Chapitre premier

Méthodologie[1]

Étant donné l'importance de l'analyse quantitative et du traitement informatique en linguistique appliquée, et considérant la nature des caractères arabes et la complexité des structures lexicales de la langue arabe, il fallait s'armer d'une méthodologie de travail qui puisse nous permettre de surmonter la difficulté de l'inadaptation de la langue aux traitements informatiques ou, plutôt, de ceux-ci à la langue[2]. Tout au long de notre étude, l'analyse théorique formelle sera accompagnée d'une analyse quantitative. À cet effet, l'une des tâches essentielles qui attend le chercheur en matière de terminologie consiste en la description du corpus de la langue. Pour parvenir à accomplir cette tâche, il faut s'outiller d'un dispositif d'analyse. Le dispositif d'analyse comprend généralement une méthode, des paramètres, un médium et un objet d'analyse.

---

1 Ce livre s'adresse d'abord aux spécialistes de la langue arabe. Les linguistes non initiés au système linguistique arabe auront avantage à lire les parties B et C du chapitre 2 avant d'entreprendre la lecture de la méthodologie.

2 L'inadaptation de la langue arabe aux traitements informatiques en est une de forme seulement. En effet, il n'y a pas encore, à notre connaissance, mis à part les programmes de traitement de texte et de bases de données, de programmes d'analyse textuelle capables de traiter des caractères et des structures schémiques de la langue arabe.

I. Méthode d'analyse

Au cours de l'élaboration de notre méthode d'analyse, nous avons suivi un certain nombre d'étapes pour surmonter les problèmes techniques posés par le traitement informatique de la langue arabe. Pour ce faire, il était impératif de mettre sur pied un système de codage assez complexe. Le codage du corpus a nécessité l'élaboration de grilles d'analyse, de codes correspondants aux paramètres d'analyse et d'un tableau de conversion pour rétablir les données dans leur forme originale une fois le travail d'analyse terminé. L'approche méthodologique s'est déroulée en trois étapes : (1) procédures préparatoires, (2) traitement, (3) analyse quantitative.

A. Procédures préparatoires

Cette étape a consisté en l'apprêt du matériel linguistique brut pour l'analyse. Elle s'est déroulée en trois phases : (1) élaboration de grilles d'analyse, (2) codage, (3) saisie.

1. Les grilles d'analyse

Le codage des données nécessite l'emploi d'un support approprié pour consigner les codes. À cet effet, et après une lecture minutieuse des données, nous nous sommes fixé deux types de grilles : (1) une grille d'analyse de la dénomination et (2) une grille d'analyse de la notion.

a. Grille d'analyse de la dénomination

La grille d'analyse de la dénomination doit représenter une structure morphosyntaxique à projection maximale, c'est-à-dire la structure morphosyntaxique la plus longue. Par ailleurs, elle doit être capable d'englober les structures les plus courtes et les structures discutables. En effet, une structure courte ne pose pas de problème dans la mesure où elle s'intègre à un des modèles de formation syntagmatique arabe. Par contre, des formations du type [ʃalä al-bärid] = *à froid*, qui ne sont d'ailleurs pas nombreuses, ne peuvent pas être omises. Nous pensons que de telles formations ne sont ni des termes simples, ni des termes complexes et posent, de ce fait, un problème d'acceptabilité. Elles constituent vraisemblablement des traits notionnels qui doivent être rattachés à une tête syntagmatique plutôt que des UTC. La solution à ce problème réside dans la considération, au niveau de la grille d'analyse, de ce genre de structures.

La grille est constituée de composants principaux variables et de composants secondaires fixes. Les composants variables concernent les substantifs, les adjectifs, les verbes et les adverbes. Les composants fixes concernent les articles et les joncteurs. Les composants ont été consignés dans des champs. Chaque champ constitue une colonne de la grille. Ainsi, nous aurons des champs variables et des champs fixes. Les champs variables sont au nombre de six, puisque le nombre maximal de composants (sans considérer l'article et les joncteurs) dans une UTC est de l'ordre de six. À cela, nous ajoutons un champ initial réservé au numéro de l'UTC et au code du domaine. Tous les champs variables ont une longueur de quatre caractères. Par ailleurs, le champ initial a cinq caractères. Les champs fixes sont au nombre de douze (six pour l'article et six pour les joncteurs). La longueur des champs d'article est d'un caractère alors que la longueur des champs de joncteurs est de deux caractères (cf. schéma suivant) :

Champs	→	Num	J	D	C1	J	D	C2	J	D	C3	J	D	C4	J	D	C5	J	D	C6				
Valeurs	→	5	2	1	4	2	1	4	2	1	4	2	1	4	2	1	4	2	1	4				
Paramètres	→	x	y	z	w	y	z	w	y	z	w	y	z	w	y	z	w	y	z	w				

L'ensemble des valeurs des champs donne une étendue potentielle de quarante-sept caractères que nous dénommons *étendue des paramètres.*

b. Grille d'analyse de la notion

La grille d'analyse de la notion est répartie elle aussi en champs variables et en champs fixes. Les champs variables et les champs fixes sont précédés par un champ initial (champ d'identification du terme). Le champ initial (cinq caractères) est réservé au numéro de l'UTC et au code du domaine. Ce champ initial est le même pour la grille d'analyse de la dénomination que pour la grille d'analyse de la notion. C'est par le biais du champ initial que nous en arrivons à établir les relations nécessaires entre les structures morphologiques et morphosyntaxiques d'une part, et les structures notionnelles d'autre part. En ce qui concerne les champs variables, ils sont au nombre de sept. Un premier champ, composé de deux caractères, est réservé à la catégorie notionnelle générale de l'UTC. Pour ce qui est des six autres champs, ils sont formés de deux caractères chacun et sont réservés aux traits notionnels

