« noms composés »
dans le choix de l'article indéfini *de* et *des* en français

Itsuko FUJIMURA[1], Hiroshi NAKAO[2]

[1] Université de Nagoya – 4648601 Nagoya –- Japon

[2]Université d'Aichi – 4700296 Miyoshi-cho – Japon

1. Introduction

Cette étude porte sur la question des "noms composés" dans le choix de l'article *de* et *des* devant les noms précédés d'épithète en français (ex. *une bonne condition* vs. *de bonnes conditions*, *une jeune fille* vs. *des jeunes filles*). Nous abordons cette question avec des méthodes statistiques, l'Information Mutuelle en particulier (cf. BARNBROOK,1996; OAKES,1998), dans un corpus journalistique homogène, très grand et doté d'informations grammaticales.

Dans nos précédents travaux (FUJIMURA et al., 2004; FUJIMURA et al. à paraître), nous avons statistiquement relevé, en analysant les divers corpus de très grande taille, plusieurs critères déterminant le choix de l'article *de* et *des* qui n'avaient guère été signalés dans la littérature, aussi bien que les conditions mentionnées à maintes reprises : le niveau de langue et la question des noms composés (cf. LE BIDOIS et LE BIDOIS, 1967; DUPRÉ, 1972; TOGEBY, 1982). Nous y avons proposé une hypothèse explicative pour rendre compte synthétiquement de ces premières conditions, soit le "poids" de l'épithète ou l'"importance" de l'épithète, dont les sous-catégories sont "poids lexical", "poids discursif" et " poids phonétique"[1]. *De* est préféré quand l'épithète est "moins légère", alors que *des* l'est quand cette dernière est "plus légère". Le "poids lexical" concerne la caractéristique lexicale des adjectifs (et /ou des adverbes qui peuvent éventuellement les qualifier comme *très* ou *tout*): *petit* est par exemple considéré comme un adjectif lexicalement "plus léger" que *nouveau,* qui est à son tour "plus léger" qu'*excellent*. Le "poids discursif" se rapporte à la nature sémantico-fonctionnelle de l'épithète dans le discours: l'adjectif qualifié par un adverbe d'intensité (ex. *très grand*) est "moins léger" que l'adjectif seul (ex. *grand*), puisque l'information transmise est plus abondante dans le premier cas que dans le second. Le "poids phonétique" correspond à la longueur phonique de l'épithète: l'adjectif avec liaison phonique (ex. *nouveaux arrivants*) est phonétiquement plus long donc "moins léger" que celui qui n'accompagne pas de liaison (ex. *nouveaux venus*).

A propos de la question des "noms composés", nous avons déjà constaté que le degré de collocation entre l'adjectif et le nom, mesuré au moyen de l'Information Mutuelle, était en corrélation avec le taux de *des*, bien que faible, auprès d'un millier de bigrammes qui ont été recueillis dans le même corpus journalistique que celui de ce présent travail mais sans étiquetage syntaxique (FUJIMURA et al., 2004). Le degré de collocation entre deux mots a été défini comme degré d'association statistiquement préférentielle entre eux. Nous avons interprété ce phénomène toujours avec la notion de "poids": un constituant d'un "nom composé" est sémantiquement et fonctionnellement plus léger qu'une épithète pleine.

---

[1] L'idée de "poids" nous a été inspirée par ABEILLÉ et GODARD,1999, 2000; ARNOLD et al., 2000; WASOW, 1997.

L'objectif de ce présent travail est double: tout d'abord la vérification descriptive de ce dernier résultat avec une procédure plus raffinée. Nous avons employé dans ce but le corpus avec les informations syntaxiques, qui nous permettrait d'obtenir un résultat plus certain. Nous avons également évalué trois indices de collocation : IM, t-score et z-score, pour employer l'IM en tant qu'indicateur le plus fiable du degré de figement sémantique des <ADJ + NOM>. Le détail méthodologique sera examiné dans les sections 2 et 3.
Notre second objectif est théorique: notre analyse porte sur le statut des "noms composés" dans le choix de *de* et *des*. Nous voudrions démontrer dans l'observation de ce fait qu'il est adéquat de considérer les "noms composés" comme phénomène dynamique et probabiliste et non comme catégorie statique et absolue. Ce qui est pertinent dans le choix de *de* et *des* est le rendement de l'adjectif qui qualifie le nom. Plus le <ADJ + NOM> est figé, moins l'ADJ est actif dans la qualification et plus *des* est choisi. Moins celui-ci est figé, plus l'ADJ est actif et plus *de* est choisi.
On va voir d'abord dans la section 4 que le choix de *de* et *des* a une corrélation continue avec le degré de figement sémantique des bigrammes, mesuré par l'IM. On n'y trouvera aucun signe supposant l'existence de la classe fermée des "noms composés". On va observer ensuite dans la section 5 que les autres conditions telles que facteur lexical ou phonétique interviennent indépendamment de ce fait, à côté du degré de figement sémantique. Cela indique que les "noms composés" n'ont pas de traitement spécifique dans ce phénomène, tel qu'il est mentionné dans les livres de grammaire. On va à la fin démontrer dans la section 6, dans la discussion qualitative sur le problème du "défigement", que l'usage figé et le non-figé des "noms composés" sont également sur l'échelle continue.

2. méthode

Nous avons établi en premier lieu le corpus de 271 millions de mots à partir des périodiques contemporains: Libération, Le Monde, Le Point, Le Monde Diplomatique et La Tribune distribués par CDROM-SNi et attribué des informations syntaxiques catégorielles à tous les mots de ces textes au moyen du Tree Tagger.
Nous avons créé en second lieu une base de données constituée d'environ 3400 bigrammes de < ADJ + NOM > (= 13000 tokens). C'étaient des < *ancien, beau, bon, grand, nouveau* et *petit* + NOM au pluriel > précédés de < PRÉPOSITION : *à, avec, dans, derrière, devant, entre, malgré, par, parmi, pour, selon, sur, suivant* et *sous* + (*de* | *des*) >. Nous avons ainsi restreint le contexte pour identifier le plus exactement possible les séquences correspondant à < ART + ADJ + NOM au pluriel >. Nous avons utilisé pour la recherche des prépositions, des adjectifs et des noms, l'information de la catégorie syntaxique donnée aux mots, alors que nous étions obligés de recourir au contexte pour la recherche de *de* et *des*, parce que la catégorisation syntaxique de ces derniers était hors de portée de la performance de Tree Tagger. Notre base de données ne représente donc qu'une sous-classe du phénomène. Nous considérons cependant que cela n'entraîne pas de conséquence erronée pour notre but, car l'examen des données que nous avons effectué dans nos précédents travaux nous a enseigné que la fonction grammaticale du < ART + ADJ + NOM> ne joue pas de rôle pertinent pour le choix entre *de* et *des*. Les informations sur *jeune fille*, "nom composé" considéré comme typique, y sont données en tant que le repère. L'adjectif *jeune* n'est pas l'objet de recherches dans ce travail, en raison de la particularité du comportement de cet adjectif qui ne qualifie en général que les êtres vivants.
Nous avons calculé le taux de *des* et l'IM, t-score et z-score pour chaque bigramme: <ADJ+NOM> au pluriel. Afin d'obtenir ces indices au moyen des expressions indiquées ci-après, il nous a fallu la fréquence de chaque adjectif, chaque nom et chaque bigramme dans tout le corpus. Nous avons aussi utilisé des informations catégorielles pour ces calculs. A la

