cette tâche devrait permettre dans un second temps d'élaborer une grammaire formelle des FLs complexes.

2. Une description détaillée des FLs syntagmatiques : les FLs simples et les fonctionnelles utilisées pour construire les FLs syntagmatiques

2.1. Les FLs syntagmatiques simples

2.1.1. Un inventaire des FLs syntagmatiques

Tout d'abord, un inventaire des FLs syntagmatiques doit être constitué, tâche qui

7. On pourra noter que la nature syntagmatique de la FL simple **Magn** n'influe pas ici dans la combinaison.

peut être effectuée à partir des trois volumes du DECFC. Ce travail, qui peut paraître absolument trivial, s'est en fait révélé plus complexe que prévu.

La nature paradigmatique ou syntagmatique de la FL peut habituellement être déduite des définitions données dans le DEC (par exemple, le DEC 2 : 91-94) et des exemples. Ainsi, **Oper**$_i$ et **Magn** sont clairement définies comme syntagmatiques :

- « **Oper**$_0$, **Oper**$_1$, **Oper**$_2$ ... : verbe sémantiquement vide qui prend le pronom personnel (*il*) ou le nom du premier, deuxième,... actant de la situation C_0 comme son sujet grammatical (SG), et le mot-clé C_0 comme son complément d'objet (CO) principal.

  Oper$_0$(*vent 1.1*) = *faire*
  Oper$_1$(*attention*) = *faire*
  Oper$_2$(*attention*) = *attirer*
  ... » (DEC 2 : 93)

- « **Magn**: ‹très›, ‹intense/intensément›, ‹à un degré élevé›.

  Magn(*mémoire 1.1*) = *prodigieuse, excellente, étonnante, d'éléphant*
  Magn(*remercier*) = *vivement, chaleureusement, de tout cœur*
  ... » (DEC 2 : 92)

Néanmoins, pour un certain nombre de FLs, le statut paradigmatique/syntagmatique demeure peu clair. Par exemple, il est difficile de décider d'après les exemples relevés si **Mult** (‹un ensemble régulier de›) est syntagmatique ou paradigmatique. En fait, nous avons relevé trois cas avec cette FL.

1. La valeur ne peut pas être utilisée sans le mot-clé (en conservant le sens de **Mult**) : **Mult** est ici clairement syntagmatique :

   Mult(*brebis*) = *troupeau*.

   Sans le mot-clé, la valeur a un sens beaucoup plus générique que celui qui est encodé par **Mult**.

2. La valeur ne peut pas apparaître en cooccurrence avec le mot-clé : **Mult** est clairement paradigmatique.

   Mult(*bleu*^{2}) = *bleusaille*.

   Dans ce cas, l'association avec le mot-clé est agrammaticale : **une bleusaille de bleus*.

3. La valeur peut apparaître soit en cooccurrence avec le mot-clé, soit sans celui-ci : **Mult** est une FL mixte.

   Mult(*abeille*) = *essaim*;

   essaim est ici un synonyme plus large pour *essaim d'abeilles*[8].

8. *Essaim* peut être considéré comme ayant pour prototype : *essaim d'abeilles*. Ce n'est pas du tout le cas avec *troupeau* et *troupeau de brebis*.

À l'heure actuelle, la codification pour **Mult** n'est donc pas cohérente, puisque **Mult** encode à la fois des relations paradigmatiques et/ou syntagmatiques. Nous proposons de considérer cette fonction comme une fonction syntagmatique. Nous la noterons de la façon suivante, en utilisant le signe de fusion[9] lorsque la FL est utilisée de façon paradigmatique :

Mult(*brebis*) = *troupeau*
Mult(*bleu*^{2}) = // *bleusaille*
Mult(*abeille*) = *essaim*; // *essaim*

La même ambiguïté apparaît dans le DEC avec quelques FLs comme $\mathbf{A_0}$, $\mathbf{A_1}$, $\mathbf{Adv_i}$, **Gener, Sing.**

2.1.2. Paramètres descriptifs

Pour décrire formellement les FLs syntagmatiques, nous proposons un ensemble de paramètres. Les valeurs obtenues pour ces paramètres varient selon la langue et les exemples fournis dans cette section sont tirés du français[10].

Une FL donnée peut s'appliquer à de nombreuses parties du discours[11]. Puisque la syntaxe de la fonction peut légèrement varier selon ce paramètre (la partie du discours de la valeur varie souvent en fonction de la partie du discours du mot-clé), nous pensons que les FLs doivent être définies selon la partie du discours de la base.

Les paramètres descriptifs engloberont :

- **La relation Syntaxique Profonde**[12] **(RSyntP) apparaissant entre la base et le collocatif** : s'agit-il d'une relation attributive ou actantielle ? Par exemple, **Magn** est une relation attributive (un modifieur), alors que **Mult** est une relation actantielle (la valeur prend la base comme complément de nom).

- Le type de **réalisation de surface pour la valeur** : les parties du discours et le type de constituant généré par la FL. Ces valeurs peuvent être : des lexèmes (a), des phrasèmes (b) ou des constituants libres (c).

  (a) **Magn**(*pluie*) = *torrentielle, diluvienne*
  (b) **Magn**(*chanter*) = *à tue-tête*
  (c) **Magn**(*manger*) = *comme un ogre*

  Il est crucial de distinguer les lexies (lexèmes et phrasèmes) des constituants « libres », car ces derniers ne possèdent pas d'entrée propre dans le DEC[13].

9. Une fonction fusionnée est une fonction paradigmatique dont la valeur est équivalente à l'ensemble mot-clé + valeur La fusion est encodée avec le signe « // »
10 La description proposée ici complète celle qui a déjà été proposée dans Alonso Ramos (1993) et Alonso Ramos et Tutin (1995)
11 Par exemple, **Magn** s'applique à des noms, des verbes, des adjectifs et des adverbes
12. Ce niveau de description est propre à la théorie Sens-Texte (voir par exemple Mel'čuk, 1981).
13. Strictement, comme cela est souligné par Alonso Ramos et Mantha dans ce volume, toutes les lexies correspondant à des collocatifs ne doivent pas faire l'objet d'entrée « sémantique » dans le DEC. Ainsi, *noir* dans *café noir* ne peut pas être considéré comme un collocatif nécessitant une entrée propre : ici, le collocatif *noir* emprunte sa forme et certains aspects sémantiques à l'adjectif *noir* Tout collocatif doit néanmoins faire l'objet d'une entrée « morphologique » qui indiquera le mode de flexion, les caractéristiques distributionnelles, etc.

- Le type d'***indices*** et d'**exposants** que chaque fonction peut avoir, et dans quelle mesure ils sont facultatifs ou obligatoires.

  Parmi les indices, on trouve des indices actantiels (sémantiques ou syntaxiques) qui indiquent l'actant concerné (par exemple, $\mathbf{Oper}_1$, $\mathbf{Magn}_1$), des indices ensemblistes ($\mathbf{Syn}_{\supset}$, $\mathbf{Ver}_{\cap}$, ...), des indices sémantiques ($\mathbf{Magn}_{\text{'conséquence'}}$).

  Parmi les exposants, on relève : des exposants sémantiques ($\mathbf{Magn}^{\text{temp}}$, $\mathbf{Magn}^{\text{quant}}$), et des exposants d'intensité ($\mathbf{Real}^{\text{II}}{}_3$, $\mathbf{Real}^{\text{I}}{}_3$).

Dans le tableau ci-dessous, quelques FLs portant sur les noms sont examinées selon ces paramètres :

FL	**Définition**	**Partie du discours du collocatif**	**Indices et exposants**	**RSyntP**	**Exemples**
Magn	'très', 'intense/ intensément/ à un degré élevé'	- adjectif ou locution adjectivale - syntagme prépositionnel	- indice sémantique facultatif - indice actantiel facultatif - exposant sémantique facultatif	Magn ←--ATTR--X	*attention soutenue* *feu d'enfer* *bruit à crever les tympans*
Epit	modifieur vide sémantiquement	- adjectif ou locution adjectivale - syntagme prépositionnel	Aucun	Epit ←--ATTR--X	*quenotte petite* *numéro appel*
Real	verbe qui signifie 'réaliser', qui prend le mot-clé comme premier complément et le nom du ième actant comme sujet grammatical	- verbe ou locution verbale	- indice actantiel obligatoire - exposant d'intensité facultatif	Real_i --II--→ X	$\text{Real}^{\text{I}}{}_3$ (*conseil I I*) *accepter* $\text{Real}^{\text{II}}{}_3$ (*conseil I I*) *suivre*

TABLEAU 1 · Description de quelques FLs syntagmatiques portant sur des noms

2.2. Les fonctionnelles

Comme on l'a déjà signalé, les fonctionnelles lexicales sont des fonctions qui s'appliquent à des fonctions simples ou complexes pour produire des fonctions complexes (le mécanisme est récursif). Les fonctionnelles peuvent avoir le même nom que quelques FLs simples paradigmatiques ou syntagmatiques, auxquelles elles empruntent leur sens, mais pas leurs caractéristiques de surface, alors que certaines fonctionnelles comme **Caus** n'ont pas de FL simple équivalente. Nous nous intéresserons seulement ici aux combinaisons engendrant des FLs complexes syntagmatiques.

Pour les fonctionnelles, les paramètres suivants devront être examinés.

- Le fait que la fonctionnelle est une **fonctionnelle régente** ou une **fonctionnelle modificatrice**.

  Dans une FL complexe, toutes les FLs n'ont pas le même statut. La **fonc-**

tionnelle régente (voir Alonso Ramos, 1993[14]) détermine habituellement :

1 - la partie du discours auquel la FL complexe s'applique ;
2 - la partie du discours que la FL complexe produit ;
3 - la nature syntagmatique/paradigmatique de la FL complexe selon le statut des FLs auxquelles elle s'applique.

Par exemple, dans la fonction complexe **AntiMagn**, **Anti** ne modifie pas nature de **Magn** qui reste un modifieur. Au contraire, la fonctionnelle **Caus** est une fonction régente parce qu'elle modifie la structure actantielle de la FL affectée.

- L'**ensemble des FLs simples et complexes** auxquelles les FLs peuvent s'appliquer. Cet ensemble de FLs peut lui-même être un ensemble de sous-ensembles. Par exemple, on utilisera le sous-ensemble des verbes supports (**Oper$_i$, Func$_i$, Labor$_{ij}$**).

- Les **types d'indices et d'exposants** que les fonctionnelles peuvent prendre. Leur usage peut ici être assez différent de l'usage que l'on repère sur les FLs simples. Par exemple, **Magn**, comme fonctionnelle, ne semble pas prendre d'indices actantiels.

- Le **type de fonction complexe** (nominale, verbale, adjectivale) produite par la fonctionnelle.

Dans le tableau suivant, nous décrivons quelques fonctionnelles à l'aide des paramètres proposés plus haut.

14. La notion de fonction régente est différente chez Alonso Ramos (1993) qui n'effectue pas de distinction entre fonction et fonctionnelle. Pour elle, la fonction régente est la fonction qui impose à la combinaison sa signification et/ou sa valeur syntaxique. Ainsi, dans **AntiMagn, Magn** est la fonction régente, dans **S$_0$Oper$_1$, S$_0$** est la fonction régente.

Notre objectif étant de produire, entre autres, des règles de réécriture, nous n'examinons que la fonction qui apparaît à gauche de la combinaison, le mécanisme de réécriture étant bien entendu récursif.

Fonctionnelle	Définition	Type de fonctionnelle	FLs sur lesquelles porte la fonctionnelle	Indices et exposants	Exemples
Incep	'Commencer à'	Fonctionnelle modificatrice qui n'affecte pas les actants de la valeur produite par la FL sur laquelle porte la fonctionnelle.	- Verbes supports : **Func$_i$**, **Oper$_i$**, **Labor$_{ij}$**. - Verbes de réalisation: **Fact$_i$**, **Real$_i$**, **Labreal$_{ij}$**. - Verbes d'expression : **Involv**, **Manif**, **Degrad**, **Son**, **Excess**. - Verbalisateur : **Pred** - FLs complexes verbales.	Aucun	**IncepOper$_1$** *(alphabétisation)* = *entreprendre* **IncepReal$_3$** *(école I.1a)* = *entrer* **IncepInvolv** *(fureur 3b)* = *se déchaîner* **IncepPredPlus** *(chagrin I 1)* = *s'accroître, augmenter, grandir* **IncepProxOper$_1$** *(objection 1)* = *concevoir*
S_0	Nominalisation	Fonctionnelle régente. Prend une FL verbale simple ou complexe en entrée.	- Verbes supports : **Func$_i$**, **Oper$_i$**, **Labor$_{ij}$**. - Verbes de réalisation: **Fact$_i$**, **Real$_i$**, **Labreal$_{ij}$**. - Verbes d'expression: **Involv**, **Manif**, **Degrad**, **Son**, **Excess**. - Verbalisateur : **Pred**. - FLs complexes verbales	Aucun	**S$_0$Oper$_2$** *(appel téléphonique)* = *réception* **S$_0$Real$_2$** *(numéro de téléphone)* = *composition* **S$_0$Son**(*tempête 1*) = *hurlement* **S$_0$PredPlus$_1$**(*pause 1*) = *augmentation, hausse, majoration*] **S$_0$Caus$_1$Func$_0$** *(scénario 1)* = *écriture, rédaction*

TABLEAU 2 : Paramètres descriptifs pour les fonctionnelles.

3. Une grammaire des FLs

Une fois que la description des FLs simples et des fonctionnelles est achevée, une grammaire des FLs complexes peut être élaborée. Le but de cette grammaire « formelle » est double.

- Permettre une vérification des combinaisons dans un éditeur électronique (voir la contribution d'Alain Polguère dans Mel'čuk, Clas et Polguère, 1995), de façon à faciliter le travail du rédacteur dans l'encodage des FLs, ce qui devrait lui permettre de se concentrer davantage sur les questions lexico-sémantiques que sur les codifications formelles.

- Permettre l'implémentation des FLs complexes dans un système de traitement automatique qui garantisse la cohérence des informations lexicales.

Les règles peuvent être facilement implémentées à l'aide d'une grammaire hors-contexte. Nous présentons ci-dessous un extrait incomplet d'une telle grammaire pour les FLs complexes. Les signes suivants sont utilisés : « , » pour la séquence, « | » pour un « ou » exclusif, « ? » pour un composant facultatif. Les règles apparaissent en gras, les commentaires en caractères romains.

ind_act --> "1" | "2" | "3".; un indice actantiel est "1", "2" ou "3".
exp_int --> "I" | "II" | "III".; un exposant est "I", "II" ou "III".
verbe_supp --> (("Oper" | "Func"), ind_act) | ("Labor", ind_act, ind_act).; un verbe support est soit Oper, soit Func suivis d'un indice actantiel, soit Labor suivi de deux indices actantiels.
verbe_real --> (("Real" | "Fact"), exp_int?, ind_act) | ("Labreal", exp_int?, ind_act, ind_act). ; un verbe de "réalisation" est soit Real, soit Fact éventuellement suivi d'un exposant de réalisation et d'un indice actantiel obligatoire, soit Labreal suivi éventuellement d'un exposant de réalisation et obligatoirement de deux indices actantiels.
fl_phasale --> "Incep" | "Fin" | "Cont".; les fonctionnelles exprimant une phase sont soit "Incep", soit "Fin", soit "Cont".
fls_verbale --> fl_phasale?, (verbe_supp | verbe_real). ; une FL syntagmatique verbale (c-à-d, qui produit un verbe) est un verbe_supp ou un verbe_real éventuellement précédés d'une fonctionnelle phasale.
fls_nominale --> s0, fls_verbale.; une FL syntagmatique verbale peut être une fonctionnelle nominale suivie d'une FL syntagmatique verbale.