véhiculés par les composants syntagmatiques. Les champs fixes concernent les joncteurs. Dans la grille d'analyse de la dénomination, ces champs, en plus d'indiquer la présence de joncteurs, indiquent aussi le type de joncteurs. Par contre, dans la grille d'analyse de la notion, ces champs indiquent seulement la présence ou l'absence du joncteur et, pour cette raison, ils comprennent un seul caractère. En fait, nous avons jugé utile de maintenir le champ des joncteurs car nous pensons qu'il est important de distinguer, dans les structures notionnelles, les structures qui s'établissent par relations implicites et celles qui s'établissent par relations explicites. Cependant, dans presque tous les cas où il y a une structure notionnelle construite par relation explicite, c'est-à-dire avec joncteur, il y en a qui sont construites par relation implicite. Cela a posé un problème au niveau de la représentation des structures notionnelles. En fait, si l'on devait faire la séparation entre relations explicites et relations implicites pour une même structure notionnelle, on arriverait à une augmentation considérable du nombre de structures notionnelles. Étant donné notre souci de systématisation et de représentation du plus grand nombre d'UTC par le moins possible de structures notionnelles, nous avons dû recourir à la parenthétisation des joncteurs. Cette opération suit le modèle suivant : supposons que nous ayons affaire à deux UTC dont les structures notionnelles sont (1) ACT + OBJ et (2) ACT + J + OBJ, nous dirions alors qu'il existe une structure (3) de laquelle dérivent (1) et (2). Dans ce cas, nous réserverons l'étiquette de structure à (3) et nous donnerons à (1) et (2) la dénomination *d'allostructure*[1]. Par ailleurs, une structure peut avoir plus de deux allostructures. En effet, certaines structures présentent une parenthétisation[2] de deux ou trois joncteurs, ce qui constitue une économie de plusieurs allostructures. Dans sa forme la plus simple, ce type de représentation peut être formulé comme suit :

(1) {ACT + OBJ} ∪ (2) {ACT + J + OBJ}⇒ (3) {ACT + [J] + OBJ}

Pour revenir donc à la grille d'analyse de la notion, nous avons quatorze champs qui totalisent vingt-cinq caractères. Nous pourrions représenter cette grille comme suit :

---

1 Nous appelons *allostructure*, toute variété qui ne diffère de la structure principale que par la présence ou l'absence de la reltion explicite.

2 La parenthétisation dans la structure devrait se lire avec ou sans joncteur.

Champs →	Num	CN	J	C1	J	C2	J	C3	J	C4	J	C5	J	C6
Valeurs →	5	2	1	2	1	2	1	2	1	2	1	2	1	2
Paramètres →	x	v	y	w	y	w	y	w	y	w	y	w	y	w

2. Codage

Nous avons signalé certains problèmes techniques reliés au traitement informatique de l'écriture arabe. En effet, à supposer que l'on transcrive les termes arabes en vue d'une application infor-matique, l'étendue des problèmes ne s'en trouve pas pour autant éliminée, et cela pour des raisons multiples. Tout d'abord, le système d'écriture de l'arabe étant de nature phonétique, le système de traitement terminotique ou textuel doit être capable d'offrir une chaîne de caractères appropriés. À la rigueur, une adjonction des caractères phonétiques manquants serait suffisante pour remédier au problème de la transcription. De plus, ce système de traitement doit être programmé pour pouvoir reconnaître, dans les mots, les racines, schèmes et les morphèmes (les marques de duel, de pluriel et de genre). En fait, il ne peut y avoir d'étude terminotique sans la réalisation de ces conditions de base. En raison de l'absence d'un tel système, nous avons dû accomplir tout ce travail de manière artisanale pour pouvoir procéder au traitement terminotique.

En matière d'analyse de la notion, le problème n'est pas de même nature dans la mesure où la difficulté, qui est toujours d'actualité, réside dans la mobilité sémantique des schème. Par ailleurs, nous pensons que ce problème n'est pas propre à l'arabe seulement, mais concerne toutes les autres langues. En effet, s'il est aujourd'hui possible de prédire dans certains programmes informatiques la catégorie syntaxique d'un constituant, on est encore loin de pouvoir en prédire la catégorie sémantique. Pour cette raison, la structure notionnelle ne peut être déterminée, à notre connaissance, de manière automatique. Le problème est d'autant plus aigu en langue arabe puisque, jusqu'à maintenant, les études sur le syntagmes terminologiques sont rares et les études sur les structures notionnelles sont inexistantes. Dans cette perspective, ce genre d'analyse constitue un terrain vierge qui devrait être exploré en profondeur.

Ce bref aperçu des problèmes amène à distinguer deux niveaux de codage : le codage de la dénomination et le codage de la notion.

a. Codage de la dénomination

Le codage de la dénomination a été effectué à trois ni-veaux : codage morphologique, codage schémique et codage morphosyntaxique.

(1) Codage morphologique

Le codage morphologique consiste en l'attribution d'un code particulier à la catégorie à laquelle appartient un constituant syntagmatique donné. Tous les codes de catégorie syntaxique des constituants des UTC ont reçu une forme alphabétique. Ainsi, nous avons établi S pour substantitf, A pour adjectif, V pour verbe, T pour adverbe[1]. Tous ces symboles occupent la première position dans les champs que nous avons qualifiés de variables, puisque nous pourrions y rencontrer toutes ces catégories syntaxiques. Ainsi, ces symboles qui figurent dans les champs variables servent à identifier la catégorie syntaxique à laquelle appartient chaque constituant. Les trois autres caractères du champ sont réservés au deuxième type de codage.

(2) Codage schémique

Après avoir identifié la catégorie syntaxique des constituants, nous nous sommes penché sur l'identification, à l'intérieur des catégories syntaxiques, des types de formes. Cette partie de l'analyse couvre deux aspects combinés en un seul code de type numérique. Nous distinguons, dans la catégorie des substantifs, les types suivants : NA, S-PA, S-PP, NI, NL, ND, S-AR et AFS. À l'intérieur des cinq premiers types, nous distinguons les schèmes selon la quantité consonantique et selon l'origine verbale, alors que pour le reste des types, ils sont différenciés selon la provenance ou la nature. Nous ne faisons pas une analyse schémique des NP (car non dérivés). Les S-AR sont classés selon leur provenance seulement, car l'étude schémique de ces formes, en plus de sa non-utilité, risque de disperser les résultats. En définitive, l'étude schémique ne porte que sur les dérivés de première dérivation. Le codage des différents types de substantifs et des différents schèmes a été fait par le biais de codes numériques ordinaux allant de 1 à n. La reconnaissance des différents types s'est faite sur la base de la