fin, nous avons ajouté à chaque bigramme les informations concernant la nature de son correspondant au singulier : l'IM, le t-score et le z-score des <ADJ+NOM> au singulier. Voici les expressions développées avec lesquelles nous avons calculé les scores:

$$IM(x,y) = \log_2 \frac{F(x,y) * \text{Nombre total de mots}}{F(x)\,F(y)}$$

$$\text{t-score}(x,y) = \frac{F(x,y) - \frac{F(x)\,F(y)}{\text{Nombre total de mots}}}{\sqrt{F(x,y)}}$$

$$\text{z-score}(x,y) = \frac{F(x,y) - \frac{F(x)F(y)}{\text{Nombre total de mots}}}{\sqrt{F(x)\,F(y)}}$$

Nombre total de mots = 271 millions, F(x) = fréquence de l'ADJ, F(y)= fréquence du NOM, F(x, y) = fréquence de l' <ADJ + NOM>

3. Information Mutuelle et "nom composé"

l'Information Mutuelle est considérée depuis son origine comme indicateur de l'association entre les deux mots, à savoir celui des "noms composés" par excellence. Voici quelques exemples calculés dans nos corpus :

exemple	F(ADJ)	F(NOM)	F(ADJ+ NOM)	IM	F (ADJ+NOM) après le contexte: < PREP + *de /des* >	taux de *des* (%)
petites annonces	66361	3665	869	9.92	16	100.0
petites entreprises	66361	129576	2472	6.28	80	32.5
petits moyens	66361	49919	65	2.41	24	20.8
jeunes filles	33155	12385	2434	10.65	53	96.2

La comparaison entre *petites annonces*, *petites entreprises* et *petits moyens* montre clairement que l'association entre ADJ et NOM est la plus forte dans *petites annonces*, vu que un quart des occurrences d' *annonces*, soit 869 sur 3665, est accompagné de *petit(e)s*. Elle est la moins forte dans *petits moyens*, seul 65 occurrences de *moyens* sur 49919 étant données avec *petit(e)s*. Le degré d'association entre deux mots n'est pas le même que la fréquence brute de l'apparition de ces mots. L'IM est moins élevée pour *petites entreprises* que *petites annonces*, bien que ce premier apparaisse trois fois plus fréquemment que le dernier. L'IM de *jeunes*

filles est plus élevée que celle de *petites annonces*, parce que la fréquence de *jeunes* est moins importante que celle de *petit(e)s*, bien que seul un sixième des occurrences de *filles* soit accompagné de *jeunes.*
Habert et Jacquemin disent par exemple : "...l'observation de paires à IM forte est sûre d'obtenir les "noms composés" dans un corpus où leur nombre d'apparition est assez élevé pour qu'ils soient repérables (HABERT et JACQUEMIN, 1993, 26-27)." On caractérise les "noms composés" premièrement par la structuration sémantique: degré de figement sémantique ou degré d'opacité (MEJRI, 2003, 30).
La figure 1 montre une autre preuve pour cet argument : forte corrélation entre les IM des <ADJ + NOM> au pluriel et au singulier (n=820, r de Pearson = 0.556). Dans ce diagramme de dispersion, chaque point correspond à un bigramme < ADJ + NOM > dont la position est déterminée par les valeurs des IM au singulier et au pluriel. La corrélation est moins forte pour le t-score et le z-score comme l'indiquent les figures 2 et 3[2]. Nous n'avons pris en

[2] Nous présentons à titre exemplaire les dix premiers IM, t-scores et z-scores dans le tableau suivant. *Jeunes filles* est parmi les dix premiers pour tous ces indices (le 4ème pour l'IM, le 9ème pour le t-score et le premier pour le z-score).

indice	ordre	ADJ+NOM au pluriel	valeur de ces bigrammes au pluriel	valeur des correspondants au singulier
IM	1	bonnes volontés	11.20	7.11
	2	petits riens	11.08	0.02
	3	petits boulots	10.95	6.54
	4	jeunes filles	10.65	9.64
	5	nouveaux venus	10.65	8.67
	6	belles empoignades	10.64	8.48
	7	petits déjeuners	10.59	8.86
	8	nouveaux arrivants	10.40	7.23
	9	petites touches	10.29	3.99
	10	petites annonces	9.92	4.42
t-score	1	grandes entreprises	75.10	24.29
	2	grands groupes	69.29	28.41
	3	nouvelles technologies	67.66	18.46
	4	grandes villes	65.14	30.83
	5	grandes lignes	57.30	-1.48
	6	grandes surfaces	52.88	27.42
	7	grandes banques	51.65	30.87
	8	grandes écoles	49.50	18.10
	9	jeunes filles	49.30	64.38
	10	petites entreprises	49.08	27.92

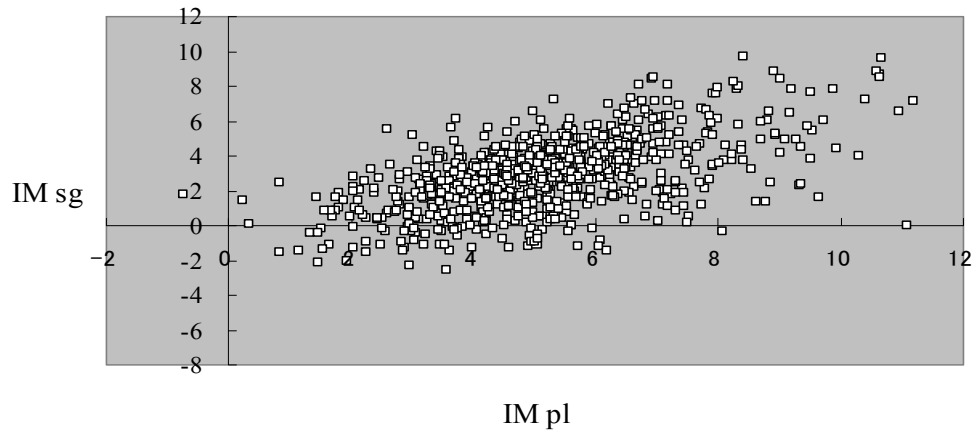
compte dans ce calcul que les bigrammes qui apparaissaient plus de 9 fois pour le pluriel aussi bien que pour le singulier dans le corpus. C'est afin d'atténuer le défaut de cette méthode qui favorise excessivement les rencontres rares (cf. FRANÇOIS et MANGUIN, à paraître). Voici des exemples:

bonnes volontés	11.20	*bonne volonté*	7.11
nouveaux venus	10.65	*nouveau venu*	8.67
bons augures	8.41	*bon augure*	9.67
anciens élèves	7.69	*ancien élève*	4.71
nouveaux marchés	5.31	*nouveau marché*	3.72
grandes familles	5.48	*grande famille*	3.55
bonnes critiques	2.75	*bonne critique*	0.87
grands résultats	0.34	*grand résultat*	0.11