FIGURE 1 : Un extrait de la grammaire formelle des FLs.

Selon cette grammaire « jouet », $\mathbf{S_0IncepReal^{II}_3}$ serait analysée comme une fonction syntagmatique nominale, alors que $\mathbf{Incep_2S_0Real^{I}}$ serait considérée comme agrammaticale. Par ailleurs, comme la grammaire ne peut être seulement construite à partir des données observées dans les trois volumes du DEC, la syntaxe et la sémantique des FLs devront être bien maîtrisées.

4. Les propriétés syntaxiques des collocations

Pour le TALN, l'information syntaxique et sémantique concernant les collocations devra être extrêmement détaillée. Bien que le DEC semble très cohérent et très formalisé pour le lecteur humain (au risque d'être parfois, avouons-le, d'une lecture rebutante), l'information syntaxique concernant les collocations n'apparaît pas suffisamment détaillée.

4.1. Le régime des collocations[15]

Le régime des lexies apparaissant comme bases dans les collocations est très exhaustivement détaillé dans le DEC. Tel n'est cependant pas le cas pour la collocation elle-même dont la structure actantielle est assez superficiellement codée. Or, la structure actantielle de la collocation n'est pas systématiquement déductible du régime de la base. Ainsi, les collocations verbales incluant des noms, en particulier les constructions à verbe support, ne prennent pas nécessairement les mêmes actants que le nom qui est à la base de cette construction :

15 Le régime est la structure actantielle des prédicats.

- de nouveaux actants peuvent apparaître, d'autres disparaître ;
- les actants peuvent être réalisés en surface pas des prépositions différentes.

Par exemple, dans le DEC 1, l'entrée ENSEIGNEMENT **1a** (p. 93) contient les collocations suivantes:

Oper$_1^{\text{actual}}$: donner, dispenser, **litt** prodiguer
Oper$_1^{\text{usual}}$: être [dans l'~]

Le régime de ENSEIGNEMENT **1a** comporte trois actants : l'enseignant (actant 1), la matière (actant 2), les étudiants ou élèves (actant 3). Les valeurs pour **Oper**$_1^{\text{actual}}$ acceptent les trois actants, alors que **Oper**$_1^{\text{usual}}$ n'accepte que le premier :

Léa dispense un enseignement de sémantique aux étudiants de troisième année.
*Léa est dans l'enseignement de la sémantique aux étudiants de troisième année.

Pour qu'un système de génération de textes puisse utiliser la collocation *être dans l'enseignement*, le lexique devrait « savoir » que les actants 2 et 3 ne peuvent pas être exprimés[16].

De plus, les actants peuvent dépendre syntaxiquement plus du nom prédicatif que du verbe. Prenons l'exemple de la collocation *donner un cours*. La lexie COURS a elle-même trois actants : le cours de quelqu'un (actant 1) sur quelque chose (actant 2) à quelqu'un (actant 3). La collocation *donner un cours* a elle aussi trois actants (*quelqu'un* (1) *donne un cours sur quelque chose* (2) *à quelqu'un* (3)), mais tous ne semblent pas dépendre du nom (Alonso Ramos, 1993). Alors que le premier et le troisième actant dépendent du verbe, le second actant semble lié au nom *cours* comme on peut le schématiser dans la représentation syntaxique profonde (simplifiée) de *Lulu a donné un cours de mathématiques à Léa.*

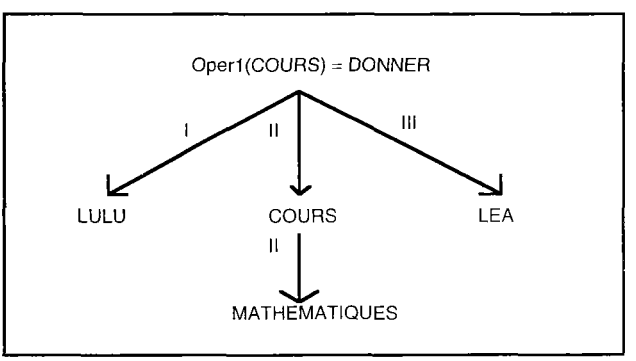


FIGURE 2 : Représentation syntaxique profonde (simplifiée) de *Lulu donne un cours de mathématiques à Léa*

16. Il serait faux de dire que cette information n'apparaît jamais dans le DEC, mais il est vrai qu'elle n'apparaît pas systématiquement. À l'article d'ENTHOUSIASME **1**, le régime comporte deux actants (l'enthousiasme de quelqu'un (1) pour quelque chose (2)). Pour la fonction lexicale **Gener**, il est indiqué que le premier actant ne se réalise pas en surface. On a ainsi :
Gener : sentiment ld'~] I $C_1 = \Lambda$

Le « déplacement » des actants 1 et 3 sur le verbe, alors que l'actant 2 reste attaché au nom semble avoir une incidence sur le comportement syntaxique de la collocation dans les transformations, comme on peut le remarquer lorsque la phrase est passivée :

Lulu a donné un cours de mathématiques à Léa.
Un cours de mathématiques a été donné à Lulu par Léa.
* Un cours a été donné de mathématiques par Léa.

Ces exemples montrent que le régime des collocations doit être extrêmement détaillé si l'on veut pouvoir manipuler ces dernières dans un véritable système informatique. Bien entendu, il apparaît possible d'effectuer quelques généralisations pour le cas des verbes opérateurs. Quelques heuristiques peuvent être utilisées comme valeurs par défaut, comme celles qui ont été proposées par Alonso Ramos (1993) pour les collocations mettant en jeu **Oper$_1$** : dans les collocations à trois actants comme *donner un cours*, le premier et le troisième actants dépendent généralement du verbe, alors que le second actant dépend du nom prédicatif.

4.2. Propriétés distributionnelles et transformationnelles des collocations

Si les collocations doivent être utilisées dans de véritables textes, elles doivent pouvoir être manipulées dans toutes sortes de contextes. Les collocations ne sont pas figées syntaxiquement dans la plupart des cas et la base et le collocatif peuvent apparaître de façon non contiguë. Par exemple, dans de nombreuses collocations de type adjectif-nom, l'adjectif peut être gradué ou apparaître en position d'attribut. On trouve ainsi :

Léa a les cheveux très blonds.
Les cheveux de Léa sont très blonds.

Ce n'est bien entendu pas toujours le cas pour les collocations nom-adjectif, comme dans l'exemple suivant où *rouge* en position d'attribut n'a pas le sens qu'il a dans *vin rouge* (l'interprétation non collocationnelle est notée par un #) :

C'est un vin très rouge.
Ce vin est rouge.

À côté de cela, il est indispensable de noter dans le lexique la position de l'adjectif par rapport au nom. *Gros* est un collocatif devant *fumeur*, mais pas en position postnominale.

Il est par ailleurs intéressant de noter que certains collocatifs productifs (c'est-à-dire, apparaissant dans de nombreuses collocations) comme *gros*, semblent imposer leurs propriétés syntaxiques à la collocation. Ainsi, *gros* quand il est un collocatif d'intensité est préposé au nom, graduable, et n'apparaît pas en position d'attribut :

Un gros fumeur.
Une grosse déprime.
Une grosse envie.

Un fumeur gros.
* Une déprime grosse.
* Une envie grosse.

Un très gros fumeur.
Une très grosse déprime.
Une très grosse envie.

Le fumeur est gros.
? Cette déprime est grosse.
* Cette envie est grosse.

Nous faisons ainsi l'hypothèse que par défaut, les propriétés distributionnelles de la collocation sont héritées du collocatif. Par souci d'économie, il apparaît ainsi plus pertinent de coder les propriétés propres au collocatif, non sur l'entrée de la base, comme cela est effectué actuellement dans le DEC (cf. 4.3.), mais sur l'entrée du collocatif.

Les propriétés « transformationnelles »[17] doivent aussi être codées dans le lexique pour les collocations verbales. Comme pour les propriétés distributionnelles, ces propriétés semblent largement dépendre du collocatif. Par exemple, le verbe support *avoir* n'accepte jamais la transformation passive, alors que tel n'est pas systématiquement le cas pour les verbes support *donner* ou *faire* :

Lulu a {faim, une grosse envie de chocolat}.
* Une {faim, grosse envie de chocolat} est eue par Lulu.

Léa a fait la vaisselle.
La vaisselle a été faite par Léa.

Léa a donné un cours.
Un cours a été donné par Léa.

Pour les collocations verbales, un nombre important de propriétés doit être codé. Parmi celles-ci, hormis la passivation, citons : la passivation réduite (a), la transformation relative (b), la transformation en *se*-passif (c), la transformation impersonnelle passive (d), la pronominalisation (e) :

(a) Le cours donné par Lulu à Léa.
(b) Le cours que Lulu a donné à Léa.
(c) Le cours se donne à l'Université de Montréal.
(d) Il a été donné le même cours à l'Université McGill.
(e) Lulu le donne à l'Université de Montréal, ce cours.

De nombreuses propriétés syntaxiques, qu'elles soient distributionnelles ou transformationnelles, semblent donc davantage héritées des collocatifs plutôt que des bases. Par souci d'économie du codage dans le lexique, il nous semble donc pertinent

17 Il nous paraît commode d'utiliser ce terme démodé pour mentionner les différents types de constructions syntaxiques quasi-synonymes que l'on peut avoir pour un verbe donné.

d'encoder ces propriétés plutôt dans l'entrée du collocatif, même si ce collocatif est sémantiquement vide, comme cela est souvent le cas pour les verbes supports. Cependant, dans le DEC, les propriétés syntaxiques, quand elles sont signalées, apparaissent codées dans l'entrée des bases.

4.3. Une proposition de codage des propriétés syntaxiques des collocations

À l'heure actuelle, le codage des propriétés syntaxiques des collocations dans le DEC s'effectue sur les articles des bases. Ainsi, la propriété [+ préposé] des collocatifs est signalée sur l'article des bases. Par exemple, à l'entrée de CHAGRIN[1]I.1, on relève :

Magn : grand, gros, **litt** noir, vif | prépos ...

Pour les collocations verbe-nom, dans la plupart des cas, l'information syntaxique est très réduite. Ainsi, dans le même article, on relève pour la FL **Oper$_1$** :

Oper$_1$: avoir, éprouver, ressentir [du ~ | C. n'a pas de dépendant]; **litt** connaître [ART ~]

Ici, il n'est pas indiqué que le passif n'est pas autorisé pour le verbe *avoir.*

Il nous paraît essentiel d'introduire des mécanismes d'héritage pour coder la syntaxe des collocations, tant pour gérer le régime des collocations que les propriétés distributionnelles et transformationnelles.

Dans le cas des collocations verbe-nom, le régime des collocations sera calculé à partir de celui de la base. Les actants de la collocation sont calculés à partir d'heuristiques telles que celles qui ont été proposées par Alonso Ramos (1993) (cf. 4.1.). Les collocations pour lesquelles le régime apparaît en contradiction avec cette heuristique sont explicitement codées dans l'entrée du collocatif.

Les propriétés syntaxiques (distribution et transformations) sont codées par défaut sur l'entrée du collocatif. La collocation hérite par défaut des propriétés du collocatif comme cela est schématisé dans la figure suivante avec l'exemple *gros fumeur.*

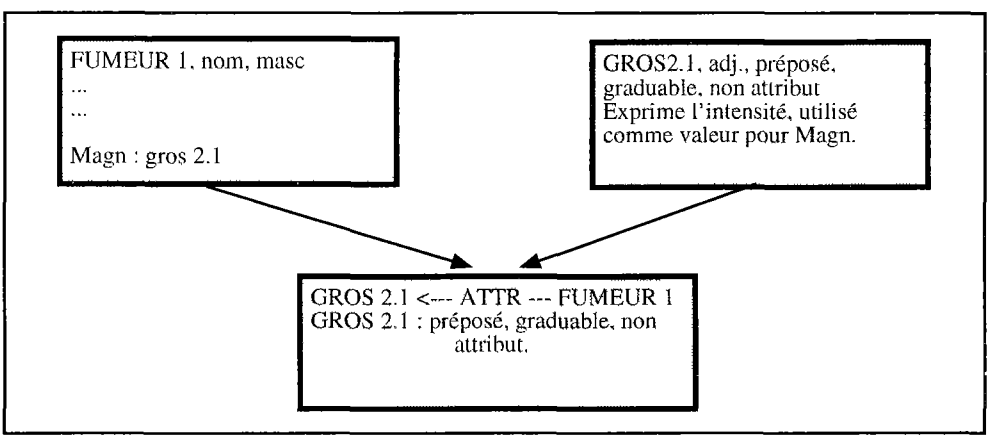


FIGURE 3 : Héritage des propriétés syntaxiques de la collocation

Cependant, les collocations n'héritent pas toujours de toutes les propriétés du collocatif. Il faut pouvoir dans ce cas prévoir un héritage « non monotone » par lequel les propriétés marquées sur la base auraient préséance sur les propriétés héritées du collocatif. Bien évidemment, ces mécanismes d'héritage ne peuvent pas être manipulés à la main. Cela implique l'utilisation d'un éditeur intégrant des mécanismes assez complexes d'héritage.

Conclusion

Dans le cadre de cette étude sur la formalisation des collocations, nous nous sommes intéressée, d'une part au mode de codage des collocations, les Fonctions Lexicales, et d'autre part, aux propriétés syntaxiques dont on devrait disposer pour les collocations. Avant que les FLs puissent être intégrées dans un système informatique, une description minutieuse des FLs (fonctions simples et fonctionnelles) et des combinaisons de celles-ci doit être effectuée. Ce travail, en cours, doit être intégré dans l'éditeur mis au point par Alain Polguère (Mel'čuk, Clas et Polguère, 1995).

Le second volet de notre étude concerne plus particulièrement le TALN, bien que l'information concernant les propriétés syntaxiques soit tout aussi indispensable dans un DEC « papier ». Il nous paraît crucial de pouvoir disposer, d'une part, d'une structure actantielle détaillée, d'autre part, de l'ensemble des propriétés syntaxiques essentielles concernant la collocation. Pour ce faire, il nous semble indispensable d'intégrer dans le lexique des mécanismes d'héritage non monotone, de façon à alléger le codage manuel concernant les collocations. Ces mécanismes d'héritage ne sont néanmoins pas facilement manipulables à la main et rendent absolument nécessaire le recours à un éditeur informatique facilitant le travail du lexicographe.