1 Dans la présentation des résultats, ce symbole est remplacé par *Ad.*

répartition séquentielle. Par exemple, nous dirons que la séquence de S001 à S062 constitue une séquence de NA et que la séquence de S066 à S076 constitue une séquence de S-PA, etc. À l'intérieur de chaque séquence, nous établissons des sous-séquences qui font que l'on peut identifier les schèmes dérivés de F-n. Nous dirons, par exemple, que la sous-séquence de S001 à S034 constitue une sous-séquence de la séquence des NA dont les schèmes sont dérivés de F-I, et ainsi de suite. Dans l'exposé de nos paramètres, nous utiliserons les codes contenus dans notre tableau de conversion afin de faciliter la tâche à nos lecteurs. Par ailleurs, un certain nombre de codes ont dû être éliminés à la fin de l'analyse en raison de l'inexistence d'attestation. En effet, comme nous ne pouvions pas faire le codage et le relevé des schèmes en même temps vu le nombre considérable de schèmes. Il a fallu faire une recherche étoffée des études portant sur les schèmes afin d'en regrouper le maximum possible. Suite à cette recherche, nous avons entrepris une opération de classification, après quoi nous avons codé ces schèmes. Les schèmes regroupés ont permis de couvrir toute la portée schémique du corpus. Enfin, les schèmes qui n'ont pas reçu d'attestation ont été tout simplement éliminés à la fin de l'analyse. Nous pourrions représenter les séquences et les sous-séquences de la manière suivante :

$S \rightarrow \{NA, S\text{-}PA, S\text{-}PP, NI, NL, S\text{-}AR, AFS\}$

$NA \rightarrow \{SCH^T, SCH^Q\}$

$SCH^T \rightarrow \{SCH^T_{-aff}, SCH^T_{+aff}\}$

$SCH^T_{-aff} \rightarrow \{x, y, n\}$

$SCH^T_{+aff} \rightarrow \{SCH^T_{1aff}, SCH^T_{2aff}, SCH^T_{3aff}\}$

SCH^Q....

S-PA....

Nous avons procédé de la même manière avec les adjectifs. Nous distinguons les A-PA, les A-PP, les AN, les AR et les AFA. Sur ces cinq types d'adjectifs, seuls les AR et les AFA n'ont pas fait l'objet d'analyse schémique. Quant aux autres types, ils ont été codés au moyen de codes numériques ordinaux par type d'adjectif, par schème et par provenance verbale. Les AR et les AFA ont été codés selon leur provenance seulement.

Les verbes ont été codés selon que leur schème est trilitère ou quadrilitère et selon qu'ils sont affixés ou non ; et dans le cas des schèmes verbaux affixés, selon le nombre d'affixes.

Les adverbes, qui ne sont d'ailleurs pas nombreux, ont reçu des numéros d'ordre identifiant chaque type. Par ailleurs, en ce qui concerne les joncteurs, nous avons seulement attribué des numéros d'ordre sans spécification de catégorie syntaxique puisqu'ils ont un champ propre à eux, ce que nous avons appelé champ fixe. Enfin, en ce qui concerne l'article, comme il n'y en a qu'un en arabe (champ D), et comme il bénéficie d'un champ fixe qui lui est propre, il est représenté par les codes binaires (0) ou (1) qui indiquent respectivement son absence ou sa présence.

(3) Codage morphosyntaxique

Le codage morphologique et le codage morphosyntaxique sont intimement liés. Alors que le premier type de codage porte sur l'analyse des constituants de l'UTC, le deuxième type de codage ne constitue que la somme linéaire des différents codages morphologiques des constituants d'une UTC donnée. Notre but était d'établir non seulement les structures morphosyntaxiques, mais aussi les structures morphosyntacticoschémiques. Nous nous sommes malheureusement résigné à ne pas le faire en raison de la taille importante des paramètres schémiques face à laquelle notre corpus, malgré son importance, s'avère limité pour ce genre d'entreprise. En effet, nous estimons qu'une analyse plus profonde des structures morphosyntacticoschémiques nécessiterait un corpus de dix fois la taille du nôtre.

b. Codage de la notion

Le codage de la notion a deux aspects. Dans le champ suivant le premier champ qui est réservé au numéro de l'UTC et au code de domaine, nous effectuons un codage de la catégorie notionnelle de l'UTC (cf. champ CN dans la grille d'analyse de la notion). Les catégories notionnelles sont les mêmes que les traits notionnels qui s'adressent aux champs des constituants. La différence réside dans la perspective. En effet, quand on parle de l'UTC en tant que totalité, on parle de notion ou de catégorie notionnelle ; quand on parle de constituant syntagmatique, on parle de trait notionnel. De ce fait, à toute UTC correspond une UNC (unité notionnelle

complexe); et toute UNC peut être UNC-1N (UNC à un trait notionnel), UNC-2N, etc. Les codes attribués aux catégories notionnelles (ou traits notionnels) sont de type numérique. Ainsi, ACT = (01), ETA = (02), OBJ = (03), etc.

Nous ne pourrions pas mettre un terme à l'examen du codage de la notion sans parler de certains problèmes. En effet, nous avons rencontré deux problèmes essentiels : (1) trait notionnel à accorder à l'anthroponymisme et (2) trait notionnel à accorder à certains termes métaphoriques. Pour le premier cas, nous pensons qu'il ne constitue pas un problème insoluble. En fait, l'anthroponymisme est presque toujours déterminant d'un outil, procédé ou objet. De ce fait, il a toutes les caractéristiques d'une propriété. Pour le deuxième cas, plus ardu celui-ci, puisque nous ne pouvons pas nous fier à la valeur apparente des constituants, nous proposons deux solutions : soit un traitement à part (si le nombre d'occurrences est élevé), soit un traitement par traduction intralinguistique ou synonymique (si le nombre d'occurrences est infime). Par ailleurs, la plupart des termes métaphoriques sont analysables en terme de structure notionnelle, à condition que la métaphore ne touche qu'un composant du groupement binaire. En fait, il faudrait distinguer deux types de métaphore: la métaphore partielle et la métaphore filée ou totale. Dans le premier cas, l'analyse en structure notionnelle est possible : par exemple, *arme blanche* → OBJ = OBJ + PRT. Par contre, dans le deuxième cas, l'analyse en structure notionnelle n'est possible que par traduction synonymique: par exemple, *oeil de poisson* → défaut de soudage → PRT = PRT + ACT. C'est d'ailleurs là le seul exemple de métaphore filée que nous avons rencontré dans notre corpus et que nous avons résolu de cette manière afin de ne pas provoquer une désorganisation dans nos données statistiques. Au demeurant, la solution adoptée a été motivée par une équation sémique. En effet, la reconstruction de la structure profonde nous permet de voir une métaphore originale suivie d'une série d'effacements : *défaut de soudage en forme d'oeil de poisson* → *défaut de soudage en oeil de poisson* → *défaut en oeil de poisson* → *oeil de poisson*. Nous constatons une identité logique qui finit par s'établir entre *défaut* et *oeil de poisson*. Cette identité nous permet, au plan sémantique, d'établir l'équation suivante : *défaut (= : oeil de poisson) de soudage*. Nous pensons que la solution n'est pas complètement