Jeune fille se positionne à l'endroit le plus proche de l'angle haut et droit dans cette figure, ayant les scores élevés aussi bien pour le pluriel que le singulier, à côté de quelques autres bigrammes mentionnés dans la note 2:

jeunes filles	10.65	*jeune fille*	9.63

Figure 1 : Corrélation entre les IM des <ADJ+NOM> au pluriel et au singulier (n=820)



z-score (x 1000)	1	jeunes filles	120.04	110.21
	2	nouvelles technologies	93.68	4.70
	3	grandes surfaces	93.37	13.64
	4	grandes villes	61.14	6.07
	5	petits porteurs	57.94	4.39
	6	grandes lignes	57.20	-0.08
	7	bonnes volontés	57.19	41.87
	8	petites annonces	55.66	4.75
	9	bonnes intentions	55.46	0.51
	10	grands groupes	54.85	3.63

On sait par intuition que le degré de figement sémantique des <ADJ + NOM> au singulier n'est pas éloigné par rapport au pluriel, parce que le nombre "ne touche pas à l'intégrité du contenu conceptuel de la séquence (MEJRI, 2003, 32)" en général. Les exemples précédemment cités sont parmi ces cas heureux. Il existe cependant également des contre-exemples comme les suivants. Les valeurs éloignées pour le pluriel et le singulier sont attribuables à une différence sémantique ou fonctionnelle de ces bigrammes au singulier et au pluriel :

grandes lignes	8.08	*grande ligne*	-0.30
grands travaux	6.17	*grand travail*	-1.47

En effet, *Grands travaux* signifie: "grands chantiers d'importance nationale" que *grand travail* ne partage pas. *Grandes lignes* est fréquemment employé pour indiquer "lignes de chemin de fer desservant de longues distances" mais *grande ligne* ne le désigne pas. *Grandes lignes* est aussi couramment utilisé sous la forme de *dans les grandes lignes* dont le sens est "en gros", tandis que *grande ligne* ne l'est pas.

Nous considérons donc l'IM comme indicateur le plus adéquat du degré de figement sémantique de <ADJ + NOM> parmi les trois indices examinés.

Figure 2: Corrélation entre les t-scores des <ADJ + NOM> au pluriel et au singulier (n=820)

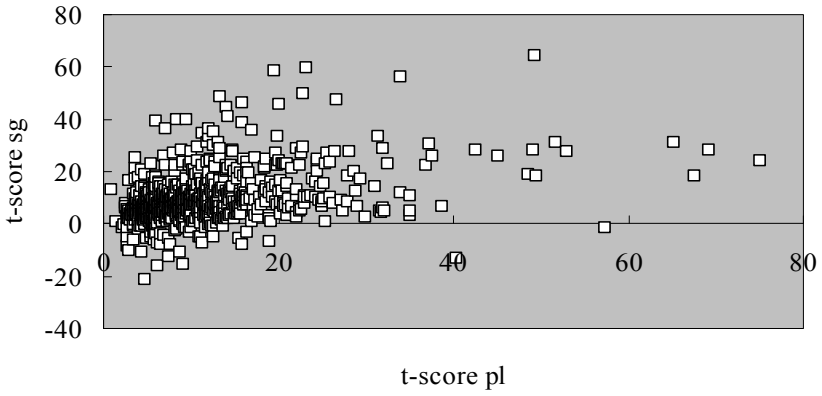
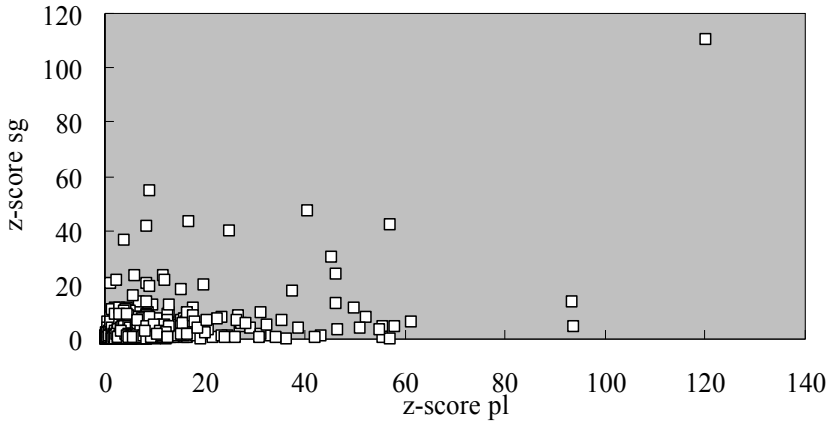


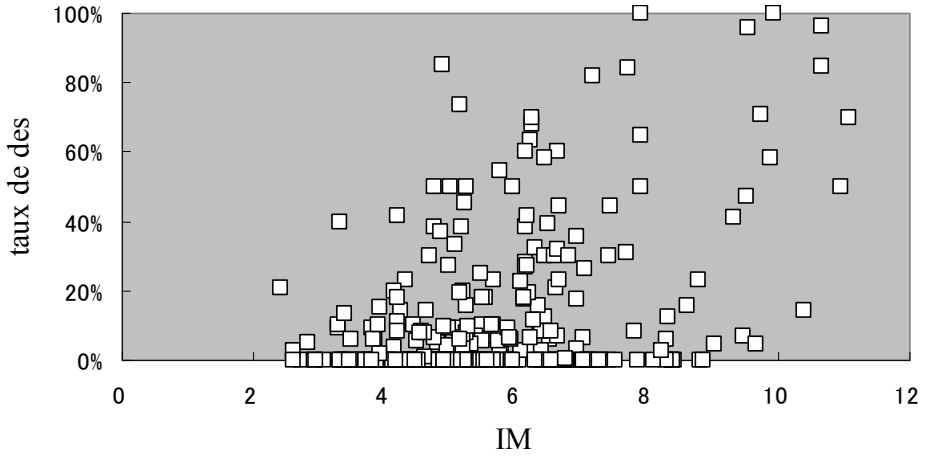
Figure 3: Corrélation entre les z-scores des <ADJ + NOM> au pluriel et au singulier (n=820)



4. Information Mutuelle et taux de *des*

Nous démontrons ensuite avec la figure 4 que le degré de figement sémantique des <ADJ + NOM au pluriel> mesuré avec l'IM est en corrélation d'intensité moyenne avec le taux de l'article: *des* auprès de 268 bigrammes de <ADJ + NOM au pluriel > (r de Pearson = 0.418). Nous avons pris en compte ici seulement les séquences qui apparaissaient plus de 9 fois, précédées de < PRÉPOSITION + (*de* | *des*) >, pour que le taux de *des* calculé ait une certaine fiabilité.