Description lexicographique des collocatifs dans un *Dictionnaire explicatif et combinatoire* : articles de dictionnaire autonomes ?[1]

Margarita ALONSO RAMOS et Suzanne MANTHA

Universidade da Coruña, Espagne et Université de Montréal, Canada

1. Présentation du problème

Dans ce travail, nous nous concentrons sur la description lexicographique des COLLOCATIONS et plus particulièrement sur celle des COLLOCATIFS dans un *Dictionnaire explicatif et combinatoire* (DEC). Notre objectif principal est d'étudier la possibilité de traiter les collocatifs comme des lexies (de plein droit), ce qui entraîne la nécessité d'avoir pour ceux-ci des articles de dictionnaire autonomes.

Nous commençons par donner quelques exemples du phénomène lexical des collocations – syntagmes spéciaux du type suivant : *OPÉRER un choix, RAVALER sa colère, MOURIR d'envie, LEVER l'ancre, prix DÉRISOIRE, café NOIR, cheveux BLONDS, peur BLEUE*, etc.

Dans tous ces syntagmes, le verbe ou l'adjectif est choisi par le locuteur en fonction du nom. Si l'on part d'autres noms, le choix des verbes ou des adjectifs pour exprimer le même sens sera différent : *PRENDRE <*OPÉRER> une décision, CONTENIR <*RAVALER> sa fureur, HISSER <*LEVER> les voiles, thé NATURE <*NOIR>, peau DORÉE <*BLONDE>, fureur NOIRE <*BLEUE>*, etc. Dans ces exemples, les noms correspondent à ce que nous appelons les « bases » des collocations et les verbes ou adjectifs à ce que nous appelons les « collocatifs »[2]. Nous nous

1. La réalisation de cette étude a été rendue possible grâce à l'aide financière accordée par le Conseil de recherches en sciences humaines du Canada (subvention n° 410-91-1844), par les Fonds FCAR (subvention n° 96-ER-0618) ainsi que par le Ministerio de Educación y Ciencia du gouvernement espagnol (bourse post-doctorale)

2. Cette correspondance (nom = base et verbe/adjectif = collocatif) n'est pas la seule possible : il existe d'autres collocations comme *coûter LA PEAU DES FESSES* et *risquer LE PAQUET* où, cette fois-ci, les verbes sont des bases et les noms, des collocatifs

intéressons à la description de ce dernier type de lexèmes – les lexèmes qui ne sont pas choisis librement.

En parlant des collocations, on insiste souvent sur l'inégalité entre les deux éléments d'une collocation : la base est sémantiquement autonome, tandis que le collocatif ne l'est pas (cf. entre autres, Hausmann, 1979 ; Heid, 1992a-b, 1994). Pour cette raison, le statut lexical des collocatifs n'est pas toujours reconnu : est-ce qu'un collocatif comme *opérer* dans *opérer un choix* est une entité lexicale suffisamment autonome pour créer un article de dictionnaire à part qui ne décrive que ce *opérer* ?

D'après notre approche (Mel'čuk, 1988, 1995a-b, 1996), ce qui caractérise spécifiquement les collocatifs, c'est leur façon spéciale d'être choisis dans le processus de la synthèse du texte. Le locuteur (humain ou machine) doit faire des choix lexicaux, c'est-à-dire qu'il doit choisir des lexies pour exprimer des sens donnés. Dans cette perspective, il faut distinguer deux types de lexies. Le premier type est constitué de lexies (la plupart, en fait) qui sont sélectionnées pour exprimer un sens donné INDÉPENDAMMENT d'autres lexies. Ainsi, si le locuteur veut exprimer le sens ‹attitude émotionnelle défavorable de X à l'égard de Y causée par le fait que X déteste Y à tel degré que X veut faire des actions agressives à Y...›, il choisit tout simplement HAINE – sans prendre en considération d'autres lexies déjà choisies. Dans ce cas, il s'agit d'un choix lexical SÉMANTIQUEMENT CONTRÔLÉ. Le deuxième type est constitué de lexies qui sont sélectionnées pour exprimer un sens donné mais qui sont sous le contrôle d'autres lexies choisies auparavant. Ainsi, pour exprimer le sens ‹intense› en parlant de la HAINE, le locuteur doit considérer la lexie à laquelle ce sens s'applique. Dans le cas de *haine*, c'est *mortelle* qui sera sélectionné pour exprimer ‹[haine] intense›. La sélection de MORTEL dans ce sens-là est doublement contrôlée : SÉMANTIQUEMENT, car on veut exprimer le sens ‹intense› et non pas le sens ‹sentir›, par exemple, et LEXICALEMENT, car le sens ‹intense› s'exprime par MORTEL auprès de HAINE mais pas auprès de RAGE, FUREUR, JOIE ou DÉSESPOIR, par exemple, où les choix lexicaux sont différents : *rage folle, fureur aveugle, vive joie, profond désespoir.*

Une collocation est donc une paire d'expressions lexicales, la base et le collocatif, telle que le choix de la base n'est contrôlé que sémantiquement, alors que le choix du collocatif est, en plus, contrôlé lexicalement – notamment, par la base. Au moment de la production du texte, la base est choisie en premier, car son sens est plus perceptible que celui du collocatif et il est exprimé de façon indépendante.

Le fait que les collocatifs soient sélectionnés en fonction de leur base affaiblit leur statut lexical et rend problématique leur description lexicographique. Les dictionnaires traditionnels n'ont pas de politique uniforme pour le traitement des collocatifs : soit qu'on crée une acception particulière pour le collocatif (p. ex., *blond* dans le *Petit Robert*), soit que le collocatif est répertorié comme un sens figuré sans marque explicite de nouvelle acception (*ravaler* [*sa colère*]), soit qu'il est consigné dans l'article de la base avec une petite description du sens de toute la collocation ([*feu*] *rouge*), ou encore qu'il est simplement consigné dans l'article de la base mais sans description (*rouge* [*de colère*]).

La création d'articles de dictionnaire pour les collocatifs présuppose qu'on leur attribue un statut de lexie de plein droit. Dans notre perspective, les notions « être une

lexie » et « avoir un article de dictionnaire » sont équivalentes. Par conséquent, une lexie est un item lexical qui a besoin d'un article de dictionnaire. Cependant, il n'est pas évident que tous les collocatifs aient besoin d'un article. Comme on le verra plus loin, les collocatifs ne sont pas tous du même type : il y en a qui se combinent avec une seule base (p. ex., *café NOIR* ; cf. section 3.1.), d'autres qui sont des syntagmes libres (p. ex., *fumer UNE CIGARETTE APRÈS L'AUTRE*). Dans ces deux cas, il n'est pas question de créer un article de dictionnaire pour les collocatifs, car ceci reviendrait à dire, dans le premier cas, que l'on attribue un statut de lexie de plein droit à un collocatif qui se combine avec une seule base et, dans le deuxième, que l'on attribue un statut de lexie à un syntagme libre.

Le problème à résoudre est de préciser le statut lexical des collocatifs de types différents et de trouver la bonne façon de les décrire lexicographiquement. Dans ce travail, nous étudierons les possibilités de description des collocatifs dans le cadre du DEC avec la rigueur et le formalisme requis.

Nous poursuivons notre discussion en procédant par les étapes suivantes :

– présentation du traitement actuel des collocations dans le DEC par l'appareil lexicographique des fonctions lexicales ;
– examen de certains traits distinctifs des collocatifs qui mesurent leur degré d'autonomie ;
– analyse des stratégies possibles de description et, pour finir, présentation d'une liste de vérification pour guider le lexicographe dans son choix de traitement à donner à un collocatif.

2. Les collocations dans le DEC : notion et description

Notre objectif principal étant de montrer comment le DEC peut décrire les collocatifs, nous devons donc commencer par caractériser, même si brièvement[3], ce type de dictionnaire. Les six propriétés de base du DEC sont les suivantes : 1) il est rédigé dans le cadre de la théorie linguistique Sens-Texte ; 2) il est orienté vers la production du texte ; 3) un article du DEC est fondé sur la définition de la lexie-vedette : sa représentation sémantique sert de base à la description de toutes ses relations paradigmatiques et syntagmatiques avec les autres lexies de la langue ; 4) il met de l'emphase sur la cooccurrence lexicale restreinte ; 5) il est formalisé, donc toutes les informations concernant le sens, la syntaxe, etc. doivent être indiquées de façon précise et explicite ; 6) il est exhaustif au niveau de la description de chaque lexie individuelle.

Une caractéristique importante du DEC est que l'unité de description lexicographique est une LEXIE, qui peut être un lexème ou un phrasème complet. Un lexème est un mot pris dans une acception déterminée, tandis qu'un phrasème complet est un syntagme non libre pris aussi dans une acception déterminée. À chaque lexie, le DEC

3 Pour une caractérisation détaillée du DEC, nous référons le lecteur aux ouvrages suivants : Mel'čuk, 1982, 1995a-b, 1996 ainsi que Mel'čuk *et al.*, 1984, 1988, 1992, 1995.

fait correspondre un seul article de dictionnaire. Dans ce sens, il se démarque radicalement des dictionnaires courants qui prennent le mot polysémique comme unité de description.

Les lexies qui sont sémantiquement liées par des ponts sémantiques (*i.e.* qui partagent des composantes sémantiques communes relativement importantes) et dont les signifiants sont identiques sont regroupées en VOCABLES. À chaque vocable correspond un superarticle qui est formé de tous les articles des lexies faisant partie de ce vocable. Comme nous le verrons plus loin à la section 3.2., il existe des vocables constitués d'une seule lexie (les « mots » monosémiques) et d'autres constitués de plusieurs (les « mots » polysémiques). Les vocables polysémiques peuvent regrouper des lexies qui ont un comportement libre, c'est-à-dire qui sont choisies d'après leur sens, et/ou d'autres lexies qui sont restreintes, c'est-à-dire qui sont contrôlées par d'autres lexies déjà choisies. En d'autres termes, un vocable polysémique peut être constitué de lexies autonomes et/ou de « lexies collocatives ». Nous appellerons les premières « contreparties libres des collocatifs ».

Les propriétés du DEC qui nous concernent le plus sont la deuxième et la quatrième. Le DEC, conformément à la théorie Sens-Texte, est orienté vers la production du texte, plutôt que vers la compréhension. Il doit donc fournir tous les moyens lexicaux pour exprimer un sens donné. Ainsi, la question à laquelle le DEC prétend répondre est « comment exprimer 'X' » plutôt que « qu'est-ce que X signifie ». En d'autres termes, le DEC est un dictionnaire adapté pour la synthèse plutôt que pour l'analyse des textes. La raison derrière cette approche est la conviction que la direction à partir du sens vers le texte (c'est-à-dire parler) est une activité linguistique par excellence, tandis que la direction inverse (comprendre) entraîne d'autres activités, qui ne sont pas purement linguistiques : le destinataire qui doit déchiffrer un énoncé doit faire appel à ses habilités logiques ainsi qu'à ses connaissances sur la situation, sur le monde, etc.

Étant donné son orientation vers la synthèse, le DEC est un dictionnaire centré sur la cooccurrence lexicale restreinte, c'est-à-dire la cooccurrence qui n'est pas déterminée sémantiquement à cent pour cent. Tous les lexèmes qui se combinent avec le lexème vedette d'un article et dont le choix est contrôlé lexicalement sont systématiquement décrits dans le DEC – au moyen des fonctions lexicales (= FL).

Avant de présenter les FL, nous rappelons ici brièvement l'approche de la théorie Sens-Texte par rapport aux collocations. Celles-ci sont comprises comme une sous-classe de phrasèmes, *grosso modo* des syntagmes non libres. Mel'čuk (1995a-b) distingue trois types de phrasèmes : 1) PHRASÈMES COMPLETS : *casser sa pipe, bas-bleu* (où le sens de l'expression n'inclut le sens d'aucun de ses constituants) ; 2) QUASI-PHRASÈMES : *donner le sein* (Mel'čuk *et al.*, 1992 : 191), *casque bleu* (où le sens de l'expression inclut les sens des deux lexèmes constituants plus un ajout sémantique imprévisible) ; 3) SEMI-PHRASÈMES qui correspondent en fait à ce que nous entendons par COLLOCATIONS (où le sens de l'expression inclut le sens d'un des lexèmes constituants et le sens de l'autre constituant est exprimé par un lexème choisi non librement).

Ces trois types de phrasèmes sont décrits dans le DEC de deux façons différentes : soit comme lexies vedettes d'articles autonomes (les phrasèmes complets et

les quasi-phrasèmes), soit comme éléments dans les articles de leurs bases (les collocations). Les deux premiers ne peuvent pas être décrits en fonction de leurs constituants, tandis que les troisièmes peuvent et doivent être décrits en fonction d'un de leurs constituants – la base – au moyen de FL dans l'article de la base.

Dans le DEC, on utilise une terminologie parallèle, où la base d'une collocation correspond au mot-clé et le collocatif, à un élément de la valeur d'une FL. Le DEC adopte cette terminologie pour mettre en relief l'aspect fonctionnel de la description des collocations.

Pour présenter le traitement des collocations dans le DEC, il nous faut introduire la notion de FL. *Grosso modo*, une FL est une fonction **f** qui associe à une lexie L – le mot-clé (= la base) de **f** – un ensemble L_i d'expressions lexicales – la valeur (= le collocatif) de **f** – qui sont choisies en fonction de L pour exprimer le sens correspondant à **f**. En d'autres mots, étant donné la collocation **A + B** signifiant 'A + C', alors le sens 'C' qui est exprimé par le collocatif **B** auprès de la base **A** correspond à ce qui est appelé une **FL**. Le lexème **A** qui garde intact son sens et détermine le choix de l'expression pour 'C' est le **mot-clé** de la FL. Le lexème **B** qui est choisi pour exprimer 'C' auprès de **A** est la **valeur** de la FL.

Il convient de distinguer deux sous-classes de FL : les standard et les non standard. Les premières doivent satisfaire les quatre conditions suivantes :

1. Pour toute paire de lexies L_1 et L_2, les lexies $\mathbf{f}(L_1)$ et $\mathbf{f}(L_2)$ montrent des relations (presque) identiques à ces lexies :

$$\frac{\text{'}\mathbf{f}(L_1)\text{'}}{(L_1)} \approx \frac{\text{'}\mathbf{f}(L_2)\text{'}}{(L_2)}$$

Ex. : Soit la FL **f** = **Magn** (= 'intensificateur'), L_1 = FROID et L_2 = SOURD, et $\mathbf{f}(L_1)$ = *de canard* et $\mathbf{f}(L_2)$ *comme un pot*. La relation sémantico-syntaxique entre *de canard* et FROID est la même que celle entre *comme un pot* et SOURD. Les deux expressions sont des modificateurs intensificateurs qui signifient, dans les contextes donnés, 'intense', 'très'.