satisfaisante, mais elle s'avère pratique pour traiter d'un cas isolé. En outre, nous pensons que la question de la métaphore filée en terminologie mérite une étude à part afin d'en déterminer les aspects et de trouver une méthode pour la représenter.

Enfin, pour clore cette partie portant sur le codage de la dénomination et de la notion, nous prendrons un exemple de codage pour illustrer les différents types de codage. Considérons l'exemple suivant : FI-01156 Ar. furn tajfīf taḥt al-tafrīġ (Fr. four à sécher sous vide, En. vacuum drying oven).

codage morphologique	codage schémique	
C1 = S	099	= nom d'emprunt
C2 = S	050	= ta-C_1C_2ïC_3
C3 = S	050	= ta-C_1C_2ïC_3

La schématisation des codages de la dénomination et de la notion nous donne la figure ci-dessous, suivie d'un exemple de conversion de codes. Sous chaque ligne de codes, nous indiquerons les conversions appropriées. On remarquera qu'au niveau de la conversion, les valeurs nulles ne figurent pas (exemple : absence de joncteur). Par ailleurs, l'article, même s'il est présent, a été omis dans la conversion parce qu'il s'est avéré inutile pour nos besoins d'analyse.

Num	CN	J		C1	J		C2	J		C3
↓	↓	↓		↓	↓		↓	↓		↓
C0553	04	0		04	0		01	1		02
↑		↑	↑	↑	↑	↑	↑	↑	↑	↑
FI11-56		0	0	furn	0	0	taʒfif	taḥta	a l-	tafrīġ
↓		↓	↓	↓	↓	↓	↓	↓	↓	↓
C0553		00	0	S099	00	0	S050	16	1	S050
↑		↑	↑	↑	↑	↑	↑	↑	↑	↑
Num		J	D	C1	J	D	C2	J	D	C3

Exemple de codage de la dénomination :

codage : C0553000S099000S050161S050

conversion : FI-1156 = S(NE) + S(NA) + J(16) + S(NA)

↑ ↓ ↑

F-II tahta F-II

Exemple de codage de la notion :

codage : C055304004001102

conversion : TS-1156 = OUT + ACT + J + ETA

Remarquons enfin que l'analyse morphosémantique, c'est-à-dire l'établissement d'un pont de correspondance entre les schèmes et les traits notionnels, se fait par l'intermédiaire du numéro de l'UTC qui est le dénominateur commun aux niveaux de la dénomination et de la notion.

3. Saisie des données

Maintenant que nos données sont mises en codes alphanumériques, elles sont prêtes pour la saisie et le traitement terminotique.

La saisie a été effectuée par le Centre de traitement de l'information de l'Université Laval. Nous avions en tout dix fichiers à saisir. En effet, notre corpus était constitué des UTC de cinq dictionnaires techniques totalisant 4671 syntagmes. Chaque dictionnaire a dû subir une analyse de la dénomination et une analyse de la notion, ce qui nous donne deux fichiers par dictionnaire. En somme, l'étude a porté sur 9342 unités que nous répartissons en 4671 UTC auxquelles correspondent 4671 UNC. Au niveau des constituants, le corpus est composé de 2596 adjectifs, 8248 substantifs, 44 adverbes, 40 verbes et 806 joncteurs, soit un total de 11 734 mots. À ces 11 734 mots correspondent 11 734 traits notionnels (ou mots sémantiques), soit, enfin, un total général de 23 468 unités de codage correspondant soit à des mots ou à des traits notionnels. Ces chiffres n'incluent bien sûr pas les codes de l'article et les codes égaux à zéro.

B. Le traitement terminotique

Le traitement terminotique a été effectué avec le logiciel dBase III plus. Nous avons effectué des comptages et des tris de formes par champ. Nous avons ensuite procédé à un traitement séquentiel des données. Le traitement séquentiel consiste en plusieurs étapes. Ces étapes ont consisté en un premier classement par nombre de constituants, puis par type de constituants, puis par présence ou absence de joncteurs. Ces étapes peuvent être représentées suivant les règles de réécriture ci-dessous.

Nous avons suivi le même principe d'analyse pour le traitement de la notion, avec la différence qu'au lieu de commencer par un classement selon le nombre de constituants, nous avons commencé par un classement selon la catégorie notionnelle.

(1) UTC ⇒ { UTC-1C, UTC-2C, UTC-nC }

(2) UTC-1C ⇒ J { A, S }

(3) UTC-2C ⇒ sans J { SS, SA, etc. }

⇒ avec J { SJS, SJ, V }

(4) UTC-nC etc..

Une fois toutes les structures établies et comptabilisées, nous avons procédé à un autre type de classement. Pour les structures morphosyntaxiques, nous avons groupé séparément les UTC-1S, les UTC-2S, les UTC-3S, etc. Pour les structures notionnelles, il fallait regrouper ce que nous avons appelé allostructures pour réduire le nombre de structures obtenues. Par ailleurs, toutes les

structures notionnelles sont présentées selon la catégorie notionnelle, selon le nombre de traits notionnels et selon le type de structure.

C. Analyse quantitative

L'analyse quantitative constitue une partie importante de l'étude des Lsp. En effet, selon Alber-DeWolf (1984 : 10),

> ce qui distingue essentiellement les Lsp de la Lc se manifeste par des choix de distribution plutôt que par des modèles spécifiques. On peut donc appliquer des méthodes d'étude quantitative à ces caractères spécifiques.