Figure 4: Corrélation entre le taux de *des* et l'IM des <ADJ + NOM> (n=268)



Les exemples sont les suivants :

	IM	taux de *des* (%)	nombre d'occurrences
jeunes filles	10.65	96.2	53
nouveaux arrivants	10.39	14.3	14
petites annonces	9.92	100.0	16
grandes surfaces	9.72	71.1	38
bonnes intentions	9.65	4.5	22
nouvelles technologies	9.02	4.7	43
grands ensembles	8.78	23.1	26
anciens élèves	7.69	31.3	16
grands noms	7.43	44.6	65
bonnes conditions	6.75	0.6	667
petites entreprises	6.28	32.5	80
grands comptes	4.88	85.2	27
petits moyens	2.41	20.8	24

La figure 4 montre que le choix de *de* et *des* est conditionné d'une façon continue et graduelle par l'IM. On peut supposer donc que les "noms composés" se situent sur une échelle continue. S'il y avait une distinction catégorielle entre le groupe des "noms composés" demandant l'usage de *des* et celui de bigrammes composés de deux mots demandant *de,* nous aurions le diagramme avec deux concentrations de dispersion. La raison pour laquelle cette corrélation

n'est pas parfaite est attribuable à l'existence d'autres facteurs qui sont complètement indépendants du degré de figement des séquences mais contribuent chacun au choix de *de* et *des,* que nous allons examiner plus bas.

5. "Noms composés" en tant qu'une des conditions pour le choix de *de* et *des*

Les facteurs relevés dans nos travaux précédents (FUJIMURA et al. 2004; FUJIMURA et al. à paraître) sont les suivants. Ce sont tous les constituants du degré du "poids" de l'épithète, qui interviennent dans ce phénomène[3].

"Poids" de l'épithète

(*Des* est préféré avec l'épithète "plus légère" alors que *de* l'est avec la "moins légère")

de <= - léger <----------> + léger => *des*

1 "poids" lexical de l'épithète

- *excellent* < *nombreux,.. nouveau* < *beau* < *grand* < *petit*

  ex. *de grands effectifs / des petits effectifs*[4]

  ex. *d'excellents travaux / des petits travaux*

- *très* < *tout*

2 "poids" discursif de l'épithète

- plus informatif < moins informatif
- avec adverbe < sans adverbe

  ex. *de très petits groupes / des petits groupes*

- adjectif plein < constituant du nom composé[5]

  ex. *de petits moyens* / *des petites annonces*

3 "poids" phonétique de l'épithète

- avec liaison entre l'adjectif et le nom < sans liaison entre ces deux

  ex. *de nouveaux arrivants* / *des nouveaux venus*

- genre féminin < genre masculin

[3] D'après l'étude diachronique que nous avons menée (FUJIMURA et al., 2004; FUJIMURA et al., à paraître), les critères relevés ici sont tous pertinents depuis le début de l'histoire de cette question au 17ème siècle. Il ne faut évidemment pas oublier le facteur de niveau de langue. On peut tout de même le mettre de côté dans la discussion de cet article, puisque nous travaillons dans le corpus le plus homogène possible pour le faire intervenir le moins possible.

[4] Il va sans dire que ces exemples sont donnés pour montrer la tendance préférentielle de l'usage. A quelques rares cas près, les formes *de* et *des* sont toutes les deux acceptables.

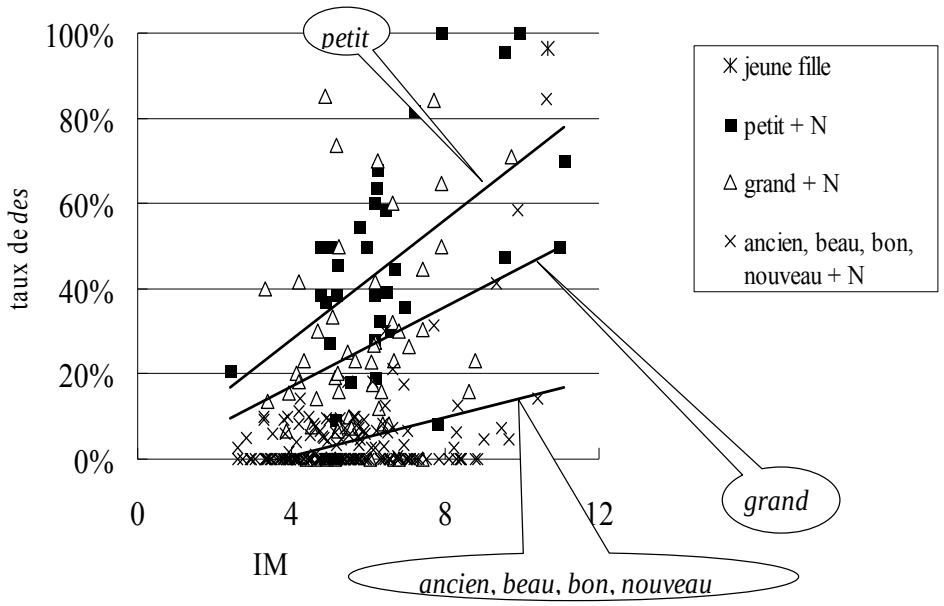
[5] Ce point est l'objet de ce présent travail, donc suspendu pour le moment.

Nous allons traiter de plus près le caractère multifactoriel de ce phénomène dans ce qui suit, en examinant en particulier l'adjectif, facteur lexical et la liaison, facteur phonétique. Nous observerons la relation entre ces facteurs et la question des "noms composés", dans le but de démontrer que cette dernière n'est qu'un des constituants du "poids" de l'épithète comme les autres.

5.1. IM, adjectif et choix de *de* et *des*

Voyons d'abord la figure 5 qui montre la relation entre l'IM, l'adjectif et le taux de *des* de l<ADJ + NOM>.