2. En règle générale, $\mathbf{f}(L_1)$ et $\mathbf{f}(L_2)$ sont différents : $\mathbf{f}(L_1) \neq \mathbf{f}(L_2)$.
Ex. : *très <*grièvement> malade* vs *grièvement <*très> blessé* ; *une grippe carabinée* vs *une fièvre de cheval ; aimer à la folie* vs *haïr à mort.* Les expressions dépendent des lexies modifiées.

3. La fonction **f** a un nombre élevé de mots-clés. Le sens 'f' est très abstrait et très général et s'applique à beaucoup d'autres sens.

4. La fonction **f** a un nombre élevé d'éléments dans sa valeur (= d'expressions).

Les FL standard sont codées dans le DEC par des symboles conventionnels comme le montrent les exemples suivants :

FL mot-clé		**valeur**
Magn(*désir*)	=	*ardent*
Magn(*envie*)	=	*folle*
Magn(*interdire*)	=	*absolument*
Mult(*abeille*)	=	*essaim* [*d'abeilles*]
Func$_0$(*silence*)	=	*régner*
Oper$_1$(*baiser*)	=	*donner* [*un baiser*]
Oper$_1$(*câlin*)	=	*faire* [*un câlin*]

Lorsque les conditions 3 et 4 données ci-dessus sont violées, on a affaire à l'autre type de FL, soit les non standard. Leurs sens sont trop spécifiques, de sorte qu'ils ne s'appliquent qu'à très peu de lexies et ne possèdent qu'un infime nombre d'expressions différentes. Les FL non standard sont spécifiées en français. Ainsi, dans l'article du lexème CAFÉ, on trouverait, entre autres, les FL suivantes :

Magn	: *fort* [FL standard]
sans lait	: *noir* [FL non standard]
avec de la crème ou un peu de lait	: *crème* [FL non standard]

Les FL sont consignées dans l'article de dictionnaire du lexème qui est le mot-clé de la FL, c'est-à-dire la base de la collocation.

Pour produire correctement une collocation, il ne suffit pas, en règle générale, d'additionner les propriétés des deux lexèmes (la base et le collocatif) qui la constituent. Dans ces expressions, soit la base, soit le collocatif peut avoir des propriétés qui ne concernent qu'une collocation particulière.

Le premier cas : propriétés particulières de la base. Nous pensons aux collocations verbales où, très souvent, le nom a un comportement déviant par rapport aux déterminants. Par exemple, en espagnol, le nom *respeto* ('respect') exige ou rejette l'article selon la collocation dans laquelle il apparaît. Quand il se combine avec le verbe *tener* ('avoir'), la collocation peut signifier ('respecter [quelqu'un]') ou ('être respecté [par quelqu'un]'). Dans le premier sens, le nom ne peut pas être déterminé :

Juan tiene (**el*) *respeto por su padre*, litt. 'Juan a respect pour son père'

tandis que, dans le second, le nom doit être déterminé :

El padre tiene *(*el*) *respeto de su hijo*, litt. 'Le père a le respect de son fils'.

Dans le DEC, cette information apparaît dans le schéma de régime réduit de la collocation, où elle est présentée de la façon suivante :

$\text{Oper}_1(respeto) = tener\ [\sim]$
$\text{Oper}_2(respeto) = tener\ [\text{ART}_{déf}\ \sim] \mid C_1 \neq \Lambda$

La notation qui apparaît après la barre verticale indique que le premier actant du mot-clé (= la désignation de celui qui respecte) ne peut pas être vide, c'est-à-dire qu'il doit être exprimé : *El padre tiene el respeto de Juan <su respeto>* vs **El padre tiene el respeto.*

Le deuxième cas : propriétés particulières du collocatif. Considérons, par exemple, l'adjectif *fort* qui est soit préposé soit postposé avec *pluie* mais seulement postposé avec *café* :

une forte pluie ~ *une pluie forte* [sans changement de sens]
**un fort café* ~ *un café fort*

Ces deux cas ne sont pas une particularité de FORT mais seulement de la collocation concernée. Par conséquent, l'emploi seulement postposé de *fort* doit être indiqué dans la description de la collocation *café fort* :

Magn(*café*) = *fort* | postposé

Chacune de ces FL avec ses valeurs constitue une description de la collocation en question – avec la particularité que cette description est enchâssée dans l'article de la base (le mot-clé).

Si le DEC dispose d'une politique bien établie pour décrire les collocations, ce n'est pas le cas pour la description des collocatifs. Comme justification de la création d'un article de dictionnaire pour un collocatif, Mel'čuk (1995a) propose le critère de généralisation suivant : si l'élément L_i (le collocatif) de la valeur d'une FL partage les mêmes particularités de régime ou de flexion avec beaucoup de mots-clés (particularités par rapport à L_i' libre), alors, afin d'éviter la répétition d'information, L_i peut avoir son propre article – avec une description exhaustive de son régime et de toutes ses autres propriétés.

Or il y a des collocatifs qui n'ont pas de contreparties employées librement (p. ex., *bissextile*) de sorte qu'on ne peut pas compter sur l'idée d'indiquer l'information dans un autre article de dictionnaire. Il y en a d'autres comme *noir* dans *café noir* qui ont de telles contreparties, mais dont le sens est exclusif à la collocation avec le nom *café*.

Dans les trois volumes du DEC publiés, le traitement des collocatifs n'est pas uniforme. Quand les collocatifs ont des contreparties libres, on peut trouver des articles pour ceux-ci, soit parmi les autres lexies du vocable (p. ex., plusieurs lexèmes collocatifs appartenant au vocable FLAMBER), soit dans un autre vocable à part (on y voit donc un cas d'homonymie, p. ex., REMPLIR[2]).

Pour standardiser le traitement des collocatifs, il faut commencer par en ébaucher la typologie. Dans la section suivante, nous allons étudier différents traits distinctifs des collocatifs, ce qui nous aidera à déterminer le traitement lexicographique adéquat pour chaque type de collocatif.

3. Traits distinctifs des collocatifs

Étant donné la nature hétérogène des collocatifs, nous les distinguerons en nous basant sur les traits distinctifs suivants :

1. à base unique *vs* non unique ;
2. isolés *vs* ayant une contrepartie libre ;

3. décrits par une FL standard *vs* non standard ;
4. plus *vs* moins phraséologisés ;
5. plus *vs* moins autonomes.

Précisons que notre liste de traits est tout à fait provisoire. Elle n'est ni exhaustive ni définitive. Cette tentative d'établir une typologie des collocatifs au moyen de traits est nouvelle dans le cadre du DEC. Étant donné la quantité de collocations à traiter dans le lexique français et la quantité d'informations à donner pour chacune d'elles, il serait nécessaire de mener une étude beaucoup plus large que la présente. De nouveaux traits devront probablement être introduits pour permettre la distinction de tous les types de collocatifs.

Un trait à lui seul n'est pas suffisant pour décider du traitement à donner à la description d'un collocatif. Par conséquent, c'est une combinaison de traits qu'il faudrait considérer pour décider du type auquel correspond un collocatif.

Les traits ne sont pas indépendants les uns des autres. Ils n'ont pas tous la même importance et donc certaines combinaisons de traits sont plus concluantes que d'autres. Une hiérarchie dans les combinaisons de traits devrait être établie.

Notre liste doit être utilisée avec prudence. Nous la croyons quand même utile pour guider le lexicographe dans sa démarche lexicographique.

3.1. Collocatifs à base unique *vs* non unique

Certains collocatifs ne se combinent qu'avec une seule base, tandis que d'autres sont productifs, c'est-à-dire qu'ils se combinent avec un nombre élevé de bases.

Examinons d'abord des collocatifs à base unique. Ainsi, l'adjectif NOIR signifiant ‹sans lait› est un collocatif à base unique ; dans le sens donné et avec les propriétés données (postposition), il ne se combine qu'avec le lexème CAFÉ en excluant même la combinaison avec ses hyponymes :

**un expresso noir, *un déca noir, *un arabica noir.*

Un autre collocatif à base unique est l'adjectif PLATE au sens ‹sans bulles› : il a ce sens seulement en combinaison avec le nom EAU. Aucune autre boisson ne peut être qualifiée de *plate* pour exprimer le sens ‹sans bulles›.

L'adjectif BISSEXTILE est aussi un cas de cooccurrence unique mais il ne s'agit pas du même type de collocatifs que ceux vus ci-dessus. Le sens ‹qui, au lieu d'avoir 365 jours, a une journée de plus, le 366e jour étant le 29 février› ne peut être appliqué qu'au lexème ANNÉE.

Les adjectifs *noir / plate* et *bissextile* sont des collocatifs à base unique qui se distinguent par les propriétés suivantes :

– Le fait que NOIR ‹sans lait› ne se combine qu'avec CAFÉ (ou PLATE ‹sans bulles› seulement avec EAU) est tout à fait arbitraire, comme c'est le cas pour la plupart

des collocations. On pourrait dire qu'il s'agit d'UNICITÉ LEXICALE. L'adjectif NOIR signifiant ‹sans lait› est SÉMANTIQUEMENT applicable à d'autres lexèmes mais pas LEXICALEMENT : **thé noir, *tisane noire.*

– Le fait que BISSEXTILE ne se combine qu'avec ANNÉE est déterminé sémantiquement. C'est le sens du collocatif qui interdit toute autre cooccurrence. On pourrait dire qu'il s'agit d'UNICITÉ SÉMANTIQUE. L'adjectif BISSEXTILE n'est pas sémantiquement applicable à d'autres lexèmes et, par conséquent, non plus lexicalement.

Par rapport aux collocatifs à base non unique, on peut établir une échelle qui va des collocatifs qui se combinent avec un petit nombre de bases (DE FER : *volonté, santé, discipline*) jusqu'à ceux dont le nombre de bases est beaucoup plus élevé, environ une centaine (FORT : *vent, médicament, lunettes, lumière, son, goût, moutarde, café, moment,* etc.).

À un extrême de cette échelle se trouve une sorte de collocatif qui, en principe, se combine avec une ou deux bases mais qui peut aussi se combiner avec des noms propres, liés aux bases « originelles » par la relation d'instanciation. C'est le cas de SE JETER, signifiant ‹[cours d'eau] se verse dans un autre cours d'eau ou une étendue d'eau›. Les bases originelles de ce collocatif sont FLEUVE, RIVIÈRE et RUISSEAU, mais tous les noms propres désignant des cours d'eau admettent aussi la combinaison avec SE JETER : *Le Saint-Laurent <La Volga, Le Duero> se jette dans la mer.* Un autre cas proche de celui-ci est la combinaison possible de BLOND avec tous les noms qui incluent le sens ‹cheveu ou poil›. Ainsi l'adjectif BLOND, qui signifie ‹[cheveu ou poil] dont la couleur est similaire à la couleur du blé›, peut aussi se combiner avec, entre autres, MÈCHE, TIGNASSE, TRESSE ou PERRUQUE. La relation entre ces noms et les bases « originelles » est une sorte de méronymie. Nous appelons ce type de collocations, « collocations par héritage ».

À l'autre extrême de l'échelle, se trouvent des collocatifs comme FORT, GRAND, GROS, FAIRE, PRENDRE, etc. qui se combinent avec un nombre très élevé de bases. On pourrait dire que ces collocatifs jouent le rôle de « joker » : ainsi, FORT, GRAND et GROS sont les « jokers » pour l'intensificateur et FAIRE, le « joker » pour le verbe support. Dans ces cas, la première question qui se pose est de savoir, d'une part, si tous ces collocatifs sont toujours le même lexème ou pas, c'est-à-dire si, par exemple, *fort* est le même lexème dans *un café fort, une moutarde forte* ou *un son fort*. D'autre part, la productivité de ces collocatifs rend plus difficile leur description lexicographique parce qu'il n'y a pas toujours moyen de faire de regroupements sémantiques pour toutes les bases possibles et il y a des résidus inclassifiables. D'ailleurs, on doit s'attendre à de tels résidus justement parce qu'il s'agit de cooccurrence lexicale restreinte.

En somme, nous croyons que plus un collocatif est unique, moins il a de chances d'avoir un article autonome.

3.2. Collocatifs isolés *vs* ayant une contrepartie libre

Un autre trait qui aura des effets sur le traitement lexicographique à donner au col-

locatif porte sur la monosémie ou la polysémie du vocable auquel le collocatif appartient. On peut observer des collocatifs qui constituent des vocables monosémiques. C'est le cas de BISSEXTILE, RAUQUE, AQUILIN qui n'existent qu'en tant que collocatifs : nous les appellerons « collocatifs isolés ». D'autres collocatifs appartiennent à des vocables polysémiques parmi lesquels il faut distinguer les vocables dont tous les lexèmes constituants sont des collocatifs (p. ex., BLOND, CLIGNER) de ceux qui sont composés à la fois de lexèmes collocatifs et de lexèmes libres (p. ex., NOIR, PLAT, FORT, OPÉRER).

Les collocatifs isolés doivent avoir un article autonome, car il est nécessaire d'indiquer leur existence dans le lexique de la langue et on ne peut pas le faire autrement. Les collocatifs qui ont d'autres partenaires collocatifs le méritent autant. Par contre, les collocatifs ayant des contreparties libres peuvent se passer d'un article autonome si leur forme phonologique, leur partie de discours et/ou toutes autres sortes de données morphologiques, syntaxiques ou lexicales sont partagées avec les lexies libres du même vocable.

La coïncidence entre les propriétés des lexies libres et celles des collocatifs ne survient pas systématiquement. Il est fréquent de trouver certaines propriétés qui se manifestent dans l'emploi libre mais non pas dans l'emploi collocationnel et *vice versa*. Ainsi, par exemple, le verbe *opérer* employé librement permet l'aspect progressif, tandis qu'employé en collocation, cet aspect n'est pas exprimable :

a. *Jean est en train d'opérer un malade.*
b. **Jean est en train d'opérer un choix.*

Le fait que le collocatif en **b** ne présente pas la même propriété que celle de sa contrepartie libre en **a** ne prouve pas nécessairement qu'il s'agit de deux lexèmes OPÉRER différents. La perte de propriétés que subit le collocatif peut être imputée au phénomène général de la phraséologisation. Cependant le cas inverse est révélateur. Prenons l'exemple de FORT : le lexème de base FORT a comme intensificateur l'expression *comme un Turc*, tandis que le collocatif FORT, lui, ne l'admet pas (**café fort comme un Turc*, **son fort comme un Turc*, etc.). De son côté, le collocatif peut avoir son propre intensificateur : [*son*, *bruit*, *cri*,...] *fort à crever les tympans* (mais pas **café fort à crever*). Dans ces cas où le collocatif possède une caractéristique que l'emploi libre n'a pas, la création d'un article de dictionnaire pour ce type de collocatif serait plus justifiable, car ses propriétés doivent être mentionnées dans un DEC.