L'étude quantitative nous a permis de remarquer des zones de concentration significative, tant au niveau de la morphologie et de la sémantique des constituants qu'à celui des structures morphosyntaxiques et notionnelles. Il s'agit en fait de déceler, par l'analyse quantitative, les schèmes ou formes et structures productives dans un domaine donné.

Bauer (1983 : 63) distingue trois niveaux d'analyse: (1) la productivité, (2) la créativité et (3) le rapport productivité / créativité. Elle définit la productivité comme étant «one of the defining features of human language, and is that property of language which allows a native speaker to produce an infinitely large number of sentences, many (or most) of which have never been produced before».

Pour ce qui est de la créativité, elle est définie (toujours selon Bauer) par « [...] the native speaker's ability to extend the language system in a motivated, but unpredictable way (non-rule-governed) way». Enfin, c'est par le rapport productivité/créativité que les néologismes naissent: «Both productivity and creativity give rise to large numbers of neologisms [...]».

Pour ce qui est du rendement d'une forme ou d'une structure donnée, Bauer le situe aux confluents de la syntaxe et de la morphologie d'une part, et de la synchronie et de la diachronie d'autre part. Bauer (1983 : 100) indique que «[...] a morphological process can be said to be more productive according to the num-ber of new words which it is used to form». La même chose peut être dite des procédés morphosyntaxiques. En effet, plus grand est le nombre d'UTC que l'on peut classer sous une structure morphosyntaxique donnée, plus grandes sont ses chances de devenir

productive et représentative d'un domaine donné. Il en va de même pour les structures notionnelles. Il y a plusieurs moyens d'évaluer la productivité d'une structure donnée. Il y a tout d'abord la fréquence et il y a aussi ce que nous appelons l'indice de répartition.

Par ailleurs, nous avons constaté que les études portant sur les syntagmes n'ont pas encore trouvé de méthode de quantification des constituants syntagmatiques. Nous nous sommes penché sur la question afin d'en trouver une appropriée qui serve à la fois à quantifier les constituants et à déterminer les frontières statistiquement significatives des formes complexes. Notre tentative de quantification des constituants nous a amené à l'établissement d'une unité de mesure que nous appelons indice de dispersion. Nous nous proposons donc d'examiner les notions de fréquence, d'indice de répartition et d'indice de dispersion.

1. Fréquence

La fréquence est le nombre total des occurrences d'une forme donnée. Il y a plusieurs types de fréquences. Nous distinguons, d'une part, la fréquence absolue (FA) et la fréquence relative (FR), et, d'autre part, la fréquence en langue et la fréquence en discours.

a. Fréquence absolue et fréquence relative

La FA est le nombre total d'occurrences reliées à une structure ou forme données dans un corpus donné. Par contre, la FR est le nombre total d'occurrences reliées à une structure ou forme données dans une tranche déterminée d'un corpus donné. En effet, pour une structure S + S, on peut avoir le nombre total d'occurrences sur la totalité du corpus ou sur le groupe des UTC à deux substantifs. Par ailleurs, un schème a une FA au niveau du corpus et des FR au niveau des champs (des constituants). Il en va de même des structures notionnelles et des traits notionnels.

b. Fréquence en langue et fréquence en discours

Selon Alber-DeWolf (1984 : 75), la fréquence en langue est le nombre d'unités lexicales formées, dans une langue donnée, sur un modèle morphologique ou morphosyntaxique donné. Par contre, la fréquence en discours est le rapport entre le nombre d'occurren-ces et un modèle de formation donné dans un ensemble de communications scientifiques et techniques (CST). Notre étude se situe au

niveau de la langue étant donné que notre corpus est de nature terminographique et non pas textuelle.

2. L'indice de répartition (IR)

Nous appelons indice de répartition la façon dont une structure ou un schème est distribué à l'intérieur d'un corpus donné. Nous distinguons l'IR moyenne (IRM) et l'IR réelle (IRR).

a. L'indice de répartition moyenne

Que ce soit pour l'analyse de répartition de structure ou de schème, l'IRM constitue une unité de mesure étalon. L'IRM suppose que tout est partout égal. Par ailleurs, l'IRM sert à calculer l'IRR.

$$\text{IRM de structure} = \frac{\text{Nombre total d'occurrences}}{\text{Nombre total de structures}}$$

$$\text{IRM de SCH} = \frac{\text{Nombre total de constituants}}{\text{Nombre total de SCH}}$$

b. L'indice de répartition réelle

L'IRR est la distribution réelle d'une structure ou d'un schème dans le cadre d'un corpus donné. Le calcul de l'IRR se fait par le rapport entre le nombre total d'occurrences par structure ou par schème et l'IRM de structure ou de schème.

$$\text{IRR de structure} = \frac{\text{Nombre total d'occurrences par structure}}{\text{IRM de structure}}$$

$$\text{IRR de SCH} = \frac{\text{Nombre total d'occurrences par SCH}}{\text{IRM de SCH}}$$

Pour qu'une structure ou un schème donné ait une productivité moyenne, il faut que son IRR soit égal à 1, c'est-à-dire que le nombre d'occurrences par structure ou par schème soit égal à l'IRM de structure ou de schème.

3. L'indice de dispersion (ID)

L'ID est une unité de mesure qui se spécialise dans l'évaluation de la distribution des constituants sur une chaîne syntagmatique. Nous distinguons l'ID moyenne (IDM) et l'ID réelle (IDR).

a. L'indice de dispersion moyenne

Nous appelons IDM la distribution moyenne d'un schème ou forme par champ syntagmatique. Le champ syntagmatique est l'espace qu'occupe un constituant syntagmatique. L'IDM est calculé par la division de l'IRM par le nombre total de champs occupés par un schème ou une forme quelconque. Supposons que, pour un schème donné, nous ayons un IRM égal à 120 et que le schème en question occupe 4 champs, nous dirons que l'IDM du schème est de 120 sur 4, soit de 30. En fait, si l'IRM est le même pour tous les schèmes et formes, l'IDM varie selon le nombre de champs occupés par le schème ou la forme en question. Cette variation vise à une juste évaluation de la dispersion schémique. L'IDM ainsi calculé sert par la suite au calcul de l'IDR des formes.