Figure 5: IM, adjectif et taux de *des*



Cette figure veut montrer que le taux de *des* d'un bigramme est déterminé par les deux facteurs : adjectif et IM, qui sont indépendants l'un de l'autre. On peut dire d'abord que l'IM est un critère pertinent dans le choix de *de* et *des* pour chacun des adjectifs : *petit, grand, ancien, beau, bon* et *nouveau.* Les courbes d'ajustement données dans le diagramme montrent que *des* a tendance à être choisi à mesure que la valeur de l'IM est élevée pour tous ces adjectifs[6] (r de Pearson: 0.53 (*petit*), 0.29 (*grand*), 0.39 (*ancien, beau, bon* et *nouveau* regroupés)). La figure montre aussi que l'adjectif est un critère approprié dans le choix de *de* et *des* à travers toute l'échelle de l'IM comme l'indiquent de nouveau les courbes d'ajustement. Dans notre base de données, le taux de *des* calculé pour chaque adjectif est: 9.5% (ancien), 2.8% (beau), 2.9% (bon), 26.4% (grand), 2.3% (nouveau) et 42.0% (petit)[7]. L'affinité

[6] Pour éviter les complications sur la figure, nous avons regroupé les adjectifs : *ancien*, *beau*, *bon* et *nouveau*.

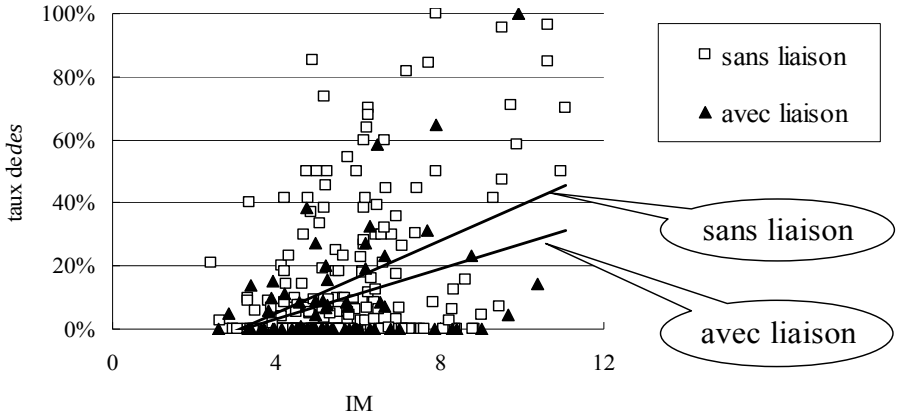
[7] Le taux de *des* pour chaque adjectif dans les textes contemporains de divers genres étudiés dans FUJIMURA et al., 2004 était : 53.3% (*petit*) , 46.4% (*jeune*) , 27.2% (*gros*) , 24.8% (*faux*) , 23.1% (*vrai*) , 23.0% (*mauvais*) , 22.7% (*vieux*) , 20.2% (*grand*) , 14.8%

particulièrement forte entre *des* et *petit* est donc à considérer comme caractéristique telle quelle de l'adjectif, indépendante de l'IM, sans rapport avec la question des "noms composés". Il ne faut donc pas traiter ce phénomène comme Togeby : « petit *forme souvent avec le substantif une sorte de mot composé* » (TOGEBY, 1982, 52), le taux de *des* pour *petit* étant toujours plus élevé que celui des autres à travers toutes les valeurs de l'IM.

5.2. IM, liaison et choix de *de* et *des*

Observons ensuite la figure 6, qui présente la relation entre l'IM, la liaison phonique et le taux de *des*.

Figure 6: IM, liaison et taux de *des*



Cette figure montre de la même manière que la précédente, que le taux de *des* d'un bigramme est conditionné par deux facteurs : liaison et IM, qui sont indépendantes l'une de l'autre. L'IM est un critère pertinent, que se produise ou non la liaison phonique entre l'ADJ et le NOM. Plus la valeur de l'IM est élevée, plus le taux de *des* l'est pour les bigrammes avec liaison et sans liaison (r de Person : 0.38 (avec liaison), 0.56 (sans liaison)). Les courbes d'ajustement données dans le diagramme indiquent que *des* a tendance à être choisi à mesure que la valeur de l'IM est élevée sous ces deux conditions phonétiques. La figure montre également que la liaison est un critère déterminant dans le choix de *de* et *des* à travers toute l'échelle de l'IM comme le montrent les courbes d'ajustement. Dans notre base de données, le taux de *des* est de 10.8% pour les bigrammes avec liaison, tandis qu'il vaut 13.8% pour ceux qui sont sans liaison[8]. Nous caractérisons cette condition phonétique que personne n'a jamais mentionnée à

(*ancien*) , 11.2% (*bon*) , 10.5% (*beau*) , 4.8% (*nouveau*) , 0.2% (*nombreux*) , 0% (*excellent*) .

[8] Nous avons constaté que le taux de *des* est 14.5% auprès des bigrammes sans liaison, tandis qu'il est 8.1% auprès des bigrammes avec liaison dans les divers textes contemporains (Fujimura et al. à paraître). Ce facteur est statistiquement significatif dans le choix de *de* et *des*.

notre connaissance, comme un facteur indépendant déterminant le choix de *de* et *des*. Elle est un constituant de la longueur phonique de l'épithète[9], dont un autre est le genre grammatical.

Il faut enfin souligner que le type d'adjectif et sa longueur, facteurs concernant l'épithète, fonctionnent comme déterminant du taux de *des,* même auprès des "noms composés" dont le degré de figement est bien avancé, dotés d'une valeur de IM bien élevée[10]. Or, si l'on définissait les "noms composés" comme bigrammes dont le figement sémantique est complet et qui constituent une liste fermée et statique déterminée dans le dictionnaire, l'adjectif dans un "nom composé" ne devrait pas jouer par définition la fonction de l'épithète, parce qu'il n'y a pas le statut indépendant[11]. Il serait a priori contradictoire de discuter sur la fonction de l'épithète dans des "noms composés" ainsi définis. Les faits examinés dans les figures 5 et 6 nous amènent donc à soutenir au contraire que le problème des "noms composés" dans le choix de *de* et *des* est à aborder dans leur acceptation graduelle, comme le dit notre hypothèse initiale.

Avant de tirer une conclusion finale, nous allons présenter notre réflexion sur le problème de "défigement" dans la section suivante.

6. Problème de "défigement"

6.1. qualité et quantité de figement

Nous avons examiné jusqu'à présent la question de degré de figement des <ADJ + NOM>, sans tenir compte de leur "défigement". Nous avons dit par exemple que *jeunes filles* est sémantiquement plus figé que *bonnes conditions*. Le premier signifie: " femme jeune non mariée (girl en anglais)" et fonctionne comme antonyme de *garçon*, alors que le dernier n'a pas de signification combinatoire spécifique. L'IM pour le premier est en effet 10.65 tandis que celle du dernier, 6.75. L'IM qui est à l'origine l'indice du degré de collocation entre les deux mots, autrement dit degré d'association statistiquement préférentielle entre eux, a été considérée comme indice du figement sémantique. La valeur quantitative a été interprétée comme qualitative, parce que nous avons pensé que l'association préférentielle des mots cause leur figement sémantique et vice versa. Nous avons donc expliqué le choix de *de* et *des* au moyen de l'IM[12]. *Jeunes filles* apparaît presque toujours avec *des* (96.2%), parce que son

---

[9] L'adjectif au féminin est plus long que celui qui est au masculin, de la même manière que l'adjectif avec liaison l'est plus que celui qui est sans liaison.