À la lumière de ces considérations, nous pouvons affirmer que les collocatifs isolés doivent avoir un article de dictionnaire. Dans les autres cas, nous croyons que plus l'emploi collocationnel considéré possède ses propres caractéristiques syntaxiques, lexicales, etc., plus le collocatif aura de chances de posséder un article autonome.

3.3. Collocatifs plus phraséologisés *vs* moins phraséologisés

On peut établir une échelle qui part des collocatifs très liés phraséologiquement à la base jusqu'à ceux qui le sont moins. Les collocatifs plus phraséologisés sont perçus par les locuteurs comme des éléments d'un tout, à savoir la collocation, qui sont inséparables. En d'autres termes, le sens d'un tel collocatif est très peu perceptible,

voire pas du tout, sans spécification de sa base. Ainsi, par exemple, un collocatif comme BLEU dans *un steak bleu* est très lié phraséologiquement à la base. Pour exprimer le sens ‹très saignant›, le français a choisi l'adjectif BLEU. La lexie de base, BLEU ‹de couleur bleue›, n'a pas de connotations qui établissent un pont sémantique avec BLEU ‹saignant›. Le choix de BLEU pour ce sens est totalement arbitraire et idiosyncratique synchroniquement.

Les sens des collocatifs moins phraséologisés sont plus facilement perçus séparément de leur base : en fait, ces collocatifs se combinent avec un nombre très élevé de bases qui *grosso modo* sont regroupables sémantiquement. Par exemple, les verbes ÉPROUVER et SOUFFRIR, comme *éprouver de la haine* et *souffrir d'une maladie* jouent le rôle de verbe support des noms d'émotions et des noms de maladies respectivement[4].

Il y existe d'autres collocatifs encore moins phraséologisés que ÉPROUVER (avec émotion) et SOUFFRIR (avec maladie). Par exemple, le verbe CONDUIRE : il peut apparaître sans sa base (*Il conduit vite*) et il garde le même sens en emploi libre comme dans *Il conduit vite* et en combinatoire restreinte comme dans *conduire une auto*. Ici, la distance sémantique entre le collocatif et sa contrepartie libre est zéro. Ceci confère au verbe CONDUIRE un double statut : il peut tantôt être une lexie collocative et tantôt une lexie de plein droit tout dépendant du point de vue à partir duquel l'expression est produite. En effet, quelquefois *conduire une auto* est bel et bien une collocation : cette expression ne peut être considérée comme entièrement libre, car, en français, pour exprimer le sens ‹faire avec l'auto ce à quoi elle est destinée›, on dit *conduire*, tandis qu'en espagnol d'Amérique du Sud, on choisirait le verbe *manejar* qui correspond littéralement à ‹manipuler› en français. D'autres fois, CONDUIRE est une lexie de plein droit : il s'emploie sans sa base, il possède sa propre combinatoire lexicale, ses propres dérivés, etc. Comme ces propriétés appartiennent à CONDUIRE et qu'elles sont assez nombreuses, alors il doit avoir son propre article de dictionnaire autonome.

En somme, moins le collocatif est phraséologisé, plus il a de chances d'avoir un article autonome.

3.4. Collocatifs décrits par une FL standard *vs* non standard

Le sens d'une FL standard est assez abstrait et général, comme ‹très›, ‹causer›, ‹faire ce qu'on est censé faire›. Il peut être appliqué à nombre très élevé de bases et le nombre de collocations obtenues l'est aussi. Le sens d'un collocatif décrit au moyen d'une FL standard a donc besoin d'être plus spécifique et les bases avec lesquelles il se combine doivent être regroupées sémantiquement. Nous pensons à la FL **Magn**, où l'intensification n'agit pas toujours sur les mêmes composantes sémantiques. Elle dépend du sémantisme de la base. Par exemple, *fort* auprès de *médicament* intensifie la

4 Même si *éprouver* signifie ‹éprouver› et *souffrir*, ‹souffrir›, ces expressions sont bel et bien des collocations. Dans le DEC, elles sont décrites par la FL **Oper** et, dans cette approche, toute valeur d'une FL combinée avec son mot-clé constitue nécessairement une collocation.

concentration de la substance tandis qu'auprès de *moutarde*, il intensifie le goût qui est assez prononcé. Il semble donc nécessaire de préciser les sens des collocatifs et un endroit approprié pour le faire serait la zone sémantique d'un article de dictionnaire. Il faut donc envisager la possibilité de créer un article de dictionnaire pour ces collocatifs.

À l'opposé, le sens d'une FL non standard est plus concret, donc limité à un nombre plus réduit de bases possibles et avec un nombre d'expressions moins élevé. Il n'est donc pas nécessaire de créer un article de dictionnaire, le sens d'un tel collocatif n'ayant pas besoin d'être spécifié davantage.

Si l'on veut rédiger une définition lexicographique pour un collocatif qui est décrit par une FL standard dans l'article de la base, elle sera formulée en termes assez abstraits et généraux (sauf quand le collocatif est sélectionné sémantiquement). Par exemple, la définition du collocatif FORT dans une *forte envie* [= Magn(envie)] serait tout simplement ‹intense›. Mais comme ce sens est très général, il faut contraindre les combinaisons (**un fort chagrin, *une forte haine, *une forte surprise*). Il s'agit de formuler une contrainte sémantique sur le nom modifié de façon à exclure tous les noms d'émotions avec lesquels FORT ne peut pas se combiner. Pour y arriver, il faudrait d'abord détenir une liste fermée de tous les noms d'émotions, ensuite voir lesquels peuvent se combiner avec FORT et lesquels ne le peuvent pas. En somme, la description lexicographique de collocatifs ayant un sens abstrait et se combinant avec un très grand nombre de bases représente un travail ardu voire même impossible.

Par contre, la définition d'un collocatif décrit par une FL non standard dans l'article de la base sera plus facile à formuler étant donné qu'il ne se combine qu'avec un nombre fort restreint de bases : il y a moins d'expressions à couvrir et donc moins de composantes sémantiques à considérer. Dans de tels cas, on aboutit à une définition plus spécifique. Ainsi, par exemple, la définition de BLOND dans *cheveux blonds* a la forme suivante :

> **I.1a.** adj. [X] *blond* = [Poil ou cheveu**I.1** X] dont la couleur est similaire à la couleur du blé et qui tend vers le jaune pâle.

Comme on peut l'observer, la contrainte ‹cheveux ou poil› fait en sorte que tout nom dont le sens inclut cette composante puisse se combiner avec BLOND.

Finalement, si la description du collocatif correspond à une FL standard, on a plus de chances d'avoir besoin de créer un article de dictionnaire. Aussi il est plus difficile de consigner les bases possibles d'un collocatif quand celui-ci correspond à une FL standard que quand il correspond à une FL non standard.

3.5. Collocatifs plus autonomes *vs* moins autonomes

Ce dernier trait a plus de poids que les précédents. L'autonomie d'une lexie est le trait par excellence pour juger du statut du collocatif comme une lexie à part entière et donc, pour nous permettre de décider si ce collocatif aura droit à un article de dictionnaire. Nous mesurons l'autonomie du collocatif en vertu de trois niveaux linguistiques : sémantique, syntaxique et lexical.

Une lexie est SÉMANTIQUEMENT AUTONOME quand elle a un sens donné indépendamment des lexies avec lesquelles elle se combine. Ainsi, CAFÉ**2** signifiant 'boisson...' a ce sens peu importe s'il est en combinaison avec *boire, faire, préparer, acheter, détester ou noir*. Dans des cas comme *grain <récolte> de café* ou *comptoir <terrasse, zinc, patron> d'un café*, il s'agit de deux autres lexies différentes : CAFÉ**1** signifiant 'plante...' et CAFÉ**3** signifiant 'établissement...'. On ne doit pas s'attendre à ce que les combinaisons libres de CAFÉ**2** soient possibles avec CAFÉ**1** et CAFÉ**3**. Ils sont des lexies sémantiquement autonomes et chacun possède sa propre combinatoire libre.

Par contre, *noir* signifiant 'sans lait' n'a ce sens qu'en combinaison avec *café*. Ainsi, il n'est pas sémantiquement autonome, car il nécessite l'« appui » d'une autre lexie pour exister. Cependant tous les collocatifs ne sont pas comme *noir* dans *café noir*. Il y en a comme *conduire* qui, lui, a un sens 'diriger un véhicule...' indépendamment des mots avec lesquels il se combine : *Elmuck conduit comme un fou, J'aime bien conduire sur l'autoroute, Les autos que je conduis sont toujours des Mercedes*. Le collocatif *conduire* est donc une lexie sémantiquement autonome – il n'a pas besoin de l'appui de sa base pour signifier ce qu'il signifie.

Une lexie est SYNTAXIQUEMENT AUTONOME quand elle peut admettre tout ce qui est prévu par les règles générales de la syntaxe française, par exemple, si le collocatif peut être linéairement séparé de sa base par d'autres lexies, changer de rôle syntaxique, recevoir des modificateurs, etc. Ainsi, le collocatif *blond* dans *des cheveux blonds* est plus autonome que *bleue* dans *une peur bleue*. Le premier est admis dans la position prédicative, mais pas le deuxième :

*Ses cheveux sont blonds ~ *Sa peur est bleue.*

Quant à la modification, on observe les différences suivantes :

*des cheveux très blonds ~ *une peur très bleue.*

Une lexie est LEXICALEMENT AUTONOME si elle a ses propres corrélats lexicaux, soit paradigmatiques, soit syntagmatiques. Dans un article du DEC, c'est l'appareil des FL qui décrit les corrélats de la lexie vedette. Alors, si un collocatif a ses propres synonymes, ses propres antonymes, ses propres dérivés, ses propres intensificateurs, etc., il sera plus autonome. Par exemple, le collocatif *brun* dans *tabac brun* est plus autonome que *nature* dans *thé nature*. En effet, on peut opposer *tabac brun* à *tabac blond*, mais on ne peut pas opposer *thé nature* à **thé blanc* ou **thé lacté*. Ainsi, ce collocatif *blond* a un opposé (traité par la FL **Contr**), tandis que *nature* n'en a pas. Pour l'exprimer, il faut utiliser la paraphrase *thé au lait*, aucune expression idiomatique n'existant pour exprimer ce sens.

En somme, plus le collocatif se comporte de façon autonome, plus il possède de propriétés qu'il faut décrire et, par conséquent, il tend à accéder au statut de lexie à part entière et nous penchons en faveur de lui attribuer un article de dictionnaire.

4. Les voies de solution : comparaison de différentes stratégies de description

À l'heure actuelle, nous ne sommes pas en mesure d'énoncer des critères définitifs

pour décider s'il faut ou non créer un article de dictionnaire autonome pour un collocatif. On ne peut que constater des tendances vers l'autonomie : plus un collocatif est autonome, plus nous sommes en faveur de créer un article ; mais les cas flous sont nombreux. Ce fait reflète une propriété typique des langues naturelles : elles résistent à l'encadrement dans des cases fermées. Dans cette section, nous montrons les stratégies alternatives de description possibles et les conséquences que chacune d'elles entraîne.

Les deux positions extrêmes consistent, d'une part, à créer des articles pour TOUS les collocatifs, peu importe leurs propriétés, et d'autre part, à ne créer d'articles pour AUCUN d'eux. Examinons ces deux positions à tour de rôle :

– La première position entraîne une multiplication des articles de dictionnaire. Il est vrai qu'avec les outils informatiques actuels, la taille du dictionnaire ne crée pas trop de difficultés (bien que les informaticiens cherchent toujours la façon la plus économique de stocker l'information), mais le problème est que le dictionnaire grossit d'une façon un peu artificielle : le dictionnaire compterait beaucoup d'articles pour des lexies irréelles, *i.e.* qui n'existent qu'en combinaison avec une seule base donnée. Dans notre perspective, une lexie doit avoir une définition (c'est-à-dire un sens bien délimité), un régime particulier et une propre combinatoire pour être une lexie de plein droit. Cependant, plusieurs collocatifs se comportent de façon « parasitique »[5], *i.e.* ils s'accrochent à un autre lexème du même vocable (par ex., *noir* ‘de couleur noire’), pour pouvoir exister comme collocatifs : ils ont un sens seulement dans des contextes très restreints, ils ont une flexion restreinte, ils n'ont pas leur propre combinatoire, ou toutes ces propriétés réunies.

Créer un article de dictionnaire pour tous les collocatifs serait équivalent à dissoudre tous les phrasèmes complets en lexies séparées[6]. Si l'on force l'unicité lexicale, on pourrait arriver à dire que tout lexème faisant partie d'un phrasème a un sens particulier qui n'apparaît qu'en combinaison avec les autres constituants du phrasème. Ainsi, par exemple, le phrasème complet *bas-bleu* serait séparé en deux lexèmes, l'un signifiant ‘femme’ (mettons *bas*) et l'autre signifiant ‘à prétentions littéraires’. De cette façon, *bas* aurait ce sens seulement avec *bleu* et *bleu* aurait ce sens seulement avec *bas*. Il semble évident qu'il n'y a aucun gain descriptif dans cette analyse.

– La deuxième position (= ne jamais créer d'articles pour les collocatifs) ne peut pas être adoptée de façon cohérente, car des lexies comme *conduire, tirer (un coup de fusil), fumer, souffrir, très, beaucoup*, etc., qui s'utilisent en dehors des collocations et qui doivent avoir un article de dictionnaire, seraient privées de leur article. Cette stratégie ne prend pas en considération le fait que plusieurs collocatifs peuvent être employés aussi comme des lexies libres. Si l'on adopte cette politique, il faudrait inclure

5 Les lexèmes constituant un phrasème complet sont aussi « parasitiques » Au moment de la production du texte, le locuteur choisit les lexèmes pour constituer un phrasème, mais il le fait sans prendre en considération leur propre sens , il emprunte la forme des lexèmes qui ont une autre vie ailleurs pour servir de blocs de construction d'une nouvelle lexie, le phrasème complet Le comportement du collocatif dans une collocation est similaire. Pour produire la collocation *café noir*, le locuteur emprunte la forme du lexème existant *noir* pour lui accrocher le sens ‘sans lait’ dans la collocation prise comme un tout

6 I Mel'čuk a discuté de cette idée dans l'article suivant · Mel'čuk, I A (1960) « O terminax "ustojčivost'" i "idiomatičnost'" » [Sur les termes « locution stable » et « locution idiomatique »] *Voprosy jazykoznanija*, n° 4, 73-80 [Il existe une traduction anglaise · FDIMT (JPRS, n° 6732)]

dans l'entrée de la base beaucoup d'informations qui ne sont pas liées à la collocation : par exemple, tous les polysèmes du vocable CONDUIRE (*Le guide nous conduit à l'intérieur du musée*) devraient être consignés sous l'entrée de *véhicule* ou de *auto*, ce qui va à l'encontre de la logique même du DEC.