$$\text{IDM} = \frac{\text{IRM}}{\text{Nombre de champs/SCH}}$$

b. L'indice de dispersion réelle

C'est par l'IRD que l'on arrive à une évaluation exacte de la dispersion d'une forme donnée dans le cadre d'un corpus donné. L'IDR est calculé en divisant le nombre total d'occurrences totalisées par un schème ou une forme dans un champ donné par l'IDM. Une dispersion réelle moyenne doit être égale à 1, c'est-à-dire que le nombre total d'occurrences par schème et par champ doit être égal à l'IDM. De ce fait, nous sommes capable de prédire, par l'IDR, les chances d'apparition d'une forme donnée dans un champ donné. Par ailleurs, nous pouvons déterminer la taille moyenne des UTC du corpus puisque tout IDR inférieur à 1 peut être considéré comme faible ou moyennement faible, et, donc, peut être rejeté comme étalon de la taille syntagmatique.

$$\text{IDR} = \frac{\text{Nombre total d'occurrences/SCH/champ}}{\text{IDM}}$$

II. Paramètres d'analyse

Les paramètres sont d'ordres structurel et sémantique. Nous distinguons les paramètres d'analyse de la dénomination et les paramètres d'analyse de la notion.

A. Paramètres d'analyse de la dénomination

Les paramètres d'analyse de la dénomination peuvent être regroupés en cinq classes : les verbes, les substantifs, les adjectifs, les adverbes et les joncteurs. Remarquons toutefois que les paramètres que nous présentons ici sont ceux qui sont attestés dans notre corpus.

1. Les verbes

Les schèmes verbaux attestés sont les suivants :

F-I	[$C_1aC_2aC_3a$]
F-II	[$C_1aC_2C_2aC_3a$]
F-IV	[?a-$C_1C_2aC_3a$]
F-IX	[?in-$C_1aC_2aC_3a$]
F-XII	[ta-$C_1aC_2C_2aC_3a$]
F-XIII	[ta-$C_1aC_2aC_3a$]
F-XVIII	[?ista-$C_1C_2aC_3a$]

2. Les substantifs

Nous distinguons huit types de substantifs : les NA, les S-PA, les S-PP, les NI, les NL, les ND, les S-AR et les AFS.

a. Les noms d'action (NA)

Les NA sont classés selon les schèmes verbaux dont ils dérivent ou selon leur origine et selon la quantité schémique du verbe générateur (c'est-à-dire schème trilitère (SCH^T) / schème quadrilitère (SCH^Q) affixés ou non affixés [= nus]).

(1) NA dérivés des verbes trilitères

Nous distinguons les NA dérivés du verbe trilitère nu et les NA dérivés des verbes trilitères affixés. Pour faciliter l'observation des rapports génératifs, nous indiquerons, dans une première colonne, le code de forme verbale génératrice ; puis, dans une deuxième colonne, le code du schème du NA (ce code sera utilisé tout au long de l'étude pour indiquer le schème du NA approprié); dans une troisième colonne enfin, nous indiquerons le schème du NA à proprement parler.

(a) NA dérivés du verbe trilitère nu (F-I)

Les NA dérivés du verbe trilitère nu (F-I) sont au nombre de trente et un.

F-I → NA 1 = [$C_1aC_2C_3$]
F-I → NA 2 = [$C_1aC_2C_3ä?$]
F-I → NA 3 = [$C_1aC_2aC_3$]
F-I → NA 4 = [$C_1aC_2äC_3$]
F-I → NA 5 = [$C_1aC_2iC_3$]
F-I → NA 6 = [$C_1aC_2ïC_3$]
F-I → NA 7 = [$C_1aC_2üC_3$]
F-I → NA 8 = [$C_1iC_2C_3$]
F-I → NA 9 = [$C_1iC_2aC_3$]
F-I → NA10 = [$C_1iC_2äC_3$]
F-I → NA11 = [$C_1uC_2C_3$]
F-I → NA12 = [$C_1uC_2aC_3$]
F-I → NA13 = [$C_1uC_2äC_3$]
F-I → NA14 = [$C_1uC_2üC_3$]
F-I → NA15 = [$C_1aC_2aC_3än$]
F-I → NA16 = [?u-$C_1C_2üC_3$]
F-I → NA17 = [$C_1aC_2C_3a$-t]
F-I → NA18 = [$C_1aC_2aC_3a$-t]
F-I → NA19 = [$C_1aC_2äC_3a$-t]
F-I → NA20 = [$C_1aC_2iC_3a$-t]
F-I → NA21 = [$C_1aC_2ïC_3a$-t]
F-I → NA22 = [$C_1aC_2iC_3C_3a$-t]
F-I → NA23 = [$C_1aC_2üC_3a$-t]
F-I → NA24 = [$C_1iC_2C_3a$-t]
F-I → NA25 = [$C_1iC_2äC_3a$-t]
F-I → NA26 = [$C_1uC_2C_3a$-t]
F-I → NA27 = [$C_1uC_2äC_3a$-t]
F-I → NA28 = [$C_1uC_2üC_3a$-t]
F-I → NA29 = [ma-$C_1C_2aC_3a$-t]
F-I → NA30 = [ma-$C_1C_2aC_3$]
F-I → NA31 = [ma-$C_1C_2iC_3$]

(b) NA dérivés des verbes trilitères affixés

Les NA dérivés des verbes trilitères affixés sont au nombre de dix-neuf. Nous pourrions les répartir en trois classes: les NA dérivés des verbes trilitères à un affixe, les NA dérivés des verbes trilitères à deux affixes et les NA dérivés des verbes trilitères à trois affixes.