[10] L'usage de *de* est attesté même à côté des "noms composés" marqués comme tels par un trait d'union, dans les textes soutenus comme suit:

- Dieu aime la justice et le bon droit et qu'on ne le trompe pas par de faux-semblants. (Oldenbourg, Z., *Les cités charnelles ou l'Histoire de Roger de Montbrun*, 1961)
- Les groupes, quoique rapprochés, vivent renfermés sur eux-mêmes, dans les cadres traditionnels qui contiennent, soit en agriculteurs, soit en artisans, tout ce que peuvent réclamer les besoins et même les ambitions de luxe, et qui, une fois complets, s' ouvrent difficilement à de nouveaux-venus. (Vidal de la Blanche, P., *Principes de géographie humaine*, 1921).

[11] Morphologiquement parlant, l'adjectif comporte en général comme mot indépendant, même dans des "noms composés" bien figés. L'exception serait *grand-mère.*

[12] D'autres facteurs y interviennent également comme nous l'avons répété.

degré de figement est très avancé, tandis que *bonnes conditions* ne le fait pas (0.6%), parce que il l'est moins. Il y a cependant le problème du "défigement" à résoudre dans cette discussion. C'est que *jeunes filles* n'est pas toujours plus figé que *bonnes conditions*. Cette expression peut être employée comme combinaison de deux mots indépendants : "young girl" en anglais comme l'exemple (1). L'expression non-figée a évidemment plus d'affinité avec l'usage de *de* que celui de *des*.

(1) On dit qu'il a fait une solide enquête sur ces repas qui permettent à de jeunes filles de continuer plus commodément leurs études et à des hommes de passer un bon moment (libre à chacune de prolonger la soirée si elle le souhaite). (Libération, 03/1997)

Le problème méthodologique est que l'IM n'est pas séparément calculée pour *jeunes filles* figé et *jeunes filles* non-figé. Nous voudrions tout de même réaffirmer ici que les "noms composés" avec l'IM élevé tels que *jeunes filles* sont plus figés que ceux avec l'IM moins élevé tels que *bonnes conditions*. On ne peut pas dire que toutes les occurrences particulières de *jeunes filles* sont plus figées que celles de *bonnes conditions*. Mais on peut dire que *jeunes filles* en tant que type l'est plus que *bonnes conditions* en tant que type, pour la raison que l'usage non-figé de *jeunes filles* serait rare à cause de l'existence même de l'usage très figé. Nous considérons que plus l'association entre deux mots est forte, plus le figement sémantique entre eux est important et moins l'usage "défigé" est fréquent. D'après nous, il n'est pas aberrant que la valeur de l'IM soit à la fois l'indice qualitatif de figement sémantique des deux mots et l'indice quantitatif de rareté proportionnelle de l'usage non-figé, comme l'indique la figure 7 ci-après.

Cela dit, on va continuer à examiner les exemples, en montrant que la question du "défigement" ne conditionne pas aussi clairement que l'on pense le choix de *de* et *des*, afin de montrer ultérieurement qu'elle n'est pas le critère privilégié mais n'est qu'une condition parmi d'autres.

6.2. figement comme condition probabiliste

Tout d'abord pour *jeunes filles* (IM : 10.65, taux de *des* : 96.2%, n: 53, *des*: 51, *de*: 2) qui est le "nom composé" typique, l'opposition entre *de* et *des* semble correspondant à celle entre le figement et le "défigement". Sur 53 occurrences de *jeunes filles*, seuls deux exemples (1) et (2) sont accompagnés de *de*, et les francophones natifs les interprètent comme "défigés":

(2) Ce sont alors de longues alternances de couplets et de répliques parlés décrivant les aventures de princes insensibles à de jeunes filles amoureuses ou prêts à partir à la guerre. (Le Monde, 03/1987)

alors que les figés qui sont majoritaires sont marqués par *des* comme l'exemple (3) :

(3) Cet établissement, qui permet d'abord de donner du travail à des jeunes filles employées comme femmes de chambre, remplit également un rôle non négligeable (La Tribune, 01/1996)

Pour *grandes surfaces* (IM : 9.73, taux de *des* : 71.1%, n: 38, *des*:27, *de*: 11), la relation entre le "défigement" et l'usage de *de* est moins transparente. *Grandes surfaces* signifie "magasins vendant de nombreux produits en libre service, sur une vaste superficie" s'il est figé, tandis qu'il réfère à la "superficie matériellement grande" dans l'usage non-figé. La différence de signification entre les usages figé et non-figé est grande dans cette combinaison par rapport à d'autres, car la "superficie" et le "magasin" sont deux concepts complètement différents. Dans

notre base de données, tous les exemples figés signifiant "magasin" sont accompagnés de *des* comme dans l'exemple (4):

(4) Une centaine de producteurs de pêches de l'Isère et du nord de la Drôme ont déversé 200 tonnes de pêches, samedi, devant des grandes surfaces à Salaise-sur-Sanne (Isère), pour protester contre la chute des cours. (Libération, 08/1996)

alors que les non-figés sont marqués par *de* :

(5) Et plutôt que de peindre penchés, coincés entre chaise et pupitre, les enfants se meuvent devant de grandes surfaces de papier "comme des danseurs". (Le Monde Diplomatique, 12/1994)

Les exemples (6), (7) et (8)[13] dans lesquels *de* est employé sont cependant intéressants par leur caractère intermédiaire entre le figé et le non-figé. Le (6) parle surtout de la grandeur de la superficie des magasins, le (7) cite plutôt des magasins avec une grande superficie et dans (8), il ne s'agit plus de la superficie matérielle des magasins mais de leur envergure abstraite qui est grande.