Il faut donc trouver des solutions moins radicales et n'accorder un article de dictionnaire qu'à certains types de collocatifs. Pour étayer tout ce qui vient d'être dit, nous procédons à la description lexicographique du verbe CLIGNER comme dans l'expression *cligner des yeux*. Il s'agit d'un collocatif unique : il ne se combine qu'avec *yeux*. Il appartient à un vocable polysémique qui n'est constitué que de lexèmes collocatifs. Il est décrit par une FL non standard. Avant d'aller plus loin dans la description de ce collocatif, examinons d'abord comment on décrit la collocation *cligner des yeux*.

Si l'on veut que le dictionnaire serve à la production de texte, l'information qu'une base donnée sélectionne un collocatif donné doit nécessairement être consignée dans l'article de la base, comme le DEC l'a toujours fait systématiquement (cf. aussi Béjoint et Thoiron, 1987 ; Benson, 1989, 1990 ; Benson *et al.*, 1986a-b) : le locuteur part du sens de la base, car celle-ci a un sens plus perceptible et il sélectionne le lexème correspondant. Ensuite, il cherche un autre lexème qui exprime un sens donné auprès de la base. Par exemple, le locuteur qui veut savoir comment se dit 'fermer et ouvrir rapidement les yeux' part du lexème ŒIL (YEUX) et cherche un verbe décrivant ce mouvement particulier des yeux. Il devrait trouver le collocatif CLIGNER, sous l'article ŒIL, ce qui nous donne la collocation *cligner des yeux*. Par conséquent, la description de la collocation est nécessairement incluse dans l'entrée de la base. Dans le DEC, elle est décrite au moyen d'une FL non standard dans l'article de ŒIL.

Nous présentons maintenant les différentes possibilités que le DEC nous offre pour décrire le collocatif CLIGNER pris seul, à savoir : 1) inclure la description du collocatif dans l'article de la base sans attribuer d'article autonome au collocatif ; 2) créer un article de dictionnaire autonome pour le collocatif et ne présenter, dans l'article de la base, que la collocation et les propriétés qui concernent la base. Les chiffres arabes encadrés ([1], [2], ...) renvoient le lecteur à certaines explications et justifications soulevées par nos descriptions. Ces commentaires figurent immédiatement après les deux propositions suivantes :

– Proposition 1 : Nous envisageons ici la possibilité de ne pas créer d'article pour le collocatif CLIGNER et, par conséquent, nous en assurons la description entière dans l'article de la base ŒIL. On obtient le résultat suivant :

ŒIL, nom, masc [pl *yeux*].
...
I.1a. *Yeux (Z) de X [permettant de voir Y]* = Deux parties du visage**I.a** d'une personne X, symétriques par rapport au nez**I.1a**, chacune étant constituée d'un globe mobile ...— organe de la vue de X, qui permet**II** à X de voir Y.
...

Fonctions lexicales

Syn	: **pop** coquillards, **pop** mirettes, **pop** quinquets
A_0	: oculaire [*globe oculaire*]

...

Mouvements et positions des yeux

les paupières couvrent
le globe oculaire : [$ART_{déf}$ ~] se fermer

fermer à demi les Y.
pour mieux voir : cligner [les / des ~] ***{1er groupe ; pas de passif}*** [1], plisser ***{1er groupe; pas de passif}*** [les ~] | O. au pl et n'a pas de dépendant [*Les myopes clignent des yeux*]

F_1 = fermer et ouvrir rapidement et involontairement plusieurs fois les Y. (— $Sympt_{21}$(*fatigue, nervosité*) ou Excess(lumière)) : cligner [2] ***{1er groupe; pas de passif}***, ciller ***{1er groupe; pas de passif}***, clignoter ***{1er groupe; pas de passif}*** [les / des ~] | O. au pl et n'a pas de dépendant [*Le soleil traversant les nuages la force à cligner des yeux*]

[3]S_0F1 : ***clignement, cillement, clignotement [des ~] | O. au pl et n'a pas de dépendant***

[4]$Oper_1(S_0F_1)$: ***// faire [des N] [Les moustiques lui faisaient faire des clignements des yeux continuels]***

[3]$Magn^{temp}F_1$: ***sans cesse***

Gestes impliquant les yeux

F_2 = en regardant une personne Y qui regarde X, X fait un signe d'entente à Y en fermant et ouvrant rapidement une fois un O. : ⌜cligner de l'~⌝ [5] ***{1er groupe}***, ⌜faire de l'~⌝ [à N = Y] | O. au sg [*Quand il est entré dans la salle, elle a cligné de l'œil à son patron*]

S_0F_2 : ***⌜clin d'~⌝ // œillade***

– Proposition 2 : Ici, nous procédons à la création d'un article de dictionnaire pour le collocatif CLIGNER. En d'autres termes, il s'agit de distribuer les informations pertinentes à la base de la collocation dans l'article de ŒIL et les informations pertinentes au collocatif lui-même dans l'article de CLIGNER. Si l'on compare avec la proposition précédente, on y constate le retrait de toutes les propriétés qui ne concernent que CLIGNER et seule la collocation comme telle demeure consignée dans l'article de ŒIL. Ainsi, par exemple, le pluriel obligatoire de ŒIL et son impossibilité d'avoir des dépendants apparaîtront dans la description de la collocation CLIGNER DES YEUX (qui figure sous ŒIL), mais les indications morphologique (1er groupe) et syntaxique (pas de passif) trouveront leur place dans l'article du collocatif CLIGNER. Ainsi les entrées de ŒIL, de CLIGNER et du quasi-phrasème CLIGNER DE L'ŒIL auront la forme suivante :

ŒIL, nom, masc [pl *yeux*].

...

I.1a. *Yeux (Z) de X [permettant de voir Y]* = Deux parties du visage**I.a** d'une personne X, symétriques par rapport au nez**I.1a**, chacune étant constituée d'un globe mobile ...— organe de la vue de X, qui permet**II** à X de voir Y.

...

Fonctions lexicales

Syn : **pop** coquillards, **pop** mirettes, **pop** quinquets

A_0 : oculaire [*globe oculaire*]

...

Mouvements et positions des yeux

les paupières couvrent le globe oculaire : [$ART_{déf}$ ~] se fermer

fermer à demi les Y. pour mieux voir : cligner**1** [les / des ~], plisser [les ~] ! O. au pl et n'a pas de dépendant [*Les myopes clignent des yeux*]

fermer et ouvrir rapidement et involontairement plusieurs fois les Y. (— $Sympt_{21}$(*fatigue*, *nervosité*) ou Excess(lumière)) : cligner**2**, ciller, clignoter [les / des ~] ! O. au pl et n'a pas de dépendant [*Le soleil traversant les nuages la force à cligner des yeux*]

Gestes impliquant les yeux

en regardant une personne Y qui regarde X, X fait un signe d'entente à Y en fermant et ouvrant rapidement une fois un O. : ⌜cligner de l'~⌝, ⌜faire de l'~⌝ [à N = Y] ! O. au sg [*Quand il est entré dans la salle, elle a cligné de l'œil à son patron*]

CLIGNER, verbe, 1er groupe, pas de passif.

1. *X cligne Y* = X ferme les yeux**I.1a** Y à demi pour mieux voir.

Régime

X = 1	Y = 2
1. N	1. N obligatoire

1) C_2 : ART = *les, des* [6]
2) C_2 : N = *yeux* [7]
$C_1 + C_2$: *Elle cligne les <des> yeux*

Fonctions lexicales

Syn : plisser

Exemples

Il regardait longuement le tableau en clignant des yeux.

2. *X cligne Y* = X ferme et ouvre rapidement et involontairement plusieurs fois les yeux**I.1a** Y sous l'effet de la lumière ou par fatigue**I.1a** ou nervosité.

Régime
Mod1

X = 1	Y = 2
1. N	1. N obligatoire

1) C_2 : ART = *les, des*
2) C_2 : N = *paupières, yeux*
$C_1 + C_2$: *Elle cligne les <des> yeux*

Mod2

Y=1
1. N

1) C_1 : N = *paupières, yeux*
C_1 : *Les yeux de Danielle clignent*

Fonctions lexicales

Syn : ciller, clignoter
S_0 : clignement [*Les moustiques lui faisaient faire des clignements des yeux continuels*]
$Magn^{temp}$: sans cesse

Exemples

Le soleil traversant les nuages la force à cligner les yeux.

CLIGNER DE L'ŒIL, loc. verbale, 1er groupe.
X ⌜*cligne de l'œil*⌝ *à Y* = En regardant une personne Y qui regarde X, X fait un signe d'entente à Y en fermant et ouvrant rapidement une fois un œil**I.1a**.

Régime

X = 1	Y = 2
1. N	1. *à* N

Fonctions lexicales

Syn : ⌜faire de l'œil⌝ [à N]
S_0 : ⌜clin d'œil⌝, œillade

Exemples

Quand Alain est entré dans la salle, Lida a cligné de l'œil à Igor.

Nous passons maintenant aux commentaires soulevés par notre démarche :

[1] Si CLIGNER n'a pas d'article autonome, il faut ajouter une description de ses propriétés morphologiques (son groupe de conjugaison) et syntaxiques (l'interdiction du passif) du collocatif dans la description de la collocation CLIGNER DES YEUX. Les informations propres au collocatif sont saisies dans une fonte différente, « ***Berling Roman*** », et figurent entre accolades, à la suite de la valeur de la FL.

[2] CLIGNER LES/DES YEUX a comme synonyme CLIGNER DES PAUPIÈRES. Strictement parlant, cette dernière doit être recensée dans l'article PAUPIÈRE. L'inconvénient de cette démarche est que l'égalité des sens n'est plus transparente : la collocation CLIGNER LES/DES PAUPIÈRES pourtant synonyme n'est plus directement disponible à partir de CLIGNER LES/DES YEUX.

[3] Si CLIGNER n'a pas d'article autonome, les valeurs correspondant aux relations paradigmatiques et syntagmatiques de la collocation entière ne peuvent être consignées ailleurs que dans l'article de ŒIL. Ainsi, les descriptions du dérivé CLIGNEMENT et du modifieur SANS CESSE deviennent en quelque sorte des « sous-entrées » de la collocation CLIGNER DES YEUX.

[4] La description du verbe support FAIRE relève de la combinatoire lexicale restreinte des substantifs CLIGNEMENT, CLIGNOTEMENT et CILLEMENT et non pas de celle de la collocation. Ces expressions ne concernent plus le collocatif CLIGNER et encore moins la lexie ŒIL dont il est question au départ. Cette façon de faire nous force à introduire des articles gigognes qui finalement n'ont pas leur place dans l'article de la lexie ŒIL. Le DEC interdit formellement cette procédure.

[5] L'expression CLIGNER DE L'ŒIL est considérée comme un quasi-phrasème, donc avec le droit à un article autonome. On y trouve le sens ‹cligner› et le sens ‹œil› et aussi un ajout de sens majeur ‹X fait un signe d'entente à Y›. À la différence de CLIGNER1 DES YEUX et de CLIGNER2 DES YEUX où le locuteur part du sens ‹œil› pour trouver un verbe qui décrive un mouvement particulier des yeux, ici le locuteur part du sens ‹signe d'entente›, le sens ‹œil› n'occupant pas une position centrale dans le sens de l'expression. Le problème qui se pose est de voir s'il faut présenter un quasi-phrasème comme valeur d'une FL. Ce procédé est contradictoire : d'une part,

les FL ne rendent compte que des collocations et jamais des quasi-phrasèmes, mais, d'autre part, étant donné que le sens ‹œil› est quand même présent dans le sens de toute l'expression, il semble nécessaire de l'introduire sous l'entrée ŒIL.

6 Par contre, les faits que ŒIL ne s'emploie qu'au pluriel et que les seuls déterminants permis sont *les* et *des* sont des propriétés qu'il faut indiquer. Ce genre de restriction n'a jamais été introduit pour les lexies de plein droit : leurs déterminants sont déduits par des règles générales de la langue. Cette nouvelle dimension de caractérisation n'est pas étonnante : les déterminants sont très souvent des victimes de la phraséologisation.

7 Il faut spécifier que le deuxième actant ‹yeux› ne se réalise en surface que par le lexème YEUX et aucun autre. Cette restriction doit être introduite pour interdire des expressions impossibles comme **cligner des mirettes <des coquillards, des quinquets>*.

Suite aux différentes solutions exposées, nous proposons une liste de vérification qui indique les points importants qu'il faut considérer lors de l'élaboration des descriptions lexicographiques des collocatifs. L'idée d'une telle liste a été proposée par Hudson (1988) puis reprise par Dostie *et al.* (1992) et par Mel'čuk *et al.* (1995). Nous nous en servons à notre tour : une telle liste est utile pour signaler au lexicographe les propriétés qui doivent être prises en considération pour décider du traitement approprié des collocatifs. Signalons que cette liste est tout à fait provisoire et qu'elle doit être utilisée avec prudence.

1. L_i, le collocatif, est-il à base unique ?
Si oui, vérifiez :
- a. s'il a des contreparties libres :
  - si oui, ne pas créer d'article (*noir, plate*) ;
  - si non, en créer un (*bissextile, cligner*).
- b. s'il permet l'héritage :
  - si oui, créer un article (*tresse blonde*) ;
  - si non, ne pas en créer un (**arabica noir*).
- c. s'il est polysémique et si tous les polysèmes sont des lexies collocatives :
  - si oui, créer un article (*cligner, blond*).

Si non, vérifiez :
- a. s'il a des contreparties libres :
  - si oui, ne pas créer d'article (*endurci*) ;
  - si non, en créer un (*invétéré*).
- b. s'il permet l'héritage :
  - si oui, créer un article (*sauce forte à partir de goût fort*) ;
  - si non, ne pas en créer un (**voleur endurci*).
- c. s'il est polysémique :
  - si oui, créer un article (*fort*).
- d. si ses bases peuvent être regroupées sémantiquement :
  - si oui, créer un article.
- e. si on peut généraliser ses propriétés :
  - si oui, créer un article.

2. L_i, le collocatif, a-t-il des contreparties libres ?
Si oui, vérifiez :
a. s'il a des propriétés particulières à l'emploi collocationnel :
si oui, créer un article.

3. L_i est-il décrit par une FL standard ?

4. À quel degré L_i est-il phraséologisé ?
Si moindre degré, créer un article (*conduire*) ;
Si degré plus élevé, vérifiez :
a. s'il a des contreparties libres, il faut déterminer la distance sémantique avec la lexie de base:
plus distant, ne pas créer d'article (*steak bleu*).

5. L_i est-il autonome ?

Remerciements

Nous ne pourrions terminer cet article avant d'avoir adressé nos remerciements à M. Igor Mel'čuk qui a lu notre article et avec qui nous avons discuté plusieurs points dans le texte. Nous lui sommes reconnaissantes de tout l'intérêt qu'il a manifesté envers notre sujet et de son enthousiasme contagieux pour le travail. Nous remercions également Alain Polguère et Agnès Tutin, nos deux victimes de ce que nous pourrions appeler du harcèlement linguistique : ils nous ont apporté des exemples et des précisions au moment où c'était nécessaire. Merci aussi à Danielle Collignon pour sa patience et sa disponibilité.