(i) NA dérivés des verbes trilitères à un affixe

F-II → NA32 = [ta-$C_1C_2ïC_3$]
F-II → NA33 = [ta-$C_1C_2iC_3a$-t]
F-II → NA34 = [ti-$C_1C_2äC_3$]
F-II → NA36 = [$C_1uC_2C_2äC_3a$-t]
F-III → NA37 = [mu-$C_1äC_2aC_3a$-t]
F-IV → NA38 = [?i-$C_1C_2äC_3$]
F-IV → NA39 = [?i-$C_1C_2äC_3a$-t]
F-VI → NA40 = [ma-$C_1C_2aC_3a$-t]
F-VII → NA41 = [C_1a-w-C_2aC_3a-t]
F-VIII → NA35 = [C_1aC_2-w-aC_3]

(ii) NA dérivés des verbes trilitères à deux affixes

F-IX → NA42 = [?in-$C_1iC_2äC_3$]
F-X → NA43 = [?i-C_1-ti-$C_2äC_3$]
F-X → NA44 = [$C_1iC_2C_2äC_3$]
F-XI → NA45 = [?i-$C_1C_2iC_3äC_3$]
F-XII → NA46 = [ta-$C_1aC_2C_2uC_3$]
F-XIII → NA47 = [ta-$C_1äC_2uC_3$]
F-XIII → NA48 = [ta-$C_1äC_2iC_3$]
F-XV → NA49 = [tama-$C_1C_2uC_3$]
F-XVII → NA50 = [ta-C_1aC_2-wu-C_3]

(iii) NA dérivés des verbes trilitères à trois affixes

F-XVIII → NA51 = [?isti-$C_1C_2äC_3$]

(2) Les NA dérivés des verbes quadrilitères

En raison de leur petit nombre, les NA dérivés des verbes quadrilitères seront regroupés dans le même paragraphe. Nous distinguons cinq formes :

F-XXI → NA52 = [$C_1aC_2C_3aC_4$a-t]
F-XXI → NA53 = [$C_1aC_2C_3äC_4$]
F-XXI → NA54 = [$C_1iC_2C_3äC_4$]
F-XXI → NA55 = [ta-$C_1aC_2C_3uC_4$]
F-XXI → NA56 = NA52 [1]

b. Participes actifs substantivés (S-PA)

Comme pour les NA, les S-PA seront classés selon le schème du verbe générateur. En raison du petit nombre de leurs structures, les S-PA seront exposés dans le même paragraphe. Les schèmes de S-PA sont au nombre de treize :

F-I → S-PA 1 = [$C_1äC_2iC_3$]
F-II → S-PA 2 = [mu-$C_1aC_2C_2iC_3$]
F-II → S-PA 3 = [$C_1aC_2C_2äC_3$]
F-III → S-PA 4 = [mu-$C_1äC_2iC_3$]
F-III → S-PA 5 = [$C_1äC_2üC_3$]
F-IV → S-PA 6 = [mu-$C_1C_2iC_3$]
F-X → S-PA 7 = [mu-C_1-ta-C_2iC_3]
F-XII → S-PA 8 = [muta-$C_1aC_2C_2iC_3$]
F-XIII → S-PA 9 = [muta-$C_1äC_2iC_3$]
F-XVIII → S-PA10 = [musta-$C_1C_2iC_3$]
F-XXI → S-PA11 = [mu-$C_1aC_2C_3iC_4$]
F-XXII → S-PA12 = [muta-$C_1aC_2C_3iC_4$]
F-XXI → S-PA13 = S-PA12 [2]

c. Les participes passifs substantivés (S-PP)

Les schèmes des S-PP attestés dans notre corpus sont au nombre de huit :

F-I → S-PP 1 = [ma-$C_1C_2üC_3$]
F-II → S-PP 2 = [mu-$C_1aC_2C_2aC_3$]

---

1 NA56 a la même forme que NA52. Nous lui avons réservé un code à part parce que NA56 indique que le NA est dérivé d'un emprunt intégré.

2 S-PA13 a la même forme que S-PA12. Nous avons dû le classer à part parce qu'il s'agit d'un S-PA dérivé d'un emprunt intégré.

F-IV	→	S-PP 3	= [mu-$C_1C_2aC_3$]
F-VII	→	S-PP 4	= [mu-C_1a-w-C_2aC_3]
F-IX	→	S-PP 5	= [mun-$C_1aC_2aC_3$]
F-X	→	S-PP 6	= [mu-C_1-ta-C_2aC_3]
F-XI	→	S-PP 7	= [mu-$C_1C_2aC_3C_3$]
F-XVIII	→	S-PP 8	= [musta-$C_1C_2aC_3$]

d. Noms d'instruments (NI)

Les schèmes de NI sont au nombre de deux :

F-I	→	NI 1	= [mi-$C_1C_2aC_3$]
F-I	→	NI 2	= [mi-$C_1C_2\ddot{a}C_3$]

e. Noms de lieux (NL)

Les schèmes de NL sont au nombre de deux :

F-I	→	NL 1	= [ma-$C_1C_2aC_3$]
F-I	→	NL 2	= [ma-$C_1C_2iC_3$]

f. Noms diminutifs (ND)

Dans la catégorie des ND, nous avons repéré un seul schème :

F-I	→	ND	= [C_1uC_2a-y-C_3]

g. Adjectifs relatifs substantivés (S-AR)

Les S-AR sont au nombre de quatre. Nous les distinguons selon leur provenance et non selon leur schème.

S-AR 1 = S [AR [NE + iyy]]
S-AR 2 = S [AR [NA + iyy]]
S-AR 3 = S [AR [NP + iyy]]
S-AR 4 = S [AR [NN + iyy]]

h. Autres formes substantivales (AFS)

Dans la catégorie des AFS, nous distinguons onze formes:

AFS 1 = Nom abstrait (NAB)
AFS 2 = Nom primaire (NP)
AFS 3 = Nom composé (NC)
AFS 4 = Nom d'emprunt (NE)
AFS 5 = Constituant alphabétique arabe (CAA)
AFS 6 = Constituant alphabétique d'emprunt (CAE)
AFS 7 = Constituant alphanumérique arabe (CANA)
AFS 8 = Nom hybride (NH)
AFS 9 = Adjectif neutre substantivé (S-AN)
AFS10 = Joncteur substantivé (S-J)
AFS11 = Suffixe pronominal (SPR)

3. Les adjectifs

Nous répartissons la catégorie des adjectifs en cinq classes: la classe des A-PA, la classe des A-PP, la classe des AN, la classe des AR et la classe des AFA.

a. Les participes actifs adjectifs (A-PA)

Nous avons relevé treize schèmes de A-PA. Nous les présenterons selon les schèmes des verbes générateurs dont ils sont dérivés:

F-I	→	A-PA 1	= [C_1ăC_2iC_3]
F-II	→	A-PA 2	= [mu-C_1aC_2C_2iC_3]
F-III	→	A-PA 3	= [mu-C_1ăC_2iC_3]
F-IV	→	A-PA 4	= [mu-C_1C_2iC_3]
F-VI	→	A-PA 5	= [muma-C_1C_2iC_3]
F-IX	→	A-PA 6	= [mun-C_1aC_2iC_3]
F-X	→	A-PA 7	= [mu-C_1-ta-C_2iC_3]
F-XII	→	A-PA 8	= [muta-C_1aC_2C_2iC_3]
F-XIII	→	A-PA 9	= [muta-C_1ăC_2iC_3]
F-XVIII	→	A-PA10	= [musta-C_1C_2iC_3]
F-XXI	→	A-PA11	= [mu-C_1aC_2C_3iC_4]
F-XXII	→	A-PA12	= [muta-C_1aC_2C_3iC_4]
F-XXI	→	A-PA13	= [A-PA11][1]

b. Les participes passifs adjectifs (A-PP)

Les schèmes des A-PP attestés dans notre corpus sont au nombre de onze :

F-I	→	A-PP 1	= [ma-C_1C_2ūC_3]
F-II	→	A-PP 2	= [mu-C_1aC_2C_2aC_3]
F-III	→	A-PP 3	= [mu-C_1ăC_2aC_3]
F-IV	→	A-PP 4	= [mu-C_1C_2aC_3]
F-VII	→	A-PP 5	= [mu-C_1a-w-C_2aC_3]
F-X	→	A-PP 6	= [mu-C_1-ta-C_2aC_3]
F-XI	→	A-PP 7	= [mu-C_1C_2aC_3C_3]
F-XII	→	A-PP 8	= [muta-C_1aC_2C_2aC_3]
F-XVIII	→	A-PP 9	= [musta-C_1C_2aC_3]
F-XXI	→	A-PP10	= [mu-C_1aC_2C_3aC_4]
F-XXI	→	A-PP11	= [A-PP11][2]

---

1 A-PA13 a la même forme schémique que A-PA11. La différence réside dans le fait que A-PA13 est un emprunt intégré.

2 A-PP12 a la même forme schémique que A-PP11. Nous faisons la distinction parce que A-PP12 est un emprunt intégré.

c. Les adjectifs neutres (AN)

Les AN sont au nombre de onze. Ils sont tous dérivés de la première forme verbale, c'est-à-dire trilitère nue (F-I):

F-I	→	AN 1	= [C_1ăC$_2$iC$_3$]
F-I	→	AN 2	= [C_1aC$_2$C$_3$]
F-I	→	AN 3	= [C_1iC$_2$C$_3$]
F-I	→	AN 4	= [C_1uC$_2$C$_3$]
F-I	→	AN 5	= [C_1aC$_2$aC$_3$]
F-I	→	AN 6	= [C_1aC$_2$iC$_3$]
F-I	→	AN 7	= [C_1aC$_2$ïC$_3$]
F-I	→	AN 8	= [C_1aC$_2$üC$_3$]
F-I	→	AN 9	= [?a-C_1C_2aC$_3$]
F-I	→	AN10	= [C_1aC$_2$ăC$_3$]
F-I	→	AN11	= [C_1uC$_2$aC$_3$]

d. Les adjectifs relatifs (AR)

Les AR sont au nombre de douze. Ils sont classés selon leur provenance:

AR 1	=	AR [AN + iyy]
AR 2	=	AR [Ad + iyy]
AR 3	=	AR [NC + iyy]
AR 4	=	AR [NE + iyy]
AR 5	=	AR [NA + iyy]
AR 6	=	AR [NP + iyy]
AR 7	=	AR [PA + iyy]
AR 8	=	AR [PP + iyy]
AR 9	=	AR [ND + iyy]
AR10	=	AR [J + iyy]
AR11	=	AR [NN + iyy]
AR12	=	AR [NH + iyy]

e. Autres formes adjectivales (AFA)

Les AFA sont au nombre de trois :

AFA 1	=	Adjectif comparatif (AC)
AFA 2	=	Adjectif possessif (AP)
AFA 3	=	Adjectif d'emprunt (AE)

4. Les adverbes

Les adverbes sont au nombre de cinq :

Ad 1	=	ʒamīʃ
Ad 2	=	kull
Ad 3	=	niš̄f
Ad 4	=	šibʃ
Ad 5	=	S + iyyan

5. Les joncteurs

Les joncteurs sont au nombre de vingt-six :

J 1	= bi-	J14	= dūna
J 2	= il-	J15	= fawqa
J 3	= min	J16	= taḥta
J 4	= ʕalā	J17	= ǧajra
J 5	= ðū/ðāt	J18	= qabla
J 6	= fī	J19	= lā
J 7	= ʕan	J20	= ʕabra
J 8	= ʔilā	J21	= maʕa
J 9	= wa	J22	= ʕinda
J10	= ḥattā	J23	= ḥawla
J11	= warāʔa	J24	= dāχila
J12	= bajna	J25	= bi-dūni
J13	= baʕda	J26	= ka-mā

B. Paramètres d'analyse de la notion

Les paramètres d'analyse de la notion ont été examinés au cours de l'étude. Ils constituent des catégories notionnelles quand ils désignent l'UNC en tant que telle et des traits notionnels quand ils sont appliqués aux constituants. Les catégories notionnelles sont au nombre de douze : action = ACT, processus = PRS, grandeur = GRD, état = ETA , propriété = PRT, espace = ESP, phénomène = PHN, objet = OBJ, occupation = OCP, procédé = PRD, outillage = OUT et modifieur = MOD.

III. Conclusion

Au cours de ce chapitre, nous avons examiné tour à tour les problèmes de traitement terminotique de la langue arabe et le processus de codage aux différents niveaux morphologique, schémique, morphosyntaxique et notionnel. Par ailleurs, nous avons attiré l'attention sur certains problèmes reliés au codage de la notion. En outre, nous avons parlé de la saisie des données puis du traitement terminotique et des problèmes que nous y avons rencontrés. Enfin, nous avons proposé une méthode d'analyse quantitative qui pourrait, espérons-le, contribuer à l'exploration du domaine de la syntagmatique. Enfin, nous avons livré nos paramètres d'analyse sous forme de répertoire catégoriel de toutes nos données. Ces paramètres serviront à une meilleure compréhension de notre démarche d'analyse.