(6) L'essor de ce commerce spécialisé, très concentré, est avant tout celui de deux géants succursalistes (Décathlon et Go Sport). Proposant un large choix de produits sur de grandes surfaces (en moyenne 900 m²), ils occupent une part croissante du marché des articles de sport hors textile (51 % du chiffre d'affaires du secteur, contre 22 % en 1988) et dominent tous leurs concurrents directs (commerces organisés en réseau, magasins indépendants, hypermarchés et autres commerces généralistes).(La Tribune, 08/1998)

(7) Enfin, en France, Leroy-Merlin espère pouvoir ouvrir trois nouvelles unités dans les six prochains mois. Mais avec un regret : l'impossibilité de créer de très grandes surfaces de 15.000 à 20.000 mètres carrés. " (La Tribune, 06/1997)

(8) J'ai longtemps fait mes courses dans de très grandes surfaces (Leclerc, Carrefour, Auchan etc..) et j'en suis revenue. (http://www.toluna.com/SUPER_U_LE_magasin-av-575582.html, 06/03/2006)

L'usage de *de* dans ce dernier exemple ne correspond pas au sens non-figé, mais il traduit le renforcement de l'épithète "très grand" sans dégrader pour autant la signification de ce "nom composé".

Nous allons ensuite voir *petites phrases* (IM: 9.51, taux de des: 47%, n: 19, des:9, de:10) dans les exemples suivants. Dans (9) celui-ci n'est pas figé, alors que dans (10), il est figé avec le sens : "extrait des propos d'un homme public et abondamment commentés par les médias".

(9) Désarmante, elle est désarmante, Danielle Darrieux. On attendait une grande dame, elle répond avec de petites phrases colorées, mâtinées d'argot : " rigolo ", " marrant ", ce sont ses mots. (Le Monde, 02/1987)

(10) Son gouverneur, Alan Greenspan, se laisse même aller à des petites phrases que ne renierait pas un partisan de "l'autre politique": "On peut tolérer une inflation de 3% afin de préserver les perspectives de croissance." (Libération, 08/1995)

[13] Les exemples (7) et (8) avec *trés* ne sont pas inclus dans notre base de données présentée dans la section 2. L'usage de *très* a affinité avec *de* (cf. FUJIMURA et al. 2004 et à paraître).

Pourtant dans (11), *petites phrases* avec *de* n'est pas "défigé". D'après nous, la raison de l'usage de *de* est ici stylistique. C'est le style de R. Fabius qui fait utiliser cette forme même avec un "nom composé", le critère du niveau de langue étant pertinent indépendamment des autres.

(11) " En un temps où la politique se résume souvent à de petites phrases, je me situerai sur un autre plan." Aux premiers mots, M. Fabius a donné le ton. L'occasion offerte par les nouvelles petites phrases de M. Rocard était trop belle pour que l'ancien premier ministre ne s'en saisisse pas, afin de peaufiner, par contraste, son image d'" anti-Rocard " à l'intérieur du PS : encore plus calme, si c'est possible, que d'habitude, écouté, comme à l'accoutumée, dans un silence religieux, ciselant des phrases imagées - " Nous vivons dans un univers où certains pays d'Asie considèrent que les Japonais sont des paresseux ", - M. Fabius n'est pas sorti de son registre, mais il en a déployé toute l'étendue. (Le Monde, 08/1987)

Nous allons enfin comparer, dans les exemples (12) -(15), les usages avec *de* et *des* auprès des "noms composés": *petits boulots* (IM; 10.95, taux de *des*: 59%, n:12, des:7, de:5) et *nouvelles technologies* (IM: 9.02, taux de *des*: 5% n: 38, des:2, de:36)[14]. On peut dire que les séquences sont plus figés dans les exemples (12) et (13), que les (14) et (15). Dans les premiers, *petits boulots* réfère à *"*jobs*"* qui est un type de travail, et *nouvelles technologies,* à *"*techniques modernes et complexes de l'information et de la communication", tandis que dans les seconds, la petitesse ou la nouveauté des référents sont soulignées.

(12) Le RMI est utile quand il les aide à surnager. Il ne l'est pas quand il devient pour de "faux exclus" une ressource minimale qu'ils complètent, selon l'humeur et les saisons, avec des petits boulots clandestins, ce qui les dissuade peu à peu de rechercher un emploi régulier. (Le Point, 02/1995)

(13) Autre chance? Lors de l'inauguration du réseau felletinois, le président de TDF, Bruno Chetaille affichait un bel optimisme, déclarant que le temps où la France jouait les "apprentis sorciers" en matière de télévision (voir notamment le crash des satellites TDF1 et TDF2) était révolu. TDF ne s'engagerait plus dans des nouvelles technologies sans avoir, au préalable, identifié un marché. Aura-t-elle plus de chance avec la télévision numérique de terre, expérimentée depuis peu à Rennes?. (Libération, 02/1999)

(14) Bonne en maths, elle était fascinée par les planeurs qui s'envolaient de l'aéroport local. Elle a financé ses cours de pilotage par de petits boulots. "Je ne dépensais rien en vêtements ni en sorties", raconte-t-elle. (Libération, 07/1999)

(15) Après quinze ans de vagues technologiques successives, les rois du jeu vidéo japonais se lancent dans une nouvelle guerre commerciale. Nintendo, Sega et Sony ont pour objectif d'élargir leur marché grâce à de nouvelles technologies toujours plus puissantes. (La Tribune, 04/1999)

Bien qu'il soit bien possible de juger ce phénomène comme figement ou "défigement" des mots, cette activité discursive n'est en fait que la mise en valeur de l'ADJ que nous avons discutée dans FUJIMURA, 2004, qui est un constituant du "poids" discursif : plus

[14] L'usage particulièrement rare de *des* avec *nouvelles technologies* n'est pas explicable avec ce facteur seul.

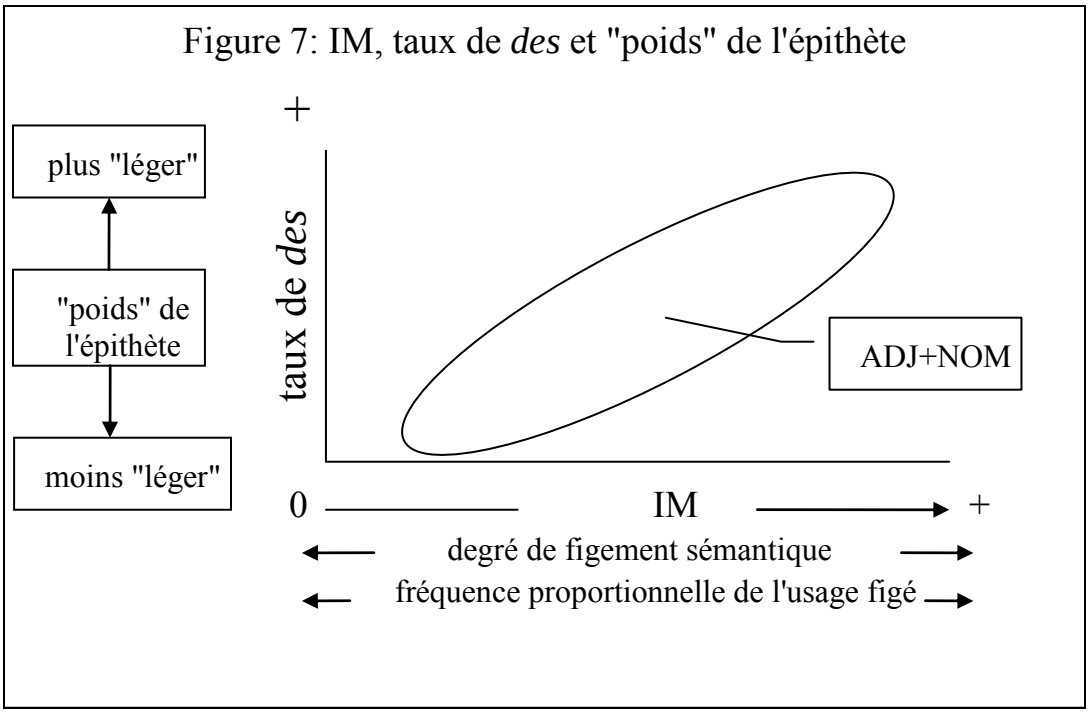
l'information que l'épithète véhicule est abondante, plus *de* est choisi ; moins elle l'est, plus *des* est sélectionné.