Vers un nouvel outil interactif d'aide à la conception de dictionnaires électroniques spécialisés

Christophe JOUIS[a,b] et Widad MUSTAFA-ELHADI[a]

a) UFR Information, Documentation, Information Scientifique et Technique, Université Charles De Gaulle-Lille III, France
b) Centre d'Analyse et de Mathématiques Sociales, Unité Mixte CNRS, EHESS, Université Paris-Sorbonne, France

Introduction

Concevoir un dictionnaire informatisé des termes d'un domaine de connaissance spécialisé n'est pas une tâche simple pour le concepteur, notamment lorsqu'il n'a pas ou peu de connaissances sur le domaine concerné. Il doit ménager des entrevues avec des spécialistes du domaine, puis analyser ces textes d'entrevues, les glossaires déjà existants, la documentation technique, etc. Il s'agit de concevoir une base de données terminologiques structurée dont chaque entrée (un terme du domaine) contient en particulier sa définition, son contexte d'utilisation et des pointeurs (hypertextes) vers d'autres entrées du dictionnaire sous forme de liens sémantiques (synonyme, hyponyme, partie/tout, localisations spatiales, temporelles, etc.). Autrement dit, le concepteur doit se construire une représentation du domaine, sa structure conceptuelle et les relations que les concepts entretiennent entre eux à partir d'un ensemble de documents textuels.

Pour analyser ces documents textuels, nous proposons une méthode linguistique et informatique : l'exploration contextuelle. Cette méthode est fondée sur le noyau de connaissances linguistiques du concepteur qui analyse des textes sans connaissances préalables sur le domaine concerné. Ce modèle se focalise non pas sur les termes spécifiques du domaine (traditionnellement appelé « mots pleins »), mais sur les connecteurs (les mots « vides » ou mots « pivots ») entre les termes. Ces derniers constituent un ensemble d'indicateurs qui sont chargés de significations. Ils traduisent *un savoir linguistique indépendant d'un domaine de connaissance particulier.*

Ainsi, nous partons de l'hypothèse suivante : *les textes contiennent des unités linguistiques qui sont des indicateurs pertinents pour structurer des connaissances.*

Le concepteur s'appuie sur ces unités linguistiques ; il utilise une stratégie d'exploration contextuelle lorsqu'il analyse les documents.

Le système SEEK[1] est une application de l'exploration contextuelle qui se présente sous la forme d'un système interactif à base de connaissances. Le système isole *d'abord* des relations sémantiques dans des textes entre les concepts d'un domaine. Puis, *dans un deuxième temps*, c'est-à-dire lorsqu'une relation est détectée, SEEK recherche les termes (concepts), arguments de la relation.

Nous testons actuellement une deuxième version de SEEK qui a comme objectif la détection des contextes définitoires à l'aide de marqueurs linguistiques.

1. Schéma de principe de construction d'un dictionnaire électronique spécialisé

Notre objectif consiste à l'élaboration d'un outil d'aide interactif visant à la construction d'un dictionnaire informatique des termes spécifiques à un domaine de connaissance à partir d'un corpus de textes décrivant le domaine en question. La structure d'un tel dictionnaire électronique peut être vue comme une base de données terminologiques dont les entrées sont les termes du domaine. Chaque enregistrement de la base décrit un terme du domaine. La description est alors composée de liens hypertextes vers des zones textuelles du corpus et de liens étiquetés vers d'autres termes (ou enregistrements) de la base (figure 1). Plus précisément, la structure d'un enregistrement est alors formée : (i) de liens hypertextes vers les contextes définitoires du terme dans le corpus (voir § 3) et (ii) d'une liste de pointeurs vers d'autres enregistrements de la base : des relations sémantiques (figure 2).

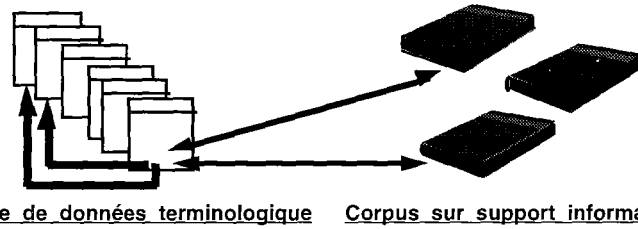


FIGURE 1 : Base de données terminologiques : liens hypertextes et liens étiquetés

Notre approche prend ses sources dans un modèle global de traitement du langage : la Grammaire Applicative et Cognitive (GAC[2]). La GAC articule plusieurs niveaux de représentations, et en particulier un niveau cognitif où l'on analyse les significations des unités linguistiques sous forme de représentations sémantico-cognitives afin de construire les représentations des connaissances associées à un texte. La GAC propose un ensemble de concepts sémantiques qui définissent un système organisé de significations. Nous distinguons les **types sémantiques** des unités linguistiques,

1. SEEK : Système Expert d'Exploration (K)contextuelle
2. Pour une description de la GAC, les auteurs renvoient par exemple à Desclés (1990), Jouis (1993).

des **relations statiques** fondamentales et des **relations évolutives** (mouvement, changement d'état, conservation d'un mouvement, itération, variation d'intensité, contraintes, causes...).

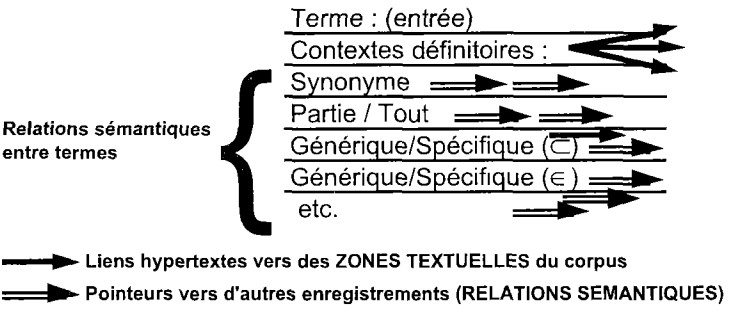


FIGURE 2 Structure d'un enregistrement.

Dans la suite, nous ne détaillerons qu'une partie des primitives de la GAC : les types sémantiques et les relations statiques car ce sont ces primitives qui sont pour le moment utilisées dans SEEK.

1.1. Les types sémantiques

Nous distinguons un certain nombre de types élémentaires permettant de classer les entités manipulées[3].

- Les entités individualisables sont celles que l'on peut désigner et montrer par pointage. On peut les compter individuellement ou les regrouper en classes. Par exemple, les entités *Jean, table, chaise, meuble, homme, enfant* sont individualisables.
- Les entités massives telles que *eau, mer, sable, vin, beurre, blé* ne sont pas des entités individualisables. Notons cependant qu'un certain nombre d'opérateurs (des classificateurs) permettent de rendre individualisable une notion massive : un *verre d'eau, un pâté de sable, un bras de mer, une bouteille de vin, un morceau de beurre.*
- Les classes distributives rassemblent des entités individuelles ayant une même propriété. Par exemple, *être-un-carré* représente une classe d'individus (ou « concept »).
- Les classes collectives se distinguent des entités individualisables parce qu'elles représentent des objets qui forment un « tout » à partir d'objets plus élémentaires. Ainsi, *foule, armée, armada, famille* sont des classes collectives.
- Le type des lieux représente des étendues ou des regroupements de positions d'une même entité (individualisable, collective ou massive) : *Paris, jardin, maison.*

3 Dans la GAC, nous pouvons construire des types plus complexes à partir des types élémentaires en utilisant des opérateurs formateurs de type : listes, n-uplets, types fonctionnels, etc. (Desclés, 1990).

1.2. Les relations statiques

Les relations statiques sont hiérarchisées et indépendantes d'un domaine particulier. Ce sont des relations binaires. Les relations statiques permettent de décrire les situations statiques du domaine, qui restent stables pendant un certain intervalle temporel où ni début, ni fin ne sont envisagés. Nous distinguons plus d'une vingtaine de relations statiques, en particulier :

- les identifications de deux entités,
- les incompatibilités entre deux entités,
- les dimensions (mesures, etc.),
- les cardinalités,
- les comparaisons de valeurs,
- les inclusions entre classes distributives (relations générique/spécifique),
- les appartenances d'une entité individuelle à une classe distributive (relations générique/spécifique),
- les relations « partie/tout » entre classes collectives,
- les localisations d'une entité par rapport à un lieu (intérieur, extérieur, frontière, fermeture, orientations, etc.).

La sémantique de chaque relation est définie par trois types de propriétés : (i) son type fonctionnel (le type sémantique des arguments de la relation) ; (ii) ses propriétés algébriques : réflexivité, symétrie, transitivité, etc. ; (iii) ses propriétés d'agencement avec les autres relations dans un même contexte. Les relations statiques s'insèrent dans un système de significations des relations de repérage entre entités[4].

En général, les systèmes de repérage de termes et de relations sémantiques entre termes sont fondés sur des approches syntaxiques, statistiques et/ou connexionnistes. Aussi, ils nécessitent pour fonctionner un dictionnaire général de la langue et parfois un lexique des termes techniques ainsi que dans certains cas une représentation sémantique préexistante spécifique au domaine de connaissances à modéliser. Dans notre problématique, c'est justement ce dernier point que nous cherchons à construire de manière semi-automatique (outil d'aide interactif).

2. Notre proposition : appliquer la méthode d'exploration contextuelle

2.1. Présentation générale de l'exploration contextuelle[5]

Pour l'analyse rapide de textes en grand nombre l'exploration contextuelle vise à construire des représentations sémantiques en se contentant d'une analyse syntaxique ap-

4 Le repérage, noté REP (ou « ε »), est un schéma général de relation : une entité X (une entité repérée) est repérée par rapport à Y (une entité repère). Le repérage se spécifie suivant les propriétés algébriques qui lui sont attribuées axiomatiquement en divers relateurs. Sur ce point, voir Desclés (1987).

5 L'exploration contextuelle a initialement été appliquée pour le traitement du temps et de l'aspect : un module de détection des valeurs sémantiques des temps de l'indicatif en français a été réalisé. Nous ne détaillerons pas ce module qui est décrit par ailleurs : Oh *et al.* (1992), Jouis (1993). D'autres applications de l'exploration contextuelle sont actuellement réalisées : résumé automatique (Leroux, Minel et Berri, 1994), analyse de la causalité (Garcia et Jackiewicz, 1995).

proximative. L'exploration contextuelle repère certains marqueurs jugés pertinents dans les phrases analysées qui deviennent des « pivots » pour établir des relations entre concepts.

D'une façon générale, un système d'exploration contextuelle se ramène à : (i) identifier des indices linguistiques pertinents (pour le problème à résoudre) ; (ii) définir un système de valeurs sémantiques ; (iii) expliciter un ensemble de règles de décision qui permettent d'associer à des cooccurrences d'unités linguistiques la valeur sémantique adéquate en fonction du contexte ; (iv) organiser l'ensemble sous forme d'un système de prise de décision sur la valeur sémantique.

L'exploration contextuelle a pour objectif de simuler une « lecture » rapide d'un texte consistant à rechercher des indices linguistiques qui permettent de construire un réseau de concepts. La « compréhension » superficielle d'un texte aboutit à la prise de décisions fondées sur le repérage des indices (marqueurs linguistiques de relations sémantiques) coprésents dans le texte. Un système d'exploration contextuelle se ramène à un ensemble de règles déclaratives qui expriment un savoir décisionnel interprétatif. Les règles (d'exploration contextuelle) se représentent sous la forme **SI** <conditions> **ALORS** <actions> ou <conclusions>. Les conditions des règles expriment la coprésence ou non d'unités linguistiques pertinentes dans le contexte. Ces indices touchent plusieurs composantes simultanément : morphologique, syntaxique, lexicale. Les conclusions de l'ensemble des règles permettent de construire progressivement des représentations sémantiques.

2.2. SEEK : une application de l'exploration contextuelle

SEEK permet, *dans un premier temps*, de rechercher des relations statiques dans des textes entre les objets d'un domaine de compétence[6]. Il produit une représentation visuelle sous forme de graphes objets/relations. Des liens hypertextes permettent de retrouver, pour chaque relation trouvée, la zone du texte ayant servi à sa construction. SEEK se présente sous la forme d'un système à base de connaissances. Il fonctionne à l'aide de règles d'exploration contextuelle, dont le but est de rechercher dans les textes des indices textuels qui permettent d'identifier une relation statique particulière, puis les arguments de la relation.

Par exemple, parmi les listes d'indices concernant la description des composants d'un objet dans un domaine (relation partie/tout ou « ingrédience ») nous avons la règle suivante, associée aux listes de marqueurs linguistiques VDecomp1, VDecomp2, VDecomp3[7] :

---

6 La base de connaissances de SEEK concernant la détection des relations statiques est composée d'une base de données de marqueurs statiques (quelques 3 300 marqueurs classés dans 240 listes) et de 220 règles d'exploration contextuelle. Cette première version de SEEK a été testée dans un contexte industriel (EDIAT/CGI, ALSTHOM/IPSé). Par ailleurs, il est soumis à une évaluation dans le cadre d'une action de recherche concertée (ARC) soutenue par l'AUPELF-UREF « Évaluation des systèmes de construction automatique de terminologie et de relations sémantiques entre termes à partir de corpus »

7 Ces indices sont utilisés dans le système SEEK (Jouis, 1993). Les listes et les règles d'exploration contextuelle concernant la relation partie-tout ont été identifiées par Agata Jackiewicz-Desbertrand *Contribution aux problèmes de l'extraction des connaissances : manifestations linguistiques et représentation informatique de la relation d'ingrédience*, Mémoire de DEA Sciences Cognitives, EHESS, 1992. Pour des raisons de place, nous ne donnerons qu'une partie du contenu des listes de marqueurs linguistiques.

SOIT	x1, x2 des unités linguistiques ; P une proposition
SI	x1 est un marqueur de la liste **VDecomp1** OU **VDecomp2** OU **VDecomp3** ET x2 est l'unité linguistique « **en** » ET x1 x2 se suivent (pas forcément immédiatement) dans la même proposition P
ALORS	Proposer la relation **partie/tout** dans P

VDecomp1 = {analyser, décomposer, démembrer, désagréger, désassembler, désintégrer, désunir, détailler, disjoindre, dissocier, réduire, séparer...} ;
VDecomp2 = {bissecter, cloisonner, compartimenter, diviser, émietter, fragmenter, graduer, morceler, parceller, partager, scinder, sectionner, subdiviser, tronçonner...} ;
VDecomp3 = {couper, découper, déchirer, fendre, trancher, casser, disloquer, rompre...}.