Il faut ajouter à la fin qu'il est difficile de mettre en lumière le problème du choix de *de* et *des* si l'on n'examine que le petit nombre d'exemples comme nous faisons dans cette section, parce que les facteurs qui le conditionnent sont très nombreux. Il nous semble que la méthode statistique avec le gros corpus est le seul moyen pour résoudre cette question multifactorielle et probabiliste.

7. Conclusion

Nous résumons les résultats obtenus dans la figure 7.

Figure 7: IM, taux de *des* et "poids" de l'épithète



Nous considérons que l'IM traduit la caractéristique potentielle et dictionnairique des bigrammes concernant le degré de figement entre l'ADJ et le NOM, qui fonctionne comme contrainte préalable lors de leur usage. Le choix de *de* et *des* est conditionné dans l'usage par ce facteur phraséologique parmi les autres. La fréquence proportionnelle de l'usage figé et non-figé est donc également indiquée par l'IM.
Le tableau suivant montre le processus du choix de l'article.

		épithète "moins légère"	épithète "plus légère"
contrainte phraséologique	↓	- figement sémantique + usage non-figé	+ figement sémantique + usage figé
activité discursive		- difficulté de mettre en valeur l'ADJ	+ difficulté de mettre en valeur l'ADJ
texte produit		+ occurrence de *de*	+ occurrence de *des*

Le problème des "noms composés" est uniformément traité comme facteur pertinent du choix de *de* et *des* dans tous les livres de grammaire. Chez certains grammairiens, les "noms composés" sont des entités lexicales complètement figées correspondant à un mot simple: "les expressions telles que *petites filles*, *gros mots* sont traitées comme un mot simple avec l'article *des*: *des petites filles, des gros mots*, parce qu'elles ont souvent pour synonymes des noms simples: *fillettes, injures* (TOGEBY 1982, 53)". En revanche, chez d'autres grammairiens, les "noms composés" sont caractérisés du point de vue plus sémantico-fonctionnel : "elle (=la langue) différenciait ainsi la qualification préalable, notoire, lexicalisée, qu'exprime parfois l'épithète antéposée (*des grands frères, des grandes sœurs, des petits pains,* etc.) de la qualification nouvelle, transitoire, voire prédicative, qui est son rôle dans bien des cas (*Elle a de grands yeux*, c'est-à-dire : Elle a les yeux grands). (GLLF, 1971, 260) ". Nos données soutiennent cette dernière caractérisation de "noms composés" et cela correspond aussi à l'avis de GROSS, 1996 et MEJRI, 2003.

Nous avons relevé par ailleurs:

1. l'IM est à considérer comme bon indicateur du figement sémantique.
2. le choix de *de* ou *des* est un composant du figement syntaxique des "noms composés". Cette étude est donc aussi à tenir pour une observation sur le rapport entre les figements syntaxique et sémantique.

Références

ABEILLÉ et GODARD, 1999: A.ABEILLÉ et D.GODARD, La place de l'adjectif épithète en français : le poids des mots, Recherches Linguistiques, 28 : 9-31.

ABEILLÉ et GODARD, 2000: A. ABEILLÉ et D.GODARD, French Word Order and Lexical Weight, dans R.BORSLEY (ed), The Nature and Function of Syntactic Categories, Syntax and Semantics 32 : 325-360.

ARNOLD et al., 2000: J.ARNOLD, Th. WASOW, A.LOSONGCO et R.GINSTRO, Heaviness vs. Newness: The effects of complexity and information structure on constituent ordering, Language 76 : 28-55.

BARNBROOK, 1996: G. BARNBROOK, Language and Computers: A Practical Introduction to the Computer Analysis of Language, Edinburgh U. P.

DUPRÉ, 1972: DUPRÉ, Encyclopédie du bon français dans l'usage contemporain, Editions de Trévise, t.II.

GLLF, 1971: Grand Larousse de la Langue Française, t.1, Larousse.

FRANÇOIS et MANGUIN, à paraître : J. FRANÇOIS et J. MANGUIN, Dispute théologique, discussion oiseuse et conversation téléphonique : les collocations adjectivo-nominales au coeur du débat.

FUJIMURA et al., 2004: I. FUJIMURA, M. UCHIDA et H. NAKAO, *De* et *des* devant les noms précédés d'épithète en français : problème de *petit*, Le Poids des mots vol.1, Presses Universitaires de Louvain, 456-48.

FUJIMURA et al., à paraître : I. FUJIMURA, M. UCHIDA et H. NAKAO, Opposition entre *de* vs *des* devant les noms précédés d'épithète en français: portée du "poids", Acte des 3èmes journées de la linguistique du corpus, Presses Universitaires de Rennes.

HABERT et JACQUEMIN, 1993 : B. HABERT et C. JACQUEMIN, Noms composés, termes, dénominations complexes; problématiques linguistiques et traitements automatiques, Tal 34-2, 5-41.

MEJRI, 2003: S. MEJRI, Le figement lexical, Cahiers de lexicologie, 82-1, 23-39.

LE BIDOIS et LE BIDOIS, 1967: G. LE BIDOIS et R. LE BIDOIS, Syntaxe du français moderne, Tome 1, Éd. A. Picard.

TOGEBY, 1982 : K. TOGEBY, Grammaire française , vol.1 : Le Nom, Akademisk Forlag.

OAKES, 1998: M. OAKES, Statistics for Corpus Linguistics, Edinburgh U. P.
WASOW, 1997: Th. WASOW, Remarks on Grammatical Weight, Language Variation and Change 9, 81-105.

Itsuko FUJIMURA
Nagoya University, GSID,
Furocho, Chikusa-ku,
Nagoya, 464-8601
JAPON
fujimura@nagoya-u.jp