Repérés dans un énoncé tel que :

> (...) la zone d'accessibilité se **décompose en** deux partie : la zone d'accessibilité principale et la zone d'accessibilité secondaire.

ces indices permettent d'orienter la décision vers la relation partie/tout. Ensuite, dans cet énoncé (où une relation a été détectée), SEEK tente d'identifier les arguments (termes) de la relation. Dans une proposition où une relation est détectée, les composants d'un argument de la relation (ou terme) (figure 3) :

- n'appartiennent pas à une liste de marqueurs statiques ;
- ne sont ni des verbes, ni des adverbes, ni des pronoms, ni des prépositions, ni des « mot-phrases », ni des conjonctions, etc.

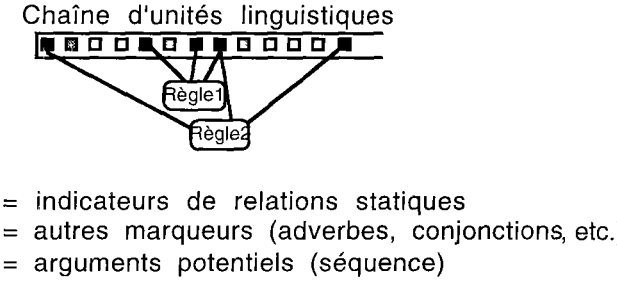


FIGURE 3 : Identification des termes par contraste.

Ainsi, grâce à la règle d'exploration contextuelle de détection d'argument ci-dessous :

SOIT	x1, x2, x3, x4 unités lexicales ; S une proposition
SI	une relation statique est détectée dans S ET x1 est l'unité linguistique [le] ET x2 est un terme primitif potentiel ET x3 est l'unité linguistique [de] ET x4 est un terme primitif potentiel ET x1 x2 x3 x4 se suivent immédiatement dans S
ALORS	Proposer x1+x2+x3+x4 comme argument dans S

on identifie l'argument [la zone d'accessibilité], comme argument possible pour la relation partie/tout. Finalement, on aboutit, après l'enchaînement de plusieurs règles puis validation par l'utilisateur, à une représentation sous forme de graphe, où sont mémorisées à l'aide de liens les zones textuelles qui ont permis de construire la représentation (figure 4).

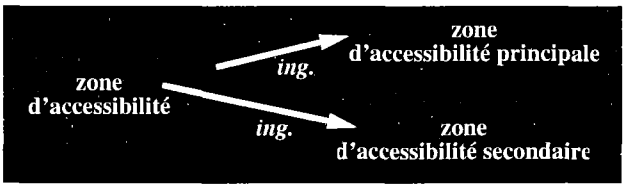


FIGURE 4. Représentation issue de SEEK

3. Extension de SEEK à la détection des contextes définitoires

Dans la perspective de la construction d'une terminologie à partir de textes de taille importante, une question se pose alors : où chercher les relations et les termes pertinents dans les textes ? Il s'agit de donner la définition de chaque concept qui est repérable dans : (i) des **zones de textes éparpillées** dans lesquelles SEEK aura repéré les relations statiques et les termes (les arguments de la relation) ; (ii) mais aussi des **zones textuelles privilégiées** associées à un terme dans lesquelles se trouvent rassemblées des informations concernant ce terme : ses **contextes définitoires.**

3.1. La définition en terminologie spécialisée

Les concepts constituent la base de toute construction terminologique. Un concept est une unité de connaissance qui comprend des propositions vérifiables concernant un élément de référence choisi, exprimé par un *terme*. La terminologie s'occupe des concepts et en conséquence des structures de la connaissance dans la mesure où celles-ci sont exprimées par le lexique des langues. Les unités lexicales ne deviennent termes que si elles sont *définies* et employées dans un contexte donné. Une théorie sur la terminologie est nécessairement liée à une structure référentielle[8] qui met en évidence les relations entre les termes et leurs référents et permet une définition correcte de l'ensemble des éléments constitutifs d'une structure référentielle.

La définition en terminologie a pour objet la délimitation du concept dans un réseau conceptuel d'un domaine donné et sert à exprimer la signification d'un terme. Ce n'est pas le terme qui fait l'objet de la définition, c'est plutôt le concept désigné par le terme. Définition, contexte, formules (mathématiques, chimiques, etc.), figures, représentations graphiques, images, etc. sont tous considérés comme constituant une information sur le concept.

A. Rey (1979 : 40) fait remarquer le lien fondamental entre le terme et sa définition. Il précise qu'ils désignent à l'origine « l'assignation d'une limite, d'une fin (*définir*) et son résultat (*terme*) ». Même si les problèmes de la définition (que ce soit en

8 Voir Dahlberg (1978, 1981)

terminologie ou en lexicologie) ont été abordés dans diverses publications[9], le champ de ces études reste largement limité aux traitements intellectuel et manuel. Peu de publications portent sur les indices de reconnaissance automatique et/ou semi-automatique des définitions.

La question que nous posons ici est de savoir comment le système doit procéder afin de repérer la triade *terme-concept-définition* lors d'une construction automatique d'une terminologie. Autrement dit, quels sont les marqueurs linguistiques susceptibles d'aiguiller le système vers les termes et les définitions ? Nous postulons que certains marqueurs serviraient à la fois pour détecter les termes et les définitions : *...se définit par..., ...se nomme..., ...c'est-à-dire....*, par exemple. Le système doit reconnaître dans une même proposition le terme et sa définition.

Nous retiendrons la distinction faite par Béjoint (1993) entre les « définitions textuelles » qu'on trouve dans les textes, manuels, articles, etc., et « les définitions dictionnairiques » qu'on trouve dans les dictionnaires. Les premières nous intéressent dans une optique de recherche de marqueurs linguistiques que SEEK va utiliser pour détecter les définitions dans les corpus. Les deuxièmes nous intéresseront en tant que « modèle ». Autrement dit, comment doit-on « aider » SEEK à compléter « idéalement » les définitions qu'il aura repérées automatiquement ?

La définition doit retenir les caractéristiques essentielles du concept et permet ainsi de le distinguer à l'intérieur d'un système conceptuel reflétant un domaine de connaissance. La littérature sur la définition « idéale » en terminologie ne manque pas. L'ISO a établi un certain nombre de règles concernant les définitions, (voir Rondeau, 1981 : 84). Mais nous savons que les règles ne sont pas toujours respectées dans les dictionnaires spécialisés. La définition par le *genre proche* et la *différence* ou la définition aristotélicienne semble être la mieux adaptée à la terminologie, car elle rend compte des relations hiérarchiques qui existent entre les concepts. Il faut signaler cependant que ce type de définition n'est pas toujours de mise ; car il y a des concepts qui s'y prêtent mal. Et c'est un des problèmes qui caractérisent la définition des concepts appartenant à des domaines techniques, ceux appartenant aux sciences exactes poseraient moins de problèmes.

Un autre problème que pose la définition par le genre et la différence est celui du choix du genre et en particulier comme nous l'avons signalé, quand il s'agit d'un domaine technique ou les genres proches s'incluent les uns les autres. Ce problème a été aussi soulevé par Béjoint (1993) qui établit une typologie de définitions génériques et montre comment leur « généricité » est douteuse.

3.2. Règles pour le repérage des termes

3.2.1. Reformulation et définition

Comme nous l'avons déjà signalé, le système doit reconnaître dans une même proposition le terme et sa définition. Nous avons établi des règles pour le repérage de termes

9 Voir Actes du colloque *La définition*, organisé par CELEX (Centre d'étude du lexique) de l'Université Paris-Nord, 1988 , *Problèmes de la définition et de la synonymie en terminologie*, Actes du Colloque international de terminologie, Université Laval, Québec mai 1982, Sager (1990 et 1982), entre autres

et les relations entre concepts. Nous exposons dans cette partie les marqueurs appropriés à la détection des contextes définitoires.

En dehors de marqueur linguistique générique tel que *définir*[10], qui détecte formellement la présence de définitions, d'autres marqueurs peuvent être retenus. Il s'agit d'éléments linguistiques qu'on trouve dans les *reformulations* (Chukwu et Thoiron, 1989 ; Thoiron et Béjoint, 1991 ; Chukwu, 1993 ; Otman, 1989). La *reformulation* peut être définie comme le processus qui permet de passer d'une forme linguistique à une autre forme porteuse de la même signification, ou de mettre en rapport des variantes d'expression que l'on souhaite considérer comme équivalentes (voir Dachelet, 1990). Pour Chukwu et Thoiron (1989 : 24) la reformulation pourrait être mise en relation avec d'autres concepts voisins tels que *paraphrase, définition, synonymie, équivalence, anaphore* et *description*. Tout en montrant la parenté entre définition et reformulation les auteurs proposent de distinguer les définitions (celles qui font référence à des documents antérieurs faisant autorité tels que les normes, les dictionnaires, les citations, etc.) des simples reformulations qui ne sont pas toujours « suffisantes » pour prétendre à des définitions à part entière (voir aussi à ce propos Chukwu, 1993 : 266).

3.2.2. Les marqueurs susceptibles de détecter des définitions

Pour établir les listes de marqueurs selon leur nature nous nous sommes largement inspirés de la classification donnée par Chukwu et Thoiron (1989), Thoiron et Béjoint (1991), Chukwu (1993), Otman (1989).

i) Les marqueurs apparaissant dans les reformulations copulatives identifiées dans Chukwu et Thoiron (1989) et Chukwu (1993) constituent une catégorie largement repérable par les marqueurs de relations statiques dans SEEK (voir la liste *supra*).

ii) Les marqueurs qui apparaissent dans les reformulations métalinguistiques appellatives (directes et inverses) identifiées dans Chukwu et Thoiron (1989). Ce sont des verbes qui appartiennent d'après les auteurs au champ sémantique de la dénomination (*appeler, nommer, désigner,* etc.) ou des expressions telles que *c'est-à-dire, autrement dit* qui signalent une explication.

iii) Les marqueurs apparaissant dans les reformulations explicatives, ou encore dans les reformulations définitoires, catégorie identifiée par Chukwu et Thoiron (1989) et Chukwu (1993), semblent être une catégorie productive pour le repérage de définitions. Chukwu et Thoiron reconnaissent à cette catégorie une fonction explicative délibérée « [...] le terme est introduit dans le discours et appelle immédiatement une définition, une explication (c'est nous qui soulignons) ou tout autre procès qui en facilite la compréhension » (Chukwu et Thoiron, 1989 : 28).

iv) Les marqueurs apparaissant dans les reformulations métalinguistiques explicatives regroupent aussi des marqueurs tels que *c'est-à-dire* ainsi que les verbes faisant ré-

10 Otman qualifie les marqueurs tels que *définir, appeler*, etc. comme étant des marqueurs « autosuffisants » pour assurer le repérage de *contextes définitoires et informatifs*. Ce terme (Otman, 1989 : 68) regroupe d'après l'auteur les notions de « contexte définitoire », « contexte encyclopédique », « contexte associatif », « contexte langagier » et « contexte métalinguistique », classification reprise par l'auteur à Boutin-Quesnel *et al.* (1985).

férence à des processus cognitifs tels que *entendre, comprendre* (Chukwu et Thoiron, 1989 : 36). Nous pouvons ajouter à cette liste aussi d'autres marqueurs que SEEK a déjà utilisé pour repérer les relations statiques *diviser, classer, distinguer, énumérer, signifier, désigner, représenter, au sens où*, etc.

De plus, nous ajoutons à ces marqueurs les listes que Otman (1989 : 71-72) a établies pour *dire, appeler, définir.*

– dire : *on dit que . , on dira que , on dit parfois que ..est, si ... on dit que c'est ., on dit que est .... si . , on dit de qu'ils sont..., nous dirons dans ce cas qu'il y a..., nous disons que . , on veut dire que. , est dit . , ...sont dits/dites , . .dits/dites, l'une dite l'autre dite... ,*
– appeler : *on appelle ., c'est ce qu'on appelle , . constitue/ent ce qu'on appelle .., dans le ...que nous appelons..., nous pourrions donc l'appeler..., ...est appelé(e)/sont appelés(es)..., l'appellation vient de ... ,*
– définir : *nous définissons ici les ...comme des ..., nous définissons que nous appelons..., est défini par ..., ...définit..., par définition...* (Listes non exhaustives).

3.3. Diversité de contextes définitoires et présence de facettes

La présence de plusieurs contextes définitoires pourrait, dans certains cas, être l'indicateur d'une facette. Autrement dit, le terme aura plus qu'une définition et par conséquent plusieurs contextes définitoires selon le point de vue ou facette sous lequel il a été envisagé. Il serait intéressant d'introduire des représentations multidimensionnelles ou à facettes du concept donné (voir Dahlberg, 1993 ; Sager, 1990 ; Mustafa-Elhadi, 1989).

3.4. Construction interactive d'une définition

La définition d'un concept est repérable dans des zones de textes éparpillées dans lesquelles SEEK aura déjà repéré les relations statiques et les termes et aussi des zones textuelles privilégiées associées à un terme dans lesquelles se trouvent rassemblées des informations concernant ce terme : ses contextes définitoires. Ces derniers sont repérables, comme pour les relations statiques, par exploration contextuelle grâce à divers indicateurs linguistiques.

À partir d'une compilation de l'ensemble des contextes définitoires il s'agit de construire ensuite une définition. Les règles d'exploration contextuelle proposent des contextes définitoires, mais il revient au spécialiste du domaine de les valider, de les intégrer et éventuellement de les compléter.

En effet, la définition terminologique doit faire apparaître les caractéristiques essentielles sans apporter des développements inutiles ou superflus. La plupart des auteurs sur la définition s'accordent sur le caractère exhaustif de celle-ci, car l'information contenue dans les définitions peut concerner divers utilisateurs potentiels qui puiseront selon leur besoin dans ce fond commun : « La définition de spécialité est une définition voulue exhaustive et dans le même temps restreinte à un seul domaine. [...] Aussi la longueur de la définition de spécialité qui est une définition nécessairement longue – ne présenterait aucun inconvénient, chaque utilisateur limitant sa recherche à la seule portion pour lui nécessaire de la définition » (Roman, 1993 : 116).

4. Conclusions

Nous avons proposé une démarche linguistique et informatique pour l'aide à la conception de bases de données terminologiques qui s'appuie sur un *savoir linguistique indépendant d'un domaine de connaissance particulier* pour repérer les définitions des concepts dans des zones de textes éparpillées. Cette démarche se focalise d'abord sur l'identification de relations statiques entre termes puis sur les termes eux-mêmes. Enfin, il s'agit de tenir compte de zones textuelles privilégiées associées à un terme dans lesquelles se trouvent rassemblées des informations concernant ce terme : ses *contextes définitoires*. Ces derniers sont repérables, comme pour les relations statiques, par exploration contextuelle grâce à divers indicateurs linguistiques.

Néanmoins, de nombreux problèmes restent en suspend : quelle serait la définition idéale ? Cette question nous conduit nécessairement à poser le problème de corpus. Quels sont les corpus « idéaux » pour une construction (semi) automatique d'une terminologie ? Est-ce que la nature des corpus (corpus descriptifs, manuels didactiques, etc.) est un paramètre déterminant quant aux performances du système ? Y a-t-il une typologie des définitions par domaine